Welcome to the Sentinel Innovation Center Webinar Series

The webinar will begin momentarily

Please visit www.sentinelinitiative.org for recordings of past sessions and details on upcoming webinars.

Sentinel

Exploring the Opportunities and Challenges of Common Data Model Representations of NLP Output of EHR Data

Michael E. Matheny, MD, MS, MPH

Co-Director, Center for Improving the Publics' Health Through Informatics
Associate Professor, Departments of Biomedical Informatics, Medicine, and Biostatistics
Vanderbilt University Medical Center

Associate Director for Data Analytics, VINCI
Associate Director, Advanced Fellowship in Medical Informatics
Tennessee Valley Healthcare System VA

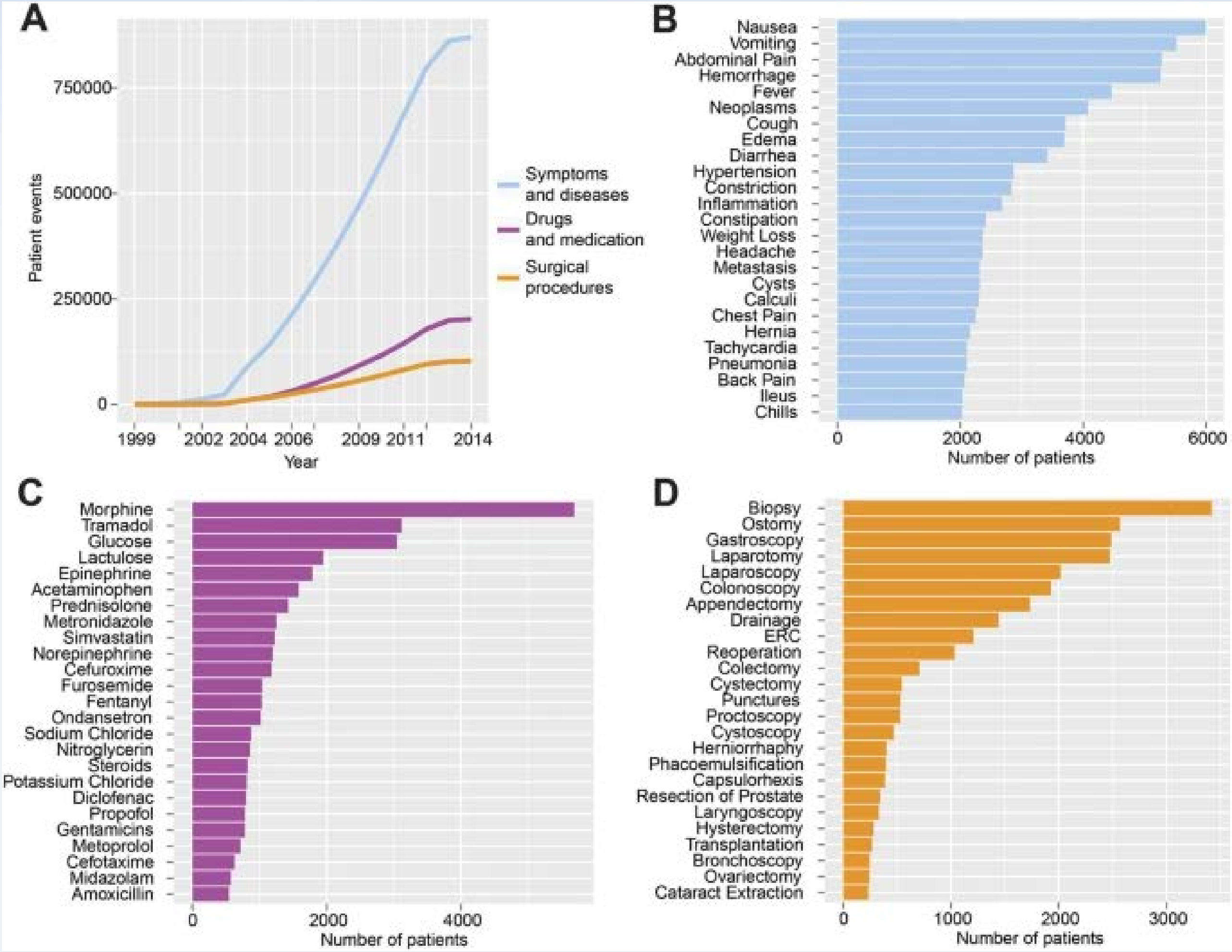
Twitter: @MichaelEMatheny
Email: michael.Matheny@Vanderbilt.edu, michael.Matheny@vumc.org, michael.Matheny@va.gov

Objectives

- Background on Types & Use of NLP
- Describe Foundational Concepts around Data Modeling for EHR-Derived Data
- Challenges In Storing Free Test In A CDM
- NLP Provenance
- Veracity & Mapping Considerations
- Same Framework Applies to Other Derived Data

Natural Language Processing

Information Content In Free Text



Norway
1.1 mil docs
7.7 k pts
GI Cancer Pts
Docs Processed to MeSH

Kasper Jensen, et al. 2017, Nature Scientific Reports. PMID: 28387314

General Types of Clinical NLP

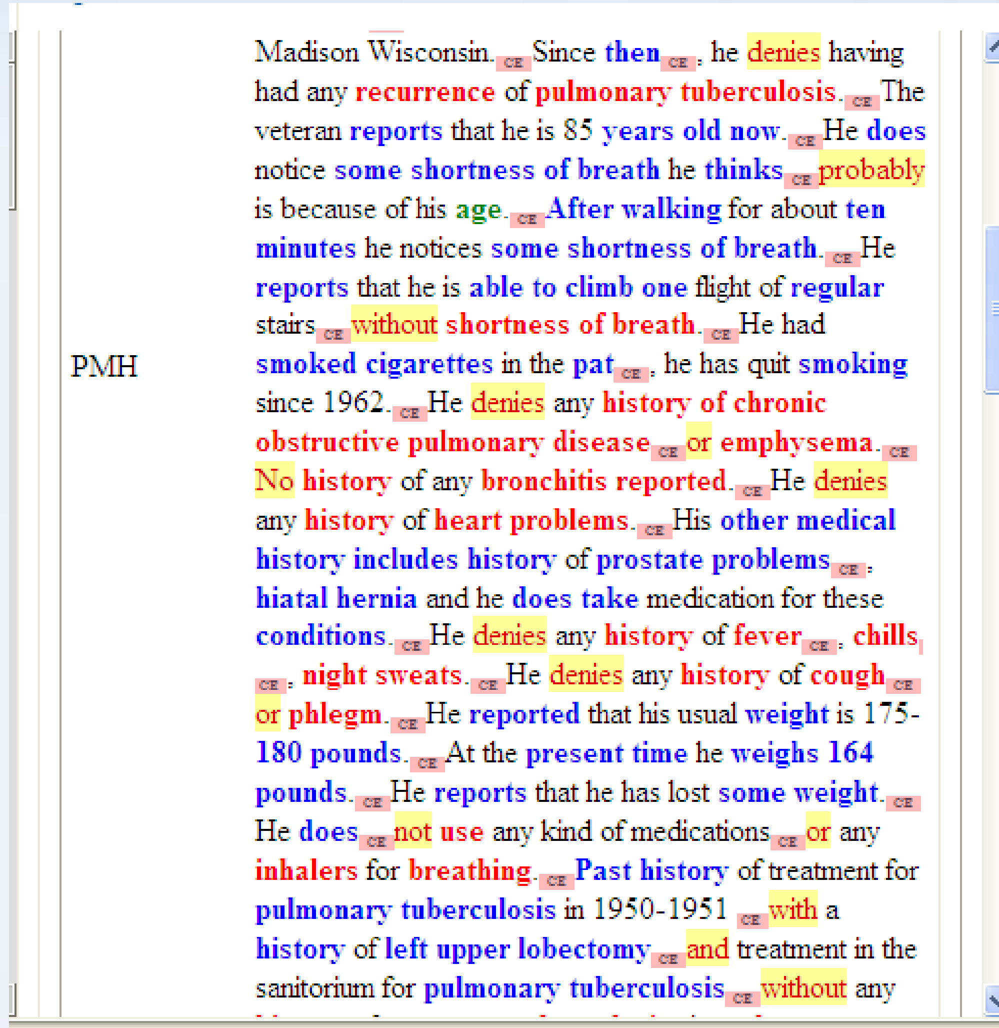
By Methods Approach
Rule-Based
Terminal Hybrid (Rule Based -> Machine Learning)
Machine Learning

By Output
Curated ("Focused") Clinical Features (Infectious Symptoms, Smoking History, etc)
Generalized Controlled Vocabular Mapping (MedLEE, MetaMap, cTakes/yTEX, KnowledgeMap, CLAMP, and others)
Feature/Vector Generation without CV Mapping (Word2Vec, Doc2Vec, Bag Of Words, and others)

Use of Generalized NLP to Detection Infectious Signs & Symptoms

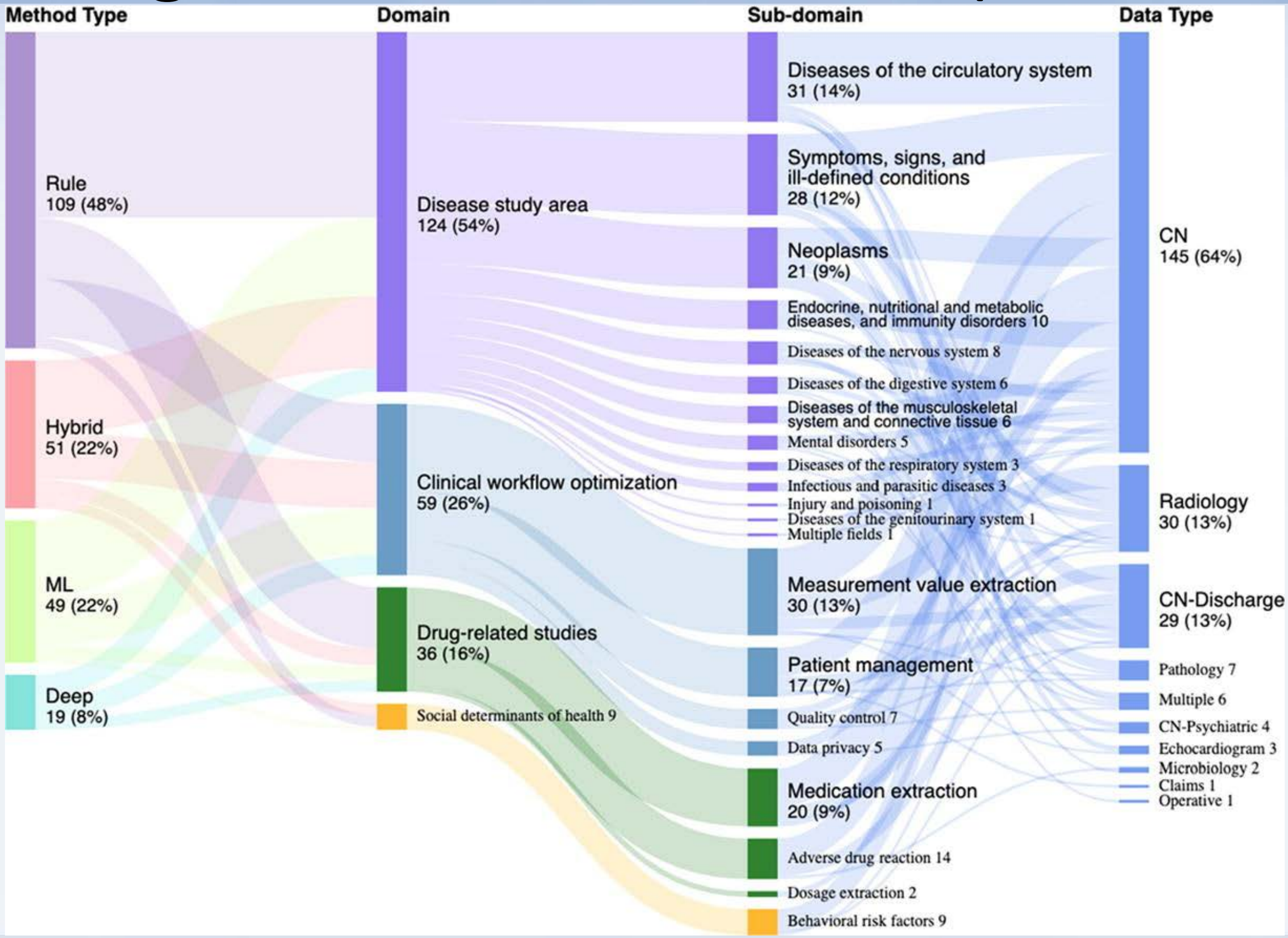
- Emergency Department & Primary Care Notes
- Annotation Reference Standard (**Supervised**)
- Symptom detection
  - Precision: 0.91
  - Recall: 0.84
  - *F* measure 0.87

Example NLP Output (Scrubbed)



Source: Matheny ME, et al, Brown SB. Int. J. Biomed. Inform, 2012;81(3):143-56.

Thorough Clinical NLP Review (2009-2019)



Sunyang Fu, et al, Hongfang Liu. Aug 2020, JBI. PMID: 32768446.

Some Current State of the Art Examples

Domain	Task	F-Measure	Method	Model
Clinical Workflow Optimization	ID of Risk Factors for CAD	0.93	Deep Learning	BioBERT
Clinical Workflow Optimization	I2b2 2006 1-B Auto De-ID of PHI	0.946	Deep Learning	BioBert
Drug-Related	I2b2 2020/VA medical problem extraction	0.903	Deep Learning	BERT-Large
Drug-Related	I2b2 2009 medications (Detailed sub-features)	0.857	Terminal Hybrid	CRF, SVM, Context Engine
Diseases	ShARe/CLEFE 2013 Named Entity Recognition	0.77	Deep Learning	BERT-Base
Diseases	SemEval 2014 Task 7: ID of Diseases and Disorders	0.807	Deep Learning	BERT-Large
Diseases	SemEval 2014 Task 14: NER & Template Slot Filling	0.817	Deep Learning	BERT-Large

Sunyang Fu, et al, Hongfang Liu. Aug 2020, JBI. PMID: 32768446.

Need for Standardized NLP Output Representation

- Expensive Computation Time For NLP
  - Late Binding (Real Time) of Free Text Large Still Infeasible
- Multi-site analyses benefit strongly from increased standardization of data representations

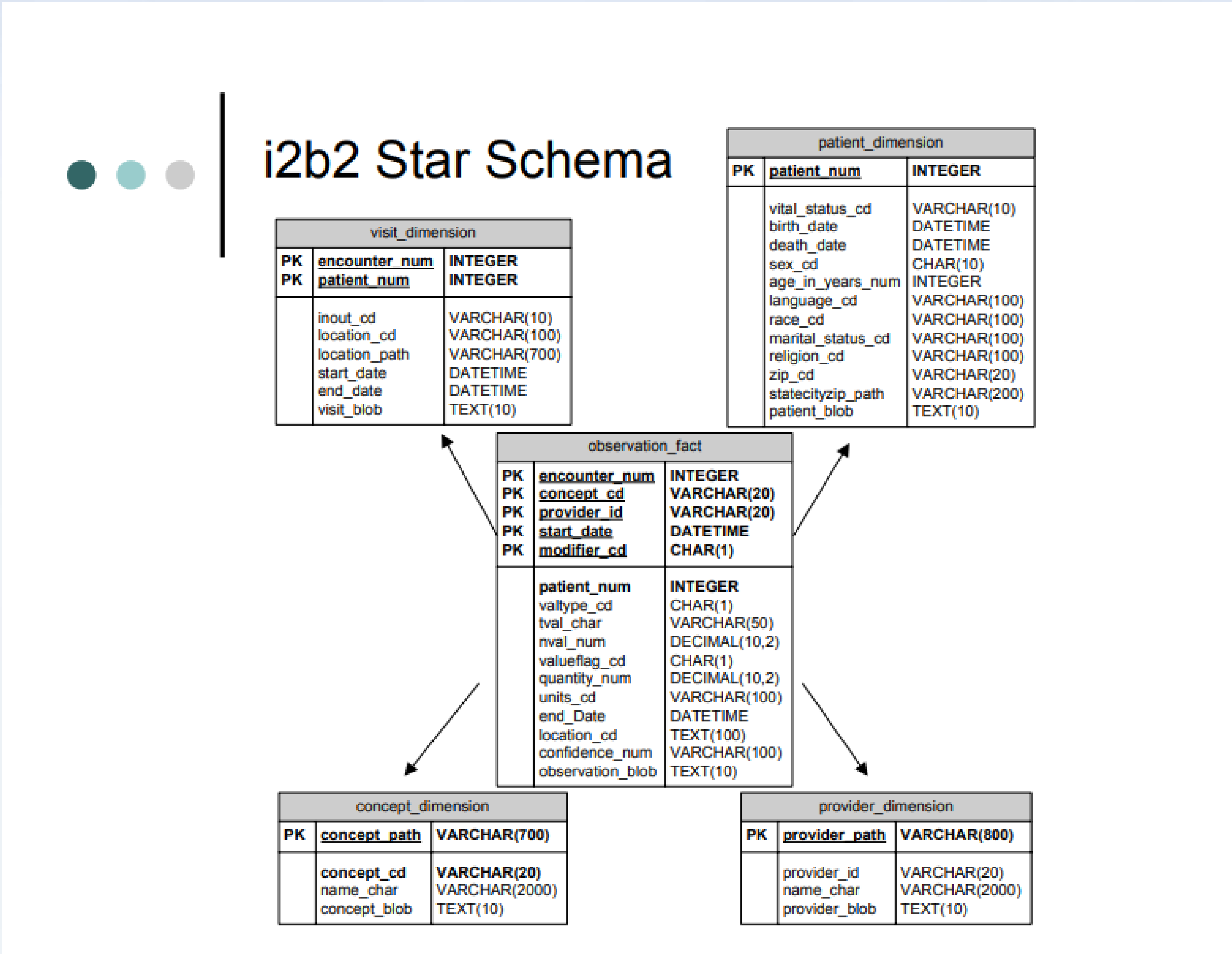
Common Data Models

Current State: Mature CDM Feature Matrix

Data Domains	I2B2 v1.7.12	PCORNet V5.1	OMOP v6	Sentinel V7.0
Person (Demographics)	Full	Full	Full	Full
Person Relationship (Family)	No	No	Full	Partial (Mom-Infant)
Enrollment	Yes	Full	Full	Full
Encounters	Yes	Full	Full ($)	Full
Medications	Partial	Full	Full (+,$)	Full
Medical Devices	Generic (EAV)	Generic (EAV)	Full ($)	No
Diagnosis	Generic	Full	Full (+)	Full
Procedures	Generic	Full	Full (+,$)	Full ($)
Provider	Yes	Yes	Yes	Embedded
Death	Generic	Full	Full	Full
Laboratory	Generic	Full	Full	Full
Free Text	Generic (EAV)	Generic (EAV)	Full	No
Vaccination	Generic	Generic	Generic	Full
Vital Signs	Generic	Full	Generic	Full
Meta Features				
Single Controlled Vocabulary Per Domain	No	Yes	Yes	No
User Community for Data Visualization and Statistical Module Building	Yes	Yes	Yes	Yes
“Catch All” Table (EAV), allows ‘all other’ facts storage (counted Generic)	Yes	Yes	Yes	No
Derived tables for Eras or Duration (represented with + in-table)	No	No	Yes	No
Health Economics / Cost Tables (represented with $ in-table)	Generic	No	Yes	No

I2B2 Star Schema

i2b2
Informatics for Integrating Biology & the Bedside



https://www.i2b2.org/events/slides/Workshop1.pdf - Shawn Murphy, Vivian Gainer (Source)

PCORNet v5.1

pcornet® The National Patient-Centered Clinical Research Network

PCORnet Common Data Model v5.1

New to v5.0

DEMOGRAPHIC
PATID
ETC...
PAT_PREF_LANGUAGE_SPOKEN

ENCOUNTER
ENCOUNTERID
PATID
ADMIT_DATE
ENC_TYPE
ETC...
PAYER_TYPE_PRIMARY
PAYER_TYPE_SECONDARY
FACILITY_TYPE

DIAGNOSIS
DIAGNOSISID
PATID
DX
DX_TYPE
DX_SOURCE
DX_DATE
ETC...
DX_POA

PROCEDURES
PROCEDURESID
PATID
PX
PX_TYPE
ETC...
PPX

CONDITION
CONDITIONID
PATID
CONDITION
CONDITION_TYPE
CONDITION_SOURCE
ETC...

LAB_RESULT_CM
LAB_RESULT_CM_ID
PATID
RESULT_DATE
LAB_RESULT_SOURCE
LAB_LOINC_SOURCE
ETC...
RESULT_SNOMED

PRESCRIBING
PRESCRIBINGID
PATID
ETC...
RX_DOSE_ORDERED
RX_DOSE_ORDERED_UNIT
RX_ROUTE
RX_SOURCE
RX_DISPENSE_AS_WRITTEN
RX_PRN_FLAG

DISPENSING
DISPENSINGID
PATID
DISPENSE_DATE
NDC
DISPENSE_SOURCE
ETC...
DISPENSE_DOSE_DISP_UNIT
DISPENSE_ROUTE

MED_ADMIN
MEDADMINID
PATID
MEDADMIN_START_DATE
ENCOUNTERID
MEDADMIN_START_TIME
MEDADMIN_STOP_DATE
MEDADMIN_STOP_TIME
PRESCRIBINGID
ETC...
MEDADMIN_SOURCE

VITAL
VITALID
PATID
MEASURE_DATE
VITAL_SOURCE
ETC...

ENROLLMENT
PATID
ENR_START_DATE
ENR_BASIS
ETC...

DEATH
PATID
DEATH_SOURCE
ETC...

DEATH_CAUSE
PATID
DEATH_CAUSE
DEATH_CAUSE_CODE
DEATH_CAUSE_TYPE
DEATH_CAUSE_SOURCE
ETC...

PROVIDER
PROVIDERID
PROVIDER_SEX
PROVIDER_SPECIALTY_PRIMARY
PROVIDER_NPI
PROVIDER_NPI_FLAG

HARVEST
NETWORKID
DATAMARTID
ETC...

PCORNET_TRIAL
PATID
TRIALID
PARTICIPANTID
ETC...

PRO_CM
PRO_CM_ID
PATID
ENCOUNTERID
PRO_DATE
PRO_TIME
PRO_TYPE
PRO_ITEM_NAME
PRO_ITEM_LOINC
PRO_RESPONSE_TEXT
PRO_RESPONSE_NUM
PRO_METHOD
PRO_MODE
PRO_CAT
PRO_SOURCE
ETC...

IMMUNIZATION
IMMUNIZATIONID
PATID
VX_CODE
VX_CODE_TYPE
VX_STATUS
ETC...

HASH_TOKEN
PATID
TOKEN_01
ETC...
TOKEN_16

OBS_CLIN
OBSCLINID
PATID
ENCOUNTERID
OBSCLIN_PROVIDERID
OBSCLIN_DATE
OBSCLIN_TIME
OBSCLIN_TYPE
OBSCLIN_CODE
ETC...
OBSCLIN_SOURCE
ETC...
RAW_OBSCLIN_UNIT

OBS_GEN
OBSGENID
PATID
ENCOUNTERID
OBSGEN_PROVIDERID
OBSGEN_DATE
OBSGEN_TIME
ETC...
OBSGEN_SOURCE
RAW_OBSGEN_UNIT

LDS_ADDRESS_HISTORY
ADDRESSID
PATID
ADDRESS_USE
ADDRESS_TYPE
ADDRESS_PREFERRED
ETC...

Bold font indicates fields that cannot be null due to primary key definitions or record-level constraints.

Sentinel v7

Administrative Data

Enrollment	Demographic	Dispensing	Encounter	Diagnosis	Procedure
Patient ID	Patient ID	Patient ID	Patient ID	Patient ID	Patient ID
Enrollment Start & End Dates	Birth Date	Dispensing Date	Service Date(s)	Service Date(s)	Service Date(s)
Drug Coverage	Sex	National Drug Code (NDC)	Encounter ID	Encounter ID	Encounter ID
Medical Coverage	Zip Code	Days Supply	Encounter Type and Provider	Encounter Type and Provider	Encounter Type and Provider
Medical Record Availability	Etc.	Amount Dispensed	Facility	Diagnosis Code & Type	Procedure Code & Type
			Etc.	Principal Discharge Diagnosis	Etc.

Clinical Data

Lab Result	Vital Signs
Patient ID	Patient ID
Result & Specimen Collection Dates	Measurement Date & Time
Test Type, Immediacy & Location	Height & Weight
Logical Observation Identifiers Names and Codes (LOINC®)	Diastolic & Systolic BP
Etc.	Tobacco Use & Type
	Etc.

Registry Data

Death	Cause of Death	State Vaccine
Patient ID	Patient ID	Patient ID
Death Date	Cause of Death	Vaccination Date
Source	Source	Admission Date
Confidence	Confidence	Vaccine Code & Type
Etc.	Etc.	Provider
		Etc.

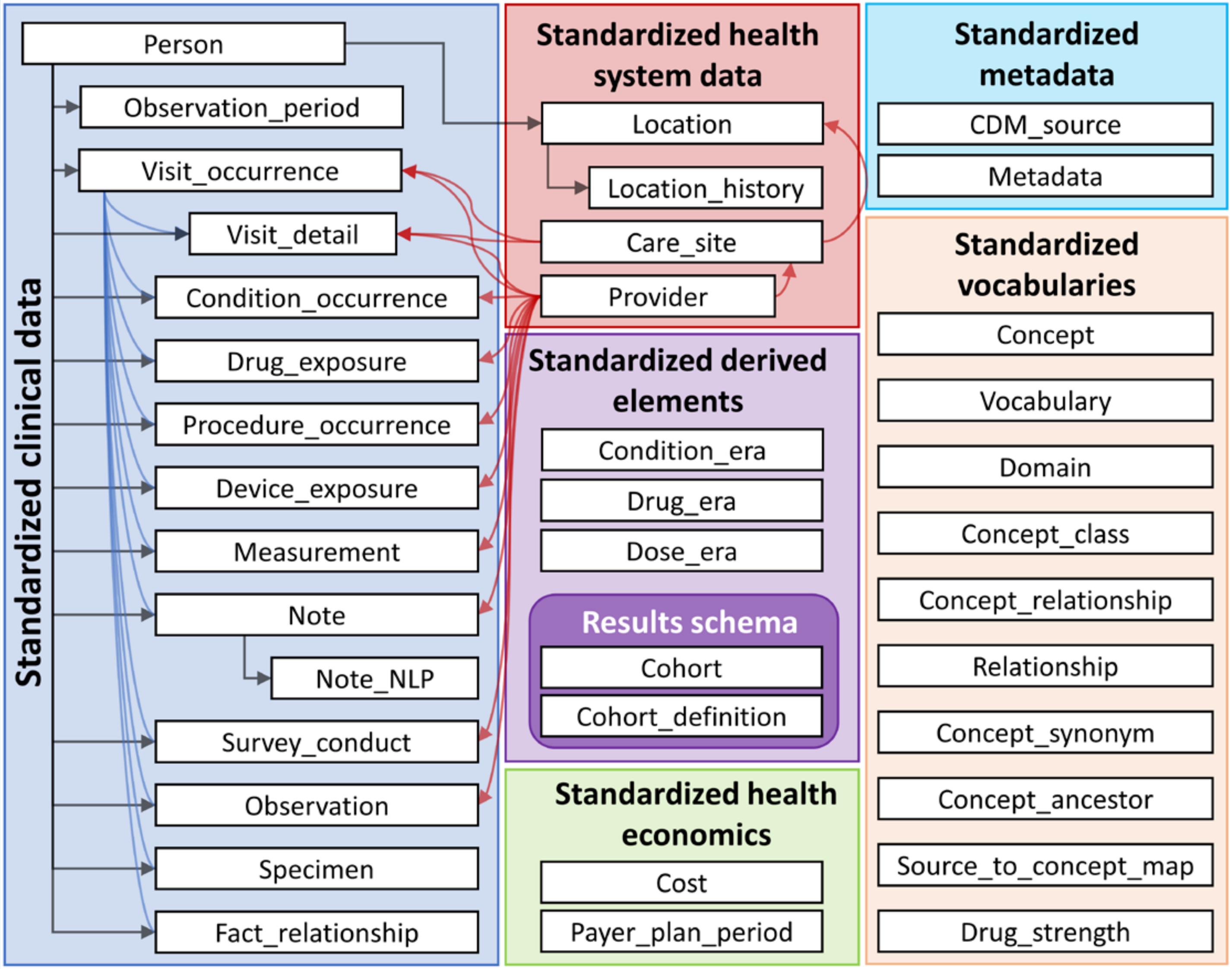
Inpatient Data

Inpatient Pharmacy	Inpatient Transfusion
Patient ID	Patient ID
Administration Date & Time	Administration Start & End Date & Time
Encounter ID	Encounter ID
National Drug Code (NDC)	Transfusion Administration ID
Route	Transfusion Product Code
Dose	Blood Type
Etc.	Etc.

Mother-Infant Linkage Data

Mother-Infant Linkage
Mother ID
Mother Birth Date
Encounter ID & Type
Admission & Discharge Date
Child ID
Child Birth Date
Mother-Infant Match Method
Etc.

OMOP CDM v6.0



https://github.com/OHDSI/CommonDataModel

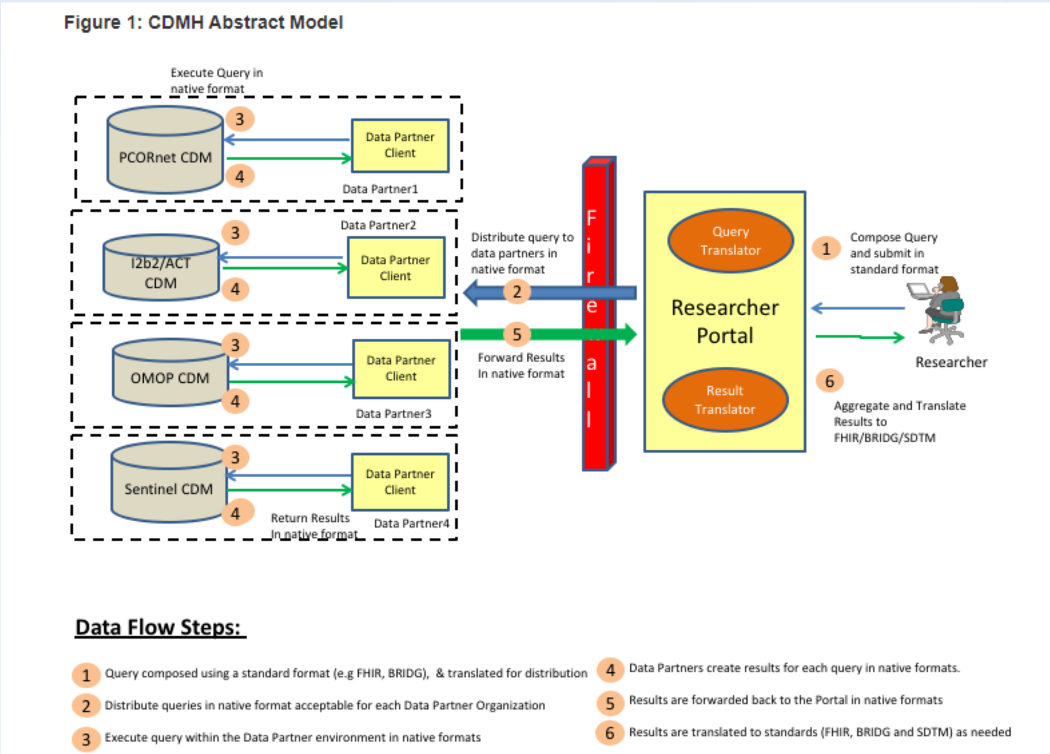
OMOP Note Domain

- Data Elements
  - Patient ID
  - Provider ID
  - Linked Encounter
  - Type of Event for Note
  - HL7 LOINC Document Type Vocabulary
  - Free Text Note Title
  - Date/Time of Authorship

OMOP Note NLP Domain

- Data Elements
  - Link to NOTE Table (Source Document)
  - Text and words around text
  - Mapping to a Standardized Vocabulary concept
  - NLP Algorithm/Tool ID
  - Date/Time of Output Concept
  - Date/Time of NLP processing
  - Free Text Note Title
  - Modifiers & Temporal Terms

Common Data Model Harmonization Project



- Interoperability Project
- FDA Led
  - NCI, NCATS, ONC, NLM participating
- Decision to create an intermediate data model: BRIDG
- Provides operability to:
  - PCORNet
  - Sentinel
  - OMOP
  - Ib2b

- Released 04/19, still uncertain how accurate

Challenges to Translating NLP to a CDM

Challenges in Framing Standardized NLP Outputs

- Relevance: Most tasks do not need or want ***ALL*** clinical text as inputs
- Standardization: NLP outputs benefit from structured mappings to be downstream usable
- Veracity: How to represent differential veracity of output?
- Portability: Need to Incorporate Local Updating in Features (Prior Sentinel Presentation) or Algorithm
- Provenance: Necessary to retain provenance of ***source*** and ***transformation process***

Representation of Source Text Data

- For Scalability and Re-usability, need to have standardized data elements:
  - Note Content
  - Date/Time of Creation
  - Note Title and Context Meta-Data
  - Author and Co-Signer (and specialty)
  - Episode of Care / Encounter / Visit

Under-Specification of Document Titles

LOINC Document Ontology Axes

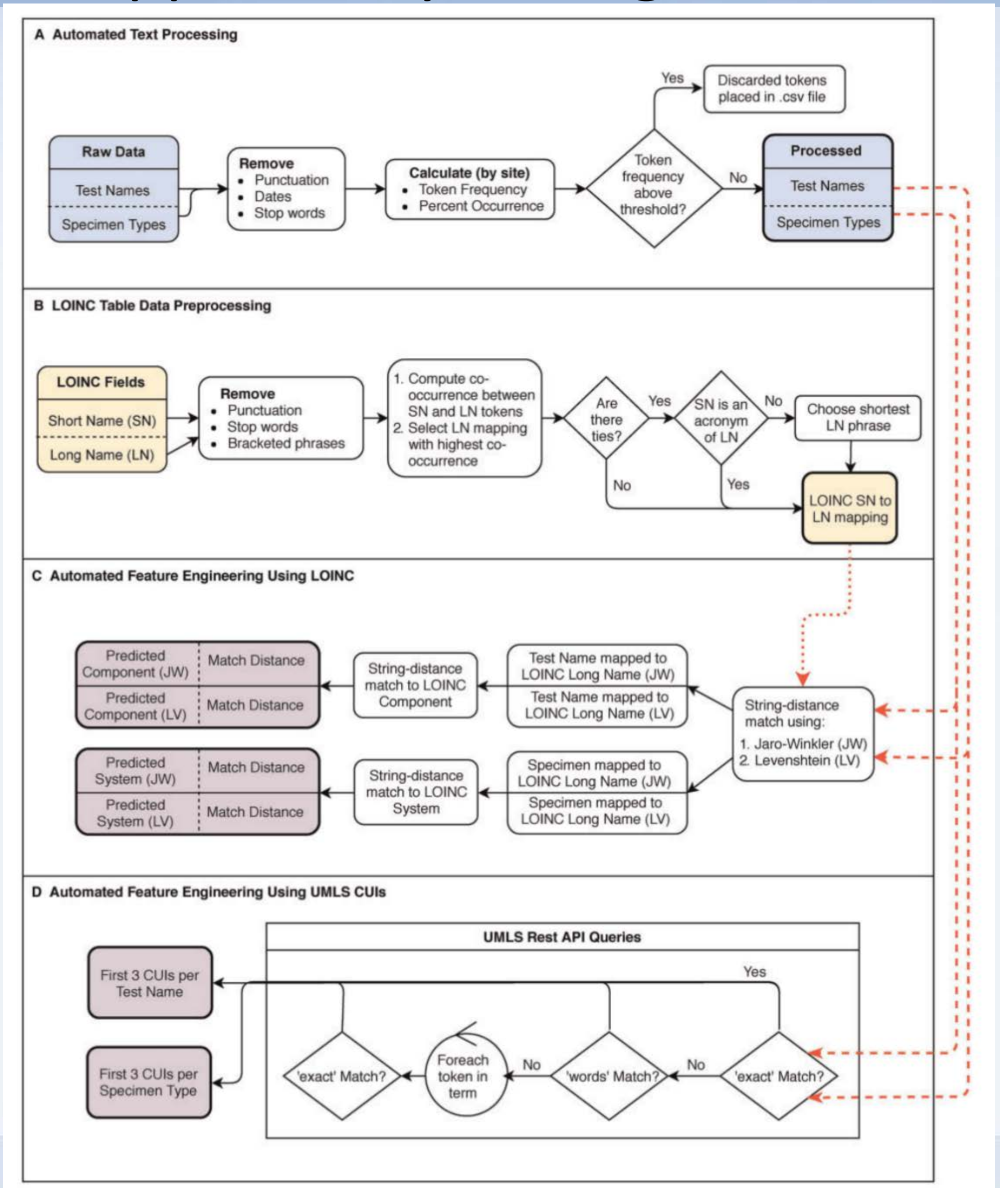
- Subject Matter Domain
- Type of Service
- Role
- Setting
- Kind of Document

LOINC DO Axis	**West Campus note titles (1644)**	**West Campus Classification Rate**	**East Campus note titles (1124)**	**East Campus Classification Rate**
Subject Matter Domain	1111	67.6%	777	69.1%
Type of Service	977	59.4%	776	69.0%
Role	569	34.6%	629	56.0%
Setting (no default value)	475	28.9%	143	12.7%
Map to no axis	135	8.2%	80	7.1%

New York Presbyterian Hospital (NYPH)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243240/

Opportunity to Augment Document Title Labeling



Lab Test LOINC Noisy Labeling

- 6.5 billion VA laboratory tests
- 130 Facilities
- 2215 LOINC Codes
- 29% LOINC Missing
- **Unsupervised** ML with partial or noisy labels, mapped a large portion of laboratory tests without LOINC to correct LOINC Codes

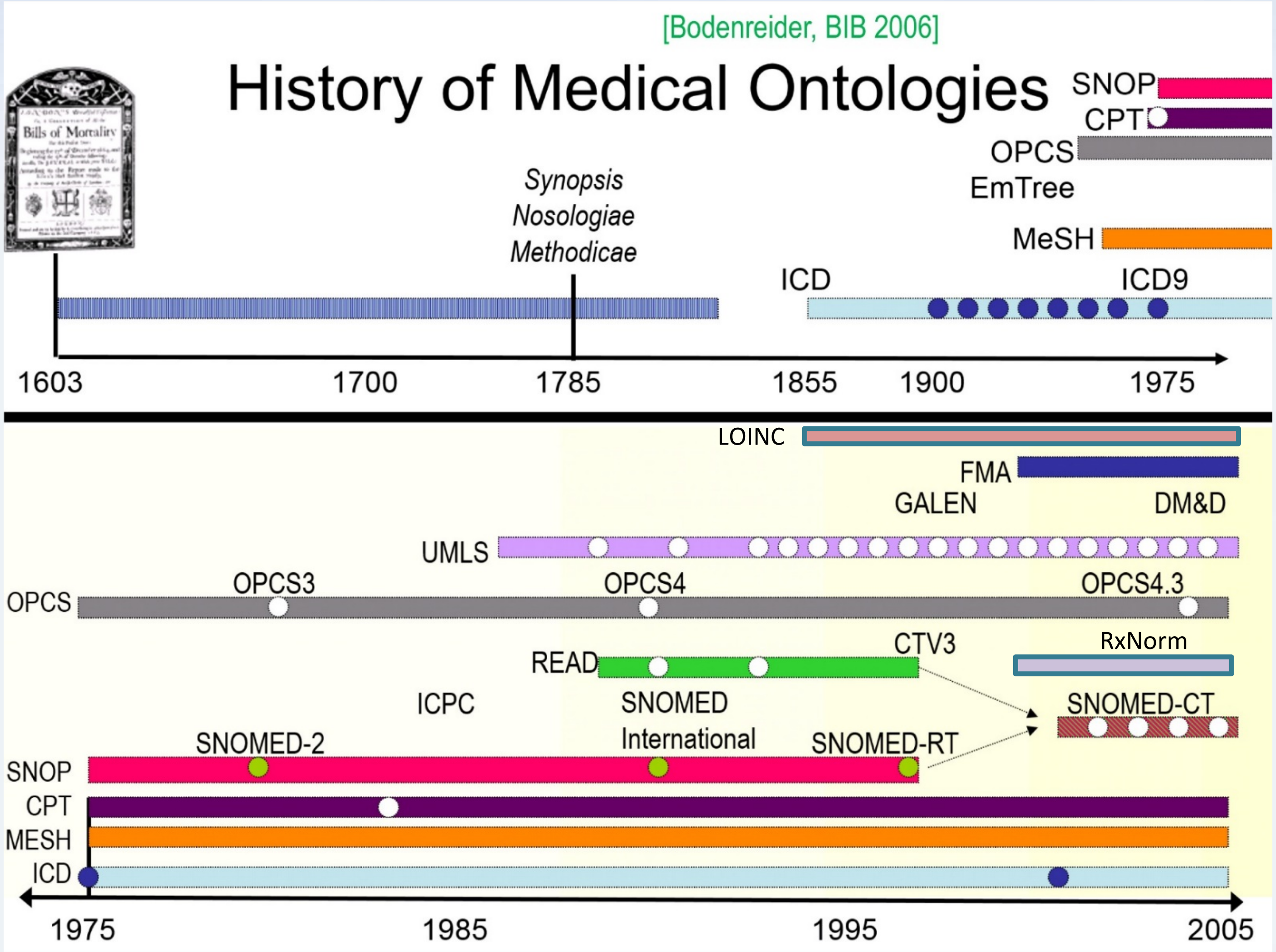
- Unlabeled Laboratory Data:
  - Correctly mapped 84.5% of tests that were not labeled
  - Fixed 1.1% mapped tests that were wrong

Source: Parr SK, et al, Matheny ME. JAMIA 2018; 25:1292-1300.

Standardization of NLP Outputs

- Word Vector or Non-Standardized Feature
  - Recommend Require Mapping to Controlled Vocabulary for use in CDM
  - Some feature generation can be used in ***supervised*** machine learning tasks downstream BUT harder to standardize outputs

Use of Controlled Terminologies



Most Common:

ICD9/10
CPT
SNOMED-CT
LOINC
RxNorm
MedDRA

Key Issues for Mapping Standardization

- Do you store mapped NLP outputs in a separate table structure?
- Do you merge NLP outputs with other data types with provenance?
- What threshold of veracity should be used to include in data model?
- Temporality – terms may represent time state different than text note (HARD!)
- Negation – negated terms… ignore or map separately?

Example: Mapping Coverage

Terminal Hybrid NLP Tool (Moonstone - Chapman)

Developed for NLP Task For Hospital Readmission Prediction

NLP-derived Variable	NLP representation	Structured Data Proxy	Vocabulary Mapping
Living Alone	Positive, negative, uncertain, no data	Marital Status (Partial)	ICD Code
Instrumental Support	Positive, negative, uncertain, no data	Health insurance type (Partial)	NONE (PRO Only)
Impaired ADL/IADL	Positive, negative, uncertain, no data	Partial / Varies	Aggregate, Base Eval SNOMED, LOINC, but impaired status represented as value
Medical Condition (causing impaired ADL)	Positive, negative, uncertain, no data	Varies / (ICD/CPT Codes)	ICD OR CPT Code
Medication Compliance	Positive, negative, uncertain, no data	Prescription fill gaps	SNOMED-CT
Depression	Positive, negative, uncertain, no data	Admin Codes	ICD code
Dementia	Positive, negative, uncertain, no data	Admin Codes	ICD code
Language Barrier	Positive, negative, uncertain, no data	None	SNOMED-CT

Ruth Reeves, Framework for Targeted NLP Output Mappings, 2020, Unpublished

Veracity: How Accurate is Enough?

- How accurate is enough?
- Concept: F-Measure: mean of Precision & Recall
Document/Patient Case: Sensitivity & Specificity
- Usually requires some silver or gold standard to evaluate performance on (annotation or noisy label)
- Storage and Reference Retention of Performance Also Important For Re-Use

Targeted Example: AKI Risk Factors

Category	Instances	TP	FP	FN	Precision (PPV)	Recall (Sensitivity)	F-Measure
Drug Exposures							
• ACE Inhibitor	575	553	8	22	0.986	0.962	0.974
• ARB	149	137	0	12	1.000	0.919	0.958
• Diuretic	733	684	4	49	0.994	0.933	0.963
• NSAID	233	201	4	32	0.980	0.863	0.918
Fluid Status							
• Diuresis	118	83	6	35	0.933	0.703	0.802
• Intake	694	412	46	282	0.900	0.594	0.715
• Intravascular Volume Condition	527	432	12	95	0.973	0.820	0.890
• Nausea/Vomiting/Diarrhea	719	674	25	45	0.964	0.937	0.951
• Weight Change	221	130	14	91	0.903	0.588	0.712
Radiographic Media Exposure							
• Contrast	2095	1858	240	237	0.886	0.887	0.886
• Potential Contrast	439	255	65	184	0.797	0.581	0.672
• Contrast Volume	4	0	0	4	-	0.000	0.000
Renal Status							
• Anatomical Kidney Status	57	9	4	48	0.692	0.158	0.257
• Nephrology Care Delivery	210	141	36	69	0.797	0.671	0.729
• Renal Function Impairment	449	368	44	81	0.893	0.820	0.855
• Renal Transplant Recipient	8	0	0	8	-	0.000	0.000
Total Concept Performance	**7231**	**5661**	**341**	**1570**	**0.921**	**0.821**	**0.868**

Category	Instances	TP	FP	FN	TN	Sensitivity	Specificity	NPV
Negation Performance		351	333	17	1049	0.954	0.759	0.984

Provenance

- What level of NLP intermediate products to retain for re-use and provenance:
  - Words mapped?
  - Position in document?
  - Algorithm used? Version? Algorithmic Coefficients?
  - Date processed?

Portability: Local Updating

Table 1. Corpus statistics of Mayo Clinic and SCH ($n = 298$ patients each)

Category	Mayo	SCH
Total no. of documents	9604	30 589
Total no. of tokens	2 212 389	10 117 963
No. of documents/patient, median (IQR)	27 (18)	80 (69.8)
No. of tokens/document, median (IQR)	186 (210)	103 (331)
No. of asthma-related concepts[a]/patient, median (IQR)	19.5 (32.8)	65.5 (88)
No. of asthma-related concepts/document, median (IQR)	2 (3)	1 (2)
No. of note types	16	32
No. of sections	17	54

[a]Each concept consists of a set of keywords. IQR: interquartile range.

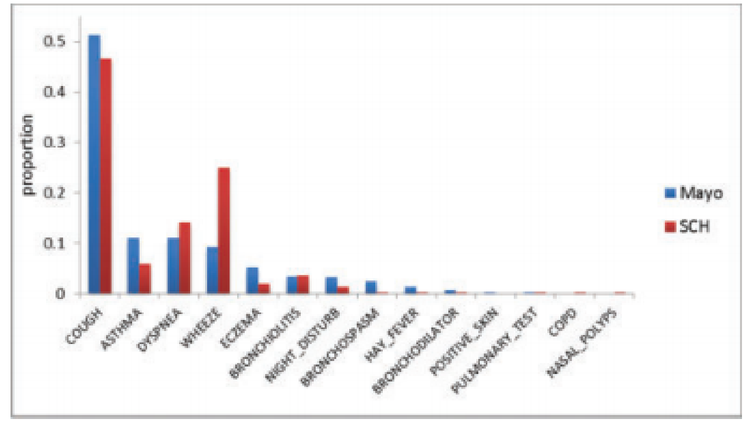


Figure 2. Distribution of asthma-related concepts.

Table 3. NLP-PAC performance for asthma ascertainment (Mayo vs Sanford)

Metrics	**Mayo**	**SCH Stage 1 (prototype)**	**SCH Stage 2 (refinement)**
Sensitivity	0.972	0.840	0.920
Specificity	0.957	0.924	0.964
PPV	0.905	0.788	0.896
NPV	0.988	0.945	0.973
F-score	0.937	0.813	0.908

Sohn S, et al. JAMIA Nov 2017. PMID: 29202185

Phenotyping: Hepatorenal Syndrome

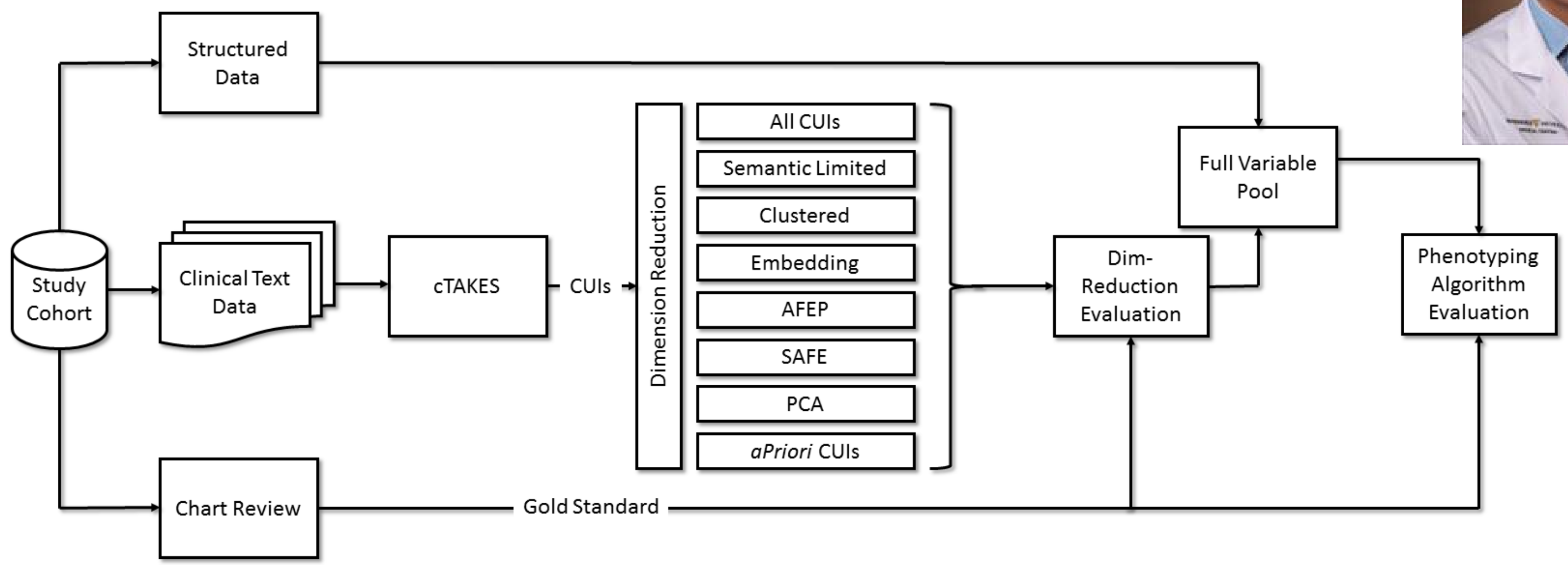
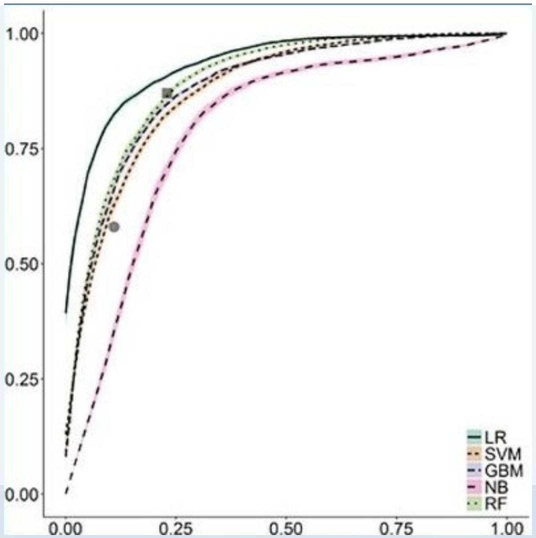
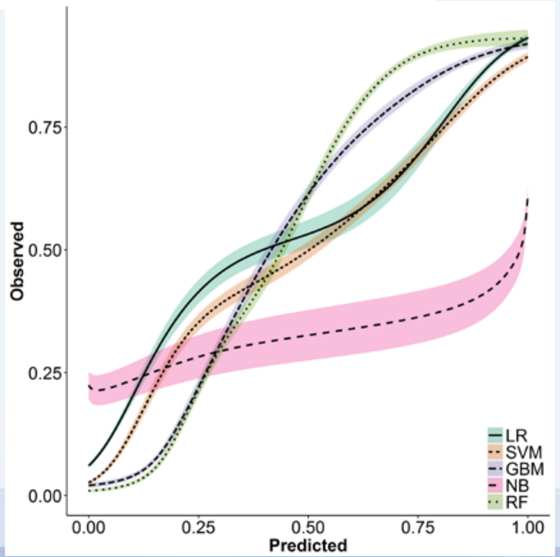


Chart Review Reference Standard (**Supervised**)

- 210/504 (41.6%) with HRS

Model	AUC (95% CI)
Logistic Regression	0.93 (0.92, 0.93)
Random Forest	0.91 (0.91, 0.91)
Support Vector Machine	0.90 (0.90, 0.91)





Koola JD, et al, Matheny ME. J. Biomed. Inform. 2018; 80:87-95.

Conclusions

- Standardized Representation of NLP Outputs is important for Scalability in Sentinel
- Key Challenges in Implementation:
  - Representation in CDM
  - Where to put Outputs (Embedded vs Separate)
  - Updating Algorithms for Local Environment
  - Documenting Performance
  - Maintaining Provenance
- Keep other Use Cases in Mind (Probabilistic Phenotyping)

Questions?

Michael E. Matheny, MD, MS, MPH
Twitter: @MichaelEMatheny
Email:
michael.Matheny@Vanderbilt.edu
michael.Matheny@vumc.org
michael.Matheny@va.gov