Natural Language Processing in OHDSI

OHDSI NLP Working Group

Noémie Elhadad; Columbia University

noemie.elhadad@columbia.edu

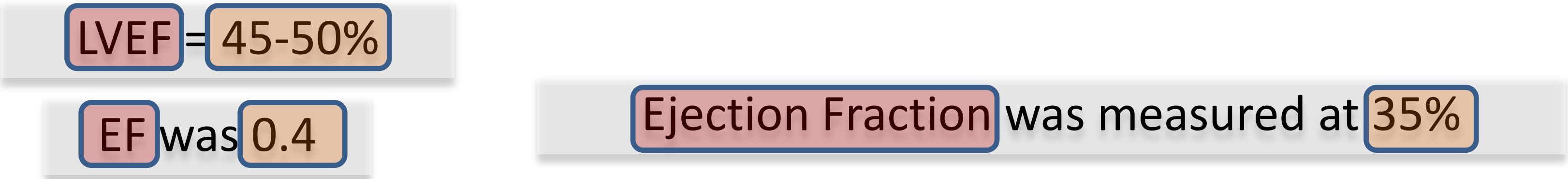
Natural Language Processing Working Group

- Promote the use of textual information from EHRs for observational studies under the OHDSI umbrella

- Schema for NLP output in the CDM
- IRBs for use of clinical texts
- NLP tools/pipelines for ETL
- Use cases and studies

Left Ventricular Ejection Fraction (LVEF)

- LVEF – Important indicator of heart disease
- **LVEF_Extractor** developed by VA VINCI NLP team
  - Processes clinical notes, outputs **LVEF values** into database
- Internal application
  - VA CDW has 2.7 billion notes, after keyword filter 165M notes
  - Resulted in a complete set of LVEF values extracted from VA documents
  - The dataset is available to any VA affiliated researcher similar to other structured sets
- External application
  - Shared **LVEF_Extractor** with other organizations



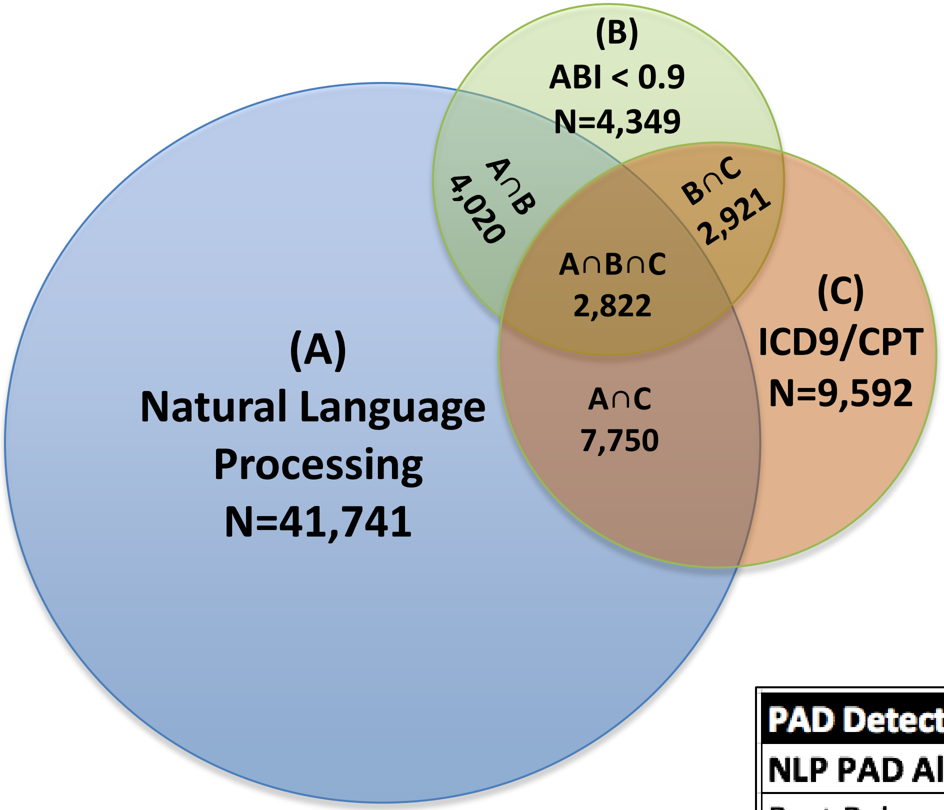
LVEF = 45-50%

EF was 0.4

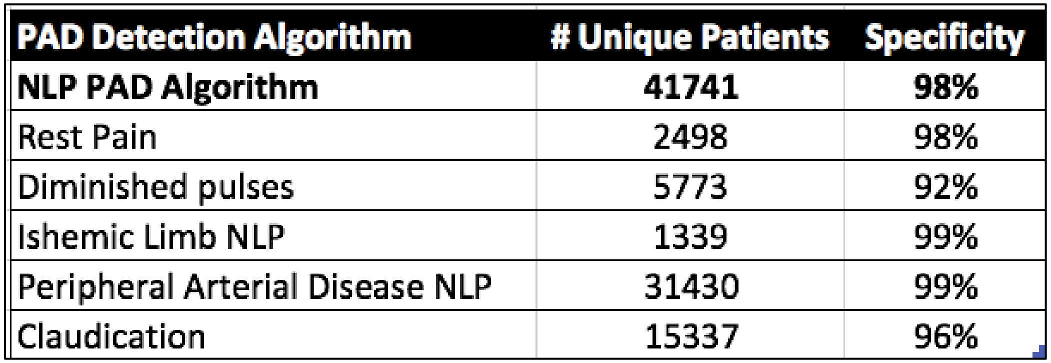
Ejection Fraction was measured at 35%

Patterson OV, Freiberg MS, Brandt C, DuVall SL. Unlocking echocardiogram report measures for heart disease research through natural language processing. In preparation.

Cohort detection – peripheral arterial disease



NLP detected 4x more patients than traditional algorithms. More importantly, many patients with PAD are missed using standard approaches.

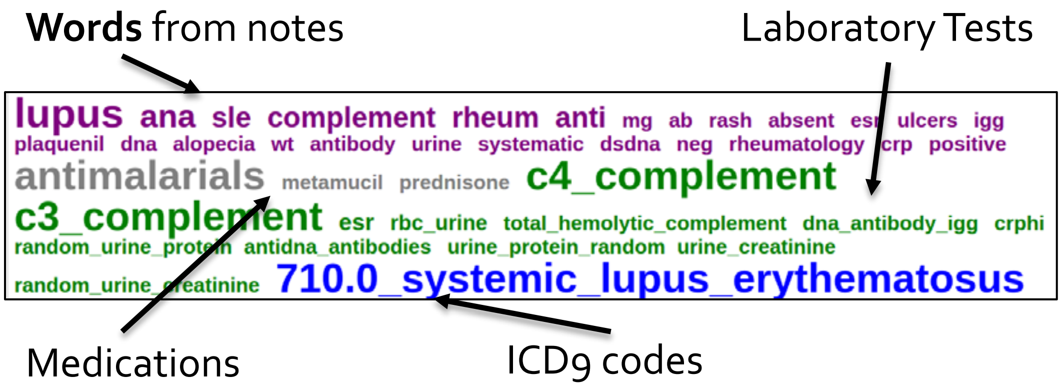


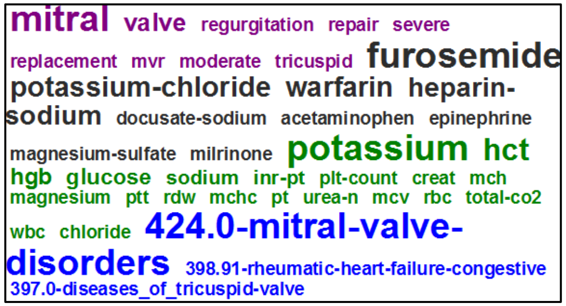
PAD Detection Algorithm	# Unique Patients	Specificity
NLP PAD Algorithm	**41741**	**98%**
Rest Pain	2498	98%
Diminished pulses	5773	92%
Ischemic Limb NLP	1339	99%
Peripheral Arterial Disease NLP	31430	99%
Claudication	15337	96%

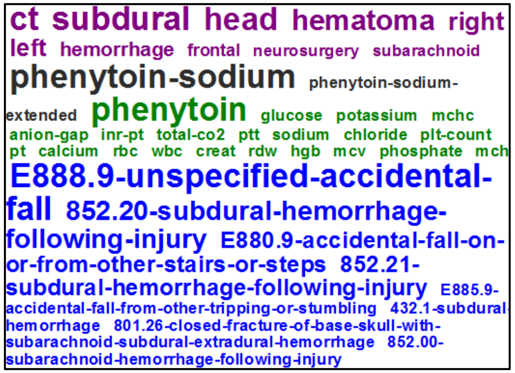
Duke JD, Chase M, Ring N, Martin J, Fuhr R, Hirch A. (2016) Natural Language Processing to Augment Identification of Peripheral Arterial Disease Patients in Observational Research. *American College of Cardiology Annual Symposium*.

Large-scale phenotyping

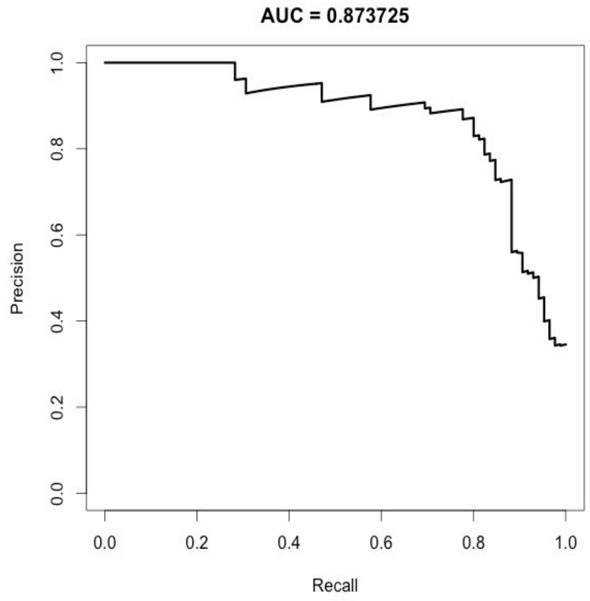
- Phenome model for joint detection of 750 phenotypes


Words from notes
Laboratory Tests
lupus ana sle complement rheum anti mg ab rash absent esr ulcers igg
plaquenil dna alopecia wt antibody urine systematic dsdna neg rheumatology crp positive
antimalarials metamucil prednisone c4_complement
c3_complement esr rbc_urine total_hemolytic_complement dna_antibody_igg crphi
random_urine_protein antidna_antibodies urine_protein_random urine_creatinine
random_urine_creatinine 710.0_systemic_lupus_erythematosus
Medications
ICD9 codes


mitral valve regurgitation repair severe
replacement mvr moderate tricuspid furosemide
potassium-chloride warfarin heparin-sodium docusate-sodium acetaminophen epinephrine
magnesium-sulfate milrinone potassium hct
hgb glucose sodium inr-pt plt-count creat mch
magnesium ptt rdw mchc pt urea-n mcv rbc total-co2
wbc chloride 424.0-mitral-valve-disorders 398.91-rheumatic-heart-failure-congestive
397.0-diseases_of_tricuspid-valve

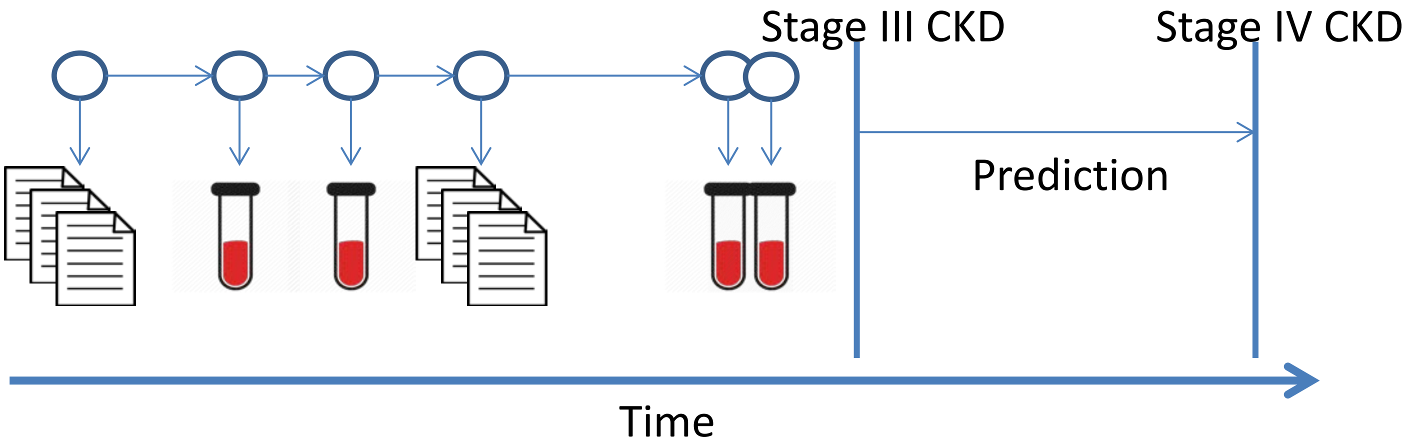

ct subdural head hematoma right
left hemorrhage frontal neurosurgery subarachnoid
phenytoin-sodium phenytoin-sodium-extended phenytoin glucose potassium mchc
anion-gap inr-pt total-co2 ptt sodium chloride plt-count
pt calcium rbc wbc creat rdw hgb mcv phosphate mch
E888.9-unspecified-accidental-fall 852.20-subdural-hemorrhage-following-injury E880.9-accidental-fall-on-or-from-other-stairs-or-steps 852.21-subdural-hemorrhage-following-injury E885.9-accidental-fall-from-other-tripping-or-stumbling 432.1-subdural-hemorrhage 801.26-closed-fracture-of-base-skull-with-subarachnoid-subdural-extradural-hemorrhage 852.00-subarachnoid-hemorrhage-following-injury

DM2 cohort identification (n=2,500)



Pivovarov R, Perotte A, Grave E, Angiolillo J, Wiggins C, Elhadad N. (2015) Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 58:156-165.

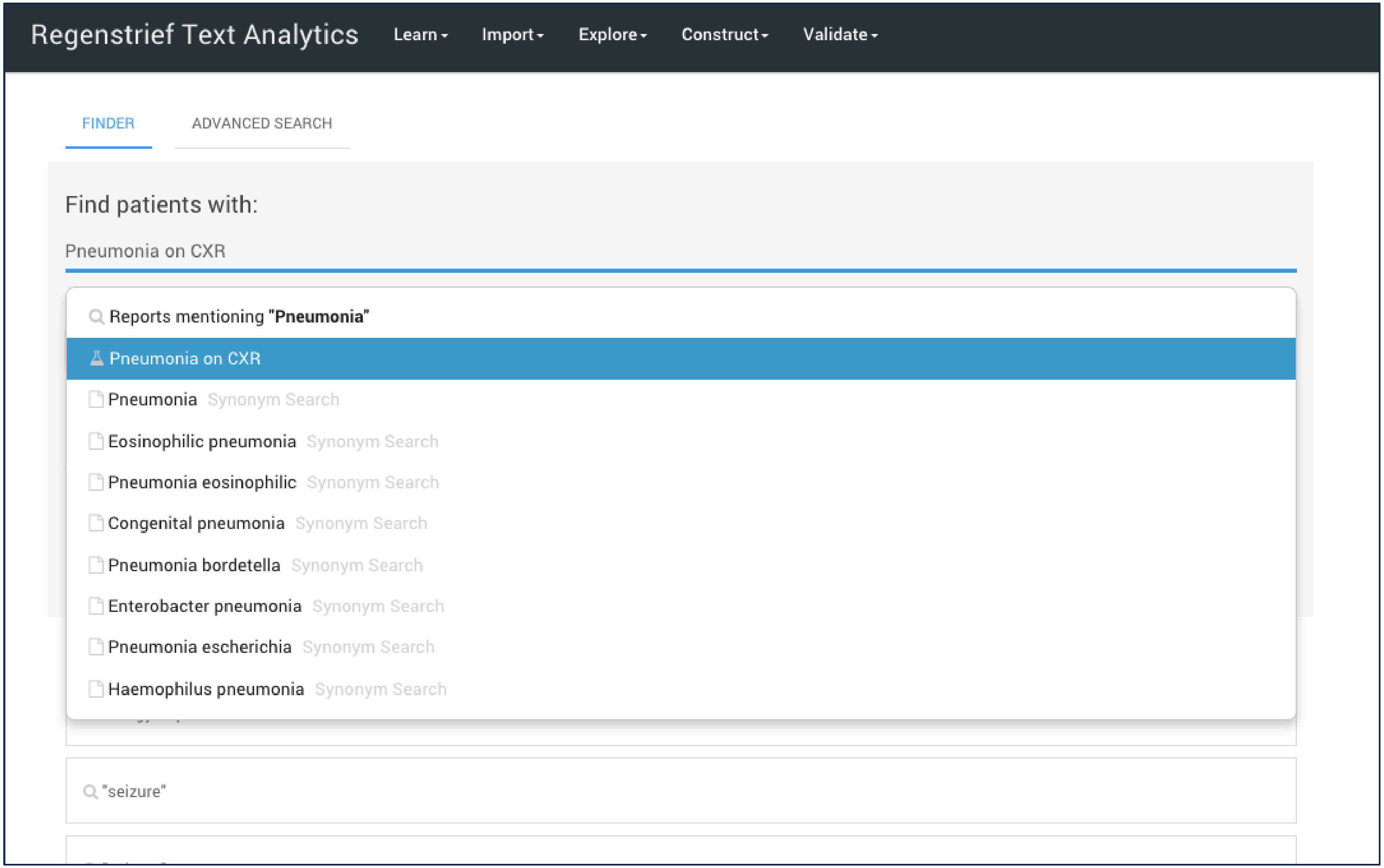
Survival analysis of CKD progression



Survival Model (n=2,617)	Concordence (n=291)
(Text + Lab) Kalman Filter	0.849
Lab Kalman Filter	0.836
Recent Labs	0.819
Text Kalman Filter	0.733
eGFR risk score	0.779

Perotte A, Ranganath R, Hirsch J, Blei D, Elhadad N (2015). Risk Prediction for Chronic Kidney Disease Progression Using Heterogeneous Electronic Health Record Data and Time Series Analysis. *J Am Med Inform Assoc*. 22(4):8720

Search and data exploration



Find patients with:

Pneumonia on CXR

Reports mentioning "**Pneumonia**"

Pneumonia on CXR

Pneumonia Synonym Search

Eosinophilic pneumonia Synonym Search

Pneumonia eosinophilic Synonym Search

Congenital pneumonia Synonym Search

Pneumonia bordetella Synonym Search

Enterobacter pneumonia Synonym Search

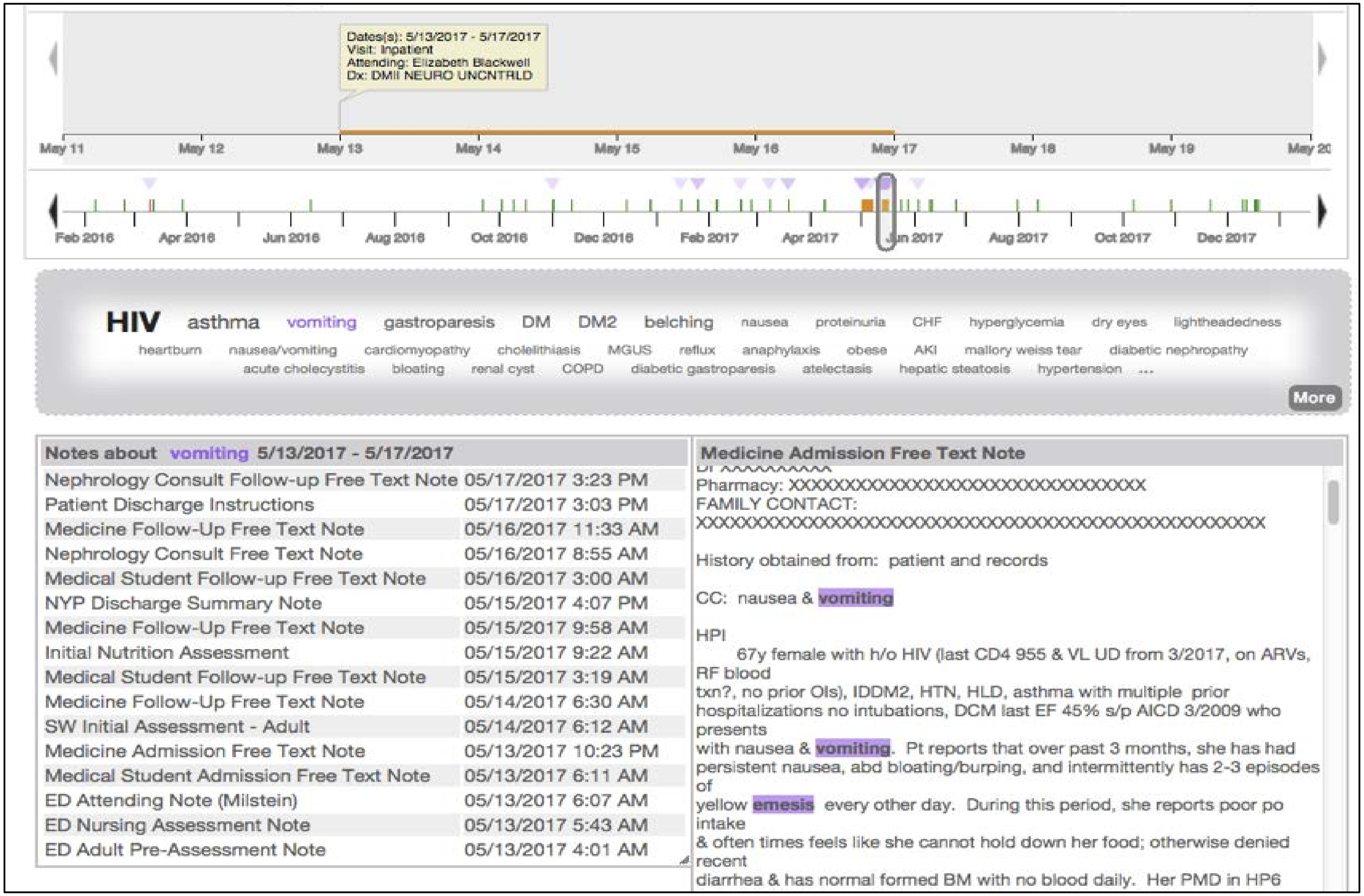
Pneumonia escherichia Synonym Search

Haemophilus pneumonia Synonym Search

"seizure"

Slide courtesy of Jon Duke

Patient-level visualization



Dates(s): 5/13/2017 - 5/17/2017
Visit: Inpatient
Attending: Elizabeth Blackwell
Dx: DMII NEURO UNCNTRLD

Notes about vomiting 5/13/2017 - 5/17/2017	
Nephrology Consult Follow-up Free Text Note	05/17/2017 3:23 PM
Patient Discharge Instructions	05/17/2017 3:03 PM
Medicine Follow-Up Free Text Note	05/16/2017 11:33 AM
Nephrology Consult Free Text Note	05/16/2017 8:55 AM
Medical Student Follow-up Free Text Note	05/16/2017 3:00 AM
NYP Discharge Summary Note	05/15/2017 4:07 PM
Medicine Follow-Up Free Text Note	05/15/2017 9:58 AM
Initial Nutrition Assessment	05/15/2017 9:22 AM
Medical Student Follow-up Free Text Note	05/15/2017 3:19 AM
Medicine Follow-Up Free Text Note	05/14/2017 6:30 AM
SW Initial Assessment - Adult	05/14/2017 6:12 AM
Medicine Admission Free Text Note	05/13/2017 10:23 PM
Medical Student Admission Free Text Note	05/13/2017 6:11 AM
ED Attending Note (Milstein)	05/13/2017 6:07 AM
ED Nursing Assessment Note	05/13/2017 5:43 AM
ED Adult Pre-Assessment Note	05/13/2017 4:01 AM

Medicine Admission Free Text Note

Pharmacy: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
FAMILY CONTACT:
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

History obtained from: patient and records

CC: nausea & vomiting

HPI
67y female with h/o HIV (last CD4 955 & VL UD from 3/2017, on ARVs, RF blood
txn?, no prior OIs), IDDM2, HTN, HLD, asthma with multiple prior
hospitalizations no intubations, DCM last EF 45% s/p AICD 3/2009 who presents
with nausea & vomiting. Pt reports that over past 3 months, she has had
persistent nausea, abd bloating/burping, and intermittently has 2-3 episodes of
yellow emesis every other day. During this period, she reports poor po intake
& often times feels like she cannot hold down her food; otherwise denied recent
diarrhea & has normal formed BM with no blood daily. Her PMD in HP6

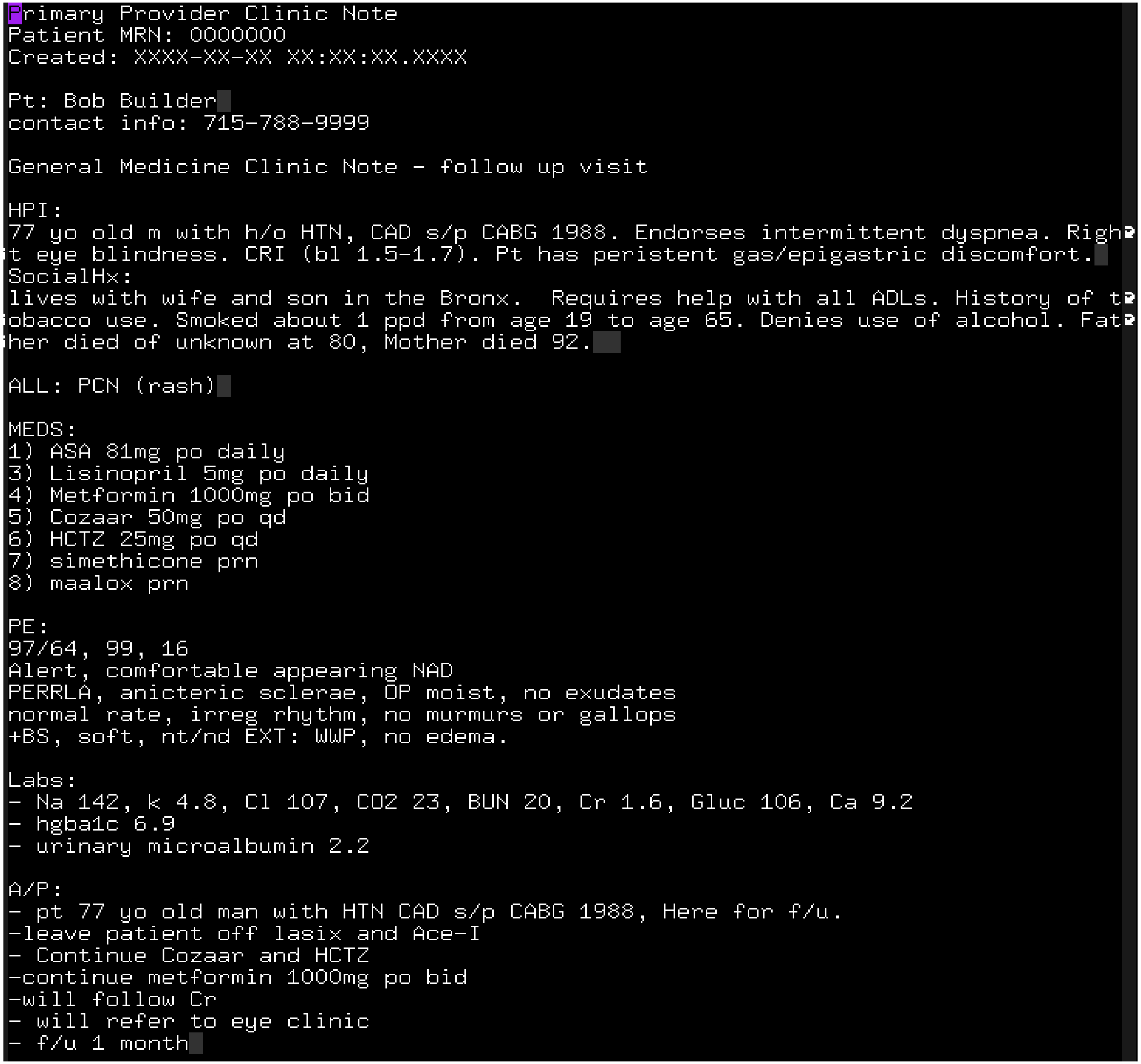
Hirsch J, Tanenbaum J, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, Ervits A, Vawdrey D, Sturm M, Elhadad N. (2015) HARVEST, a longitudinal patient record summarizer. *J Am Med Inform Assoc*. 22(2):263-274.

Natural Language Processing Working Group

- Promote the use of textual information from EHRs for observational studies under the OHDSI umbrella

- Schema for NLP output in the CDM
- IRBs for use of clinical texts
- NLP tools/pipelines for ETL
- Use cases and studies

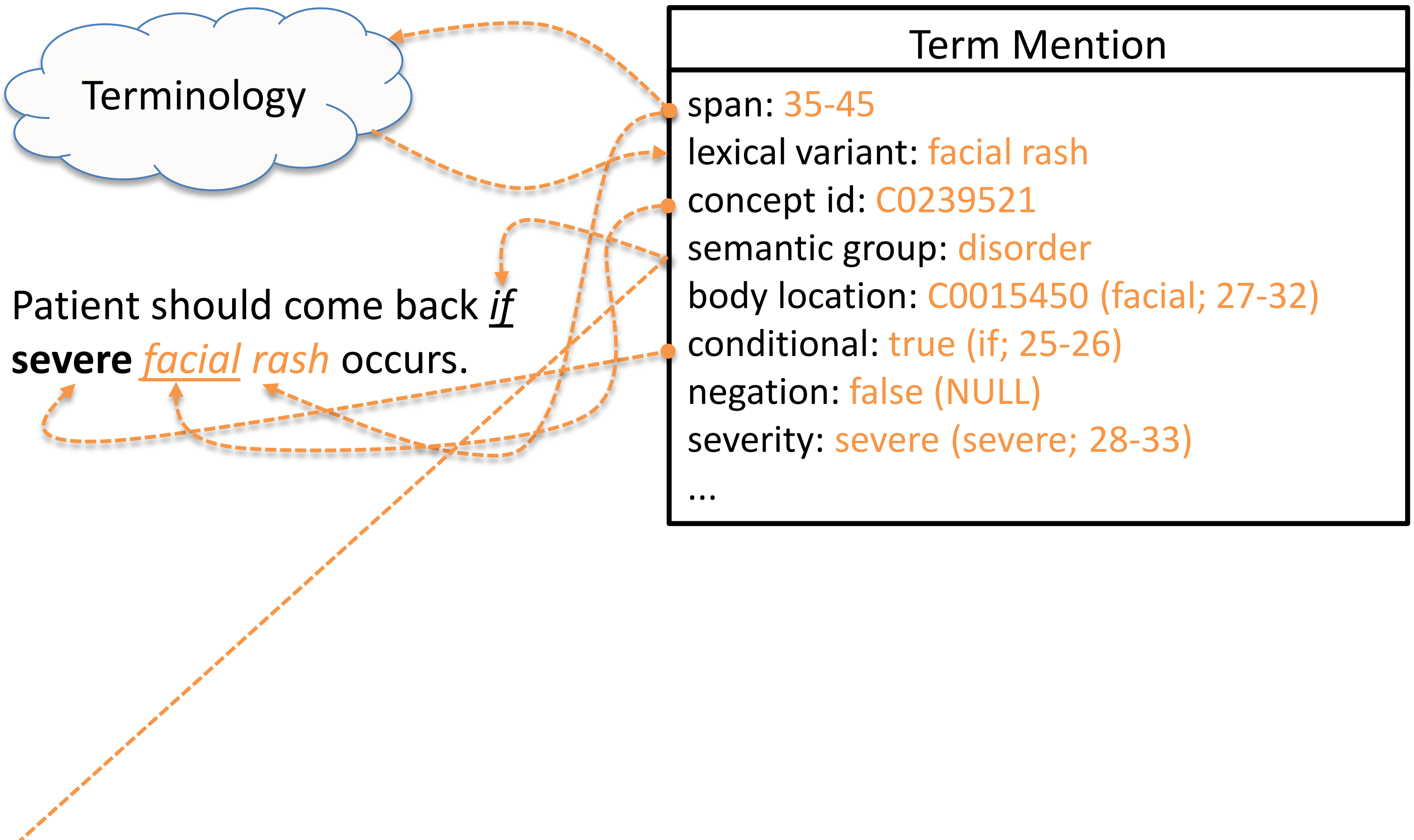
Unstructured clinical text



```
Primary Provider Clinic Note
Patient MRN: 0000000
Created: XXXX-XX-XX XX:XX:XX.XXXX

Pt: Bob Builder
contact info: 715-788-9999

General Medicine Clinic Note - follow up visit

HPI:
77 yo old m with h/o HTN, CAD s/p CABG 1988. Endorses intermittent dyspnea. Righ
t eye blindness. CRI (bl 1.5-1.7). Pt has peristent gas/epigastric discomfort.
SocialHx:
lives with wife and son in the Bronx.  Requires help with all ADLs. History of t
obacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Fat
her died of unknown at 80, Mother died 92.

ALL: PCN (rash)

MEDS:
1) ASA 81mg po daily
3) Lisinopril 5mg po daily
4) Metformin 1000mg po bid
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable appearing NAD
PERRLA, anicteric sclerae, OP moist, no exudates
normal rate, irreg rhythm, no murmurs or gallops
+BS, soft, nt/nd EXT: WWP, no edema.

Labs:
- Na 142, k 4.8, Cl 107, CO2 23, BUN 20, Cr 1.6, Gluc 106, Ca 9.2
- hgba1c 6.9
- urinary microalbumin 2.2

A/P:
- pt 77 yo old man with HTN CAD s/p CABG 1988, Here for f/u.
-leave patient off lasix and Ace-I
- Continue Cozaar and HCTZ
-continue metformin 1000mg po bid
-will follow Cr
- will refer to eye clinic
- f/u 1 month
```

Structured output

- Clinical NLP pipeline output



Terminology

Patient should come back *if*
severe *facial rash* occurs.

Term Mention

span: 35-45
lexical variant: facial rash
concept id: C0239521
semantic group: disorder
body location: C0015450 (facial; 27-32)
conditional: true (if; 25-26)
negation: false (NULL)
severity: severe (severe; 28-33)
...

Common Da types

Sign/Symptom

- Alleviating Factor
- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- End Time
- Exacerbating Factor
- *Generic*
- *Negation Indicator*
- Relative Temporal Context
- Severity
- Start Time
- *Subject*
- *Uncertainty Indicator*

Procedure

- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Device
- End Date
- *Generic*
- Method
- *Negation Indicator*
- Relative Temporal Context
- Start Date
- *Subject*
- *Uncertainty Indicator*

Lab

- Abnormal
- Interpretation
- **Associated Code**
- Conditional
- Delta Flag
- Estimated flag
- *Generic*
- Lab Value
- *Negation Indicator*
- Ordinal Interpretation
- Reference Range Narrative
- *Subject*
- *Uncertainty Indicator*

Anatomical Site

- **Associated Code**
- Body Laterality
- Body Site
- Conditional
- *Generic*
- *Negation Indicator*
- *Subject*
- *Uncertainty Indicator*

Disease/Disorder

- Alleviating Factor
- Associated Sign or Symptom
- **Associated Code**
- Body Laterality
- Body Location
- Body Side
- Conditional
- Course
- Duration
- End Time
- Exacerbating Factor
- *Generic*
- *Negation Indicator*
- Relative Temporal Context
- Severity
- Start Time
- *Subject*
- *Uncertainty Indicator*

Medication

- **Associated Code**
- Change Status
- Conditional
- Dosage
- Duration
- End Date
- Form
- Frequency
- *Generic*
- *Negation Indicator*
- Route
- Start Date
- Strength
- *Subject*
- *Uncertainty Indicator*

Slide courtesy of Guergana Savova

ShARe disorder annotations

- CUI (normalization)

  "presented with facial rash"

  Facial rash (CUI C0239521)

- Negation

  "patient denies numbness"

- Subject

  "son has schizophrenia"

- Uncertainty

  "evaluation of MI"

- Course

  "The cough got worse over the next two weeks."

- Severity

  "slight bleeding"

- Conditional

  "Pt should come back if any rash occurs"

- Generic

  "she went to the HIV clinic"

- Body Location

  "patient presented with facial rash"

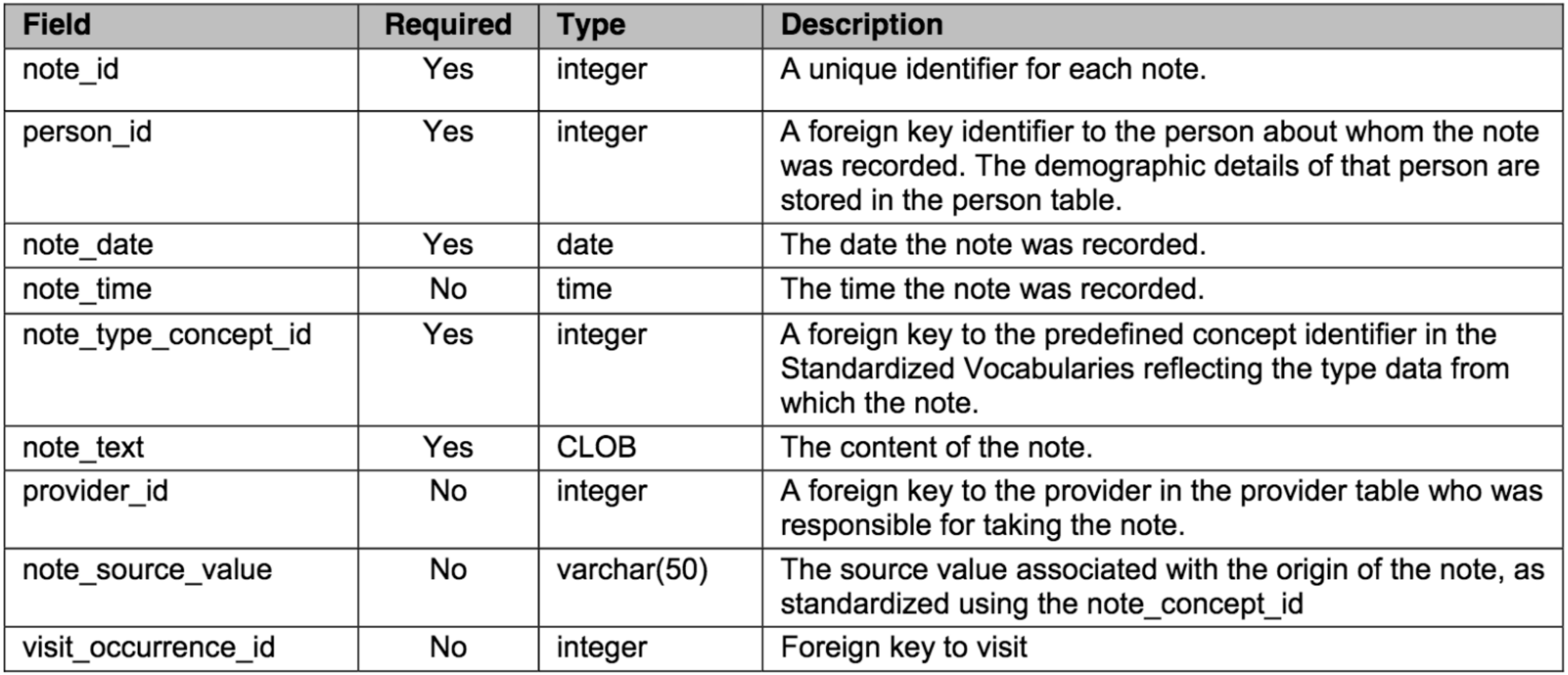
  Face (CUI: C0015450)

Elhadad et al (2015) SemEval-2015 Task 14: Analysis of Clinical Text. Proc. SemEval'15.

Proposed edits to CDM

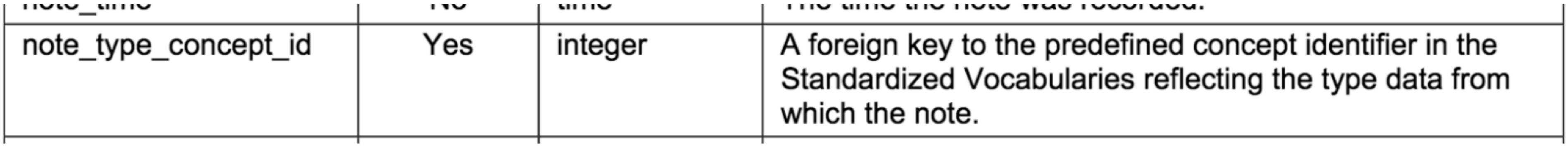
- Edits to the Note table
- New table: Note_NLP

Note table – CDM v5.0



Field	Required	Type	Description
note_id	Yes	integer	A unique identifier for each note.
person_id	Yes	integer	A foreign key identifier to the person about whom the note was recorded. The demographic details of that person are stored in the person table.
note_date	Yes	date	The date the note was recorded.
note_time	No	time	The time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.
note_text	Yes	CLOB	The content of the note.
provider_id	No	integer	A foreign key to the provider in the provider table who was responsible for taking the note.
note_source_value	No	varchar(50)	The source value associated with the origin of the note, as standardized using the note_concept_id
visit_occurrence_id	No	integer	Foreign key to visit

Note table – CDM v5.0



note_time	No	time	The time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined concept identifier in the Standardized Vocabularies reflecting the type data from which the note.

Pathology Report
Discharge Summary
Nursing Report
Outpatient Note
ED Note
Inpatient Note
Radiology
Ancillary Report
Note
Admission Note

Proposed edits to Note table

- Note_source_value:
  - extend the string to 250 chars
  - remove reference to standardized terminology
  - maybe change name to note_title_source_value or title_source_value, so that it is clear that it should be the title of the note
- Proposed 5 elements instead of note_type_concept_id and their potential values/LOINC codes

Note Table proposed edits

- Replace Note_type_concept_id with 5 elements
  - Note_role_concept_id (Role)
  - Note_domain_concept_id (Subject Matter Domain)
  - Note_setting_concept_id (Setting)
  - Note_service_concept_id (Type of Service)
  - Note_kind_concept_id (Document Kind)

Note – Role proposed

- High-level LOINC taxonomy of roles
- Filtered based on note type frequency at CUMC

Physician
Nurse
Assistant
Student
Therapist_Technician
Case Manager
Patient

Note – Domain proposed

- High-level LOINC taxonomy of subject matter domains
- Filtered based on note type frequency at CUMC

- 53 original domains or slightly filtered out?
  - Filter out Ethics, Forensic, Pastoral Care, Pharmacy?

Note – Setting proposed

- High-level LOINC taxonomy of settings

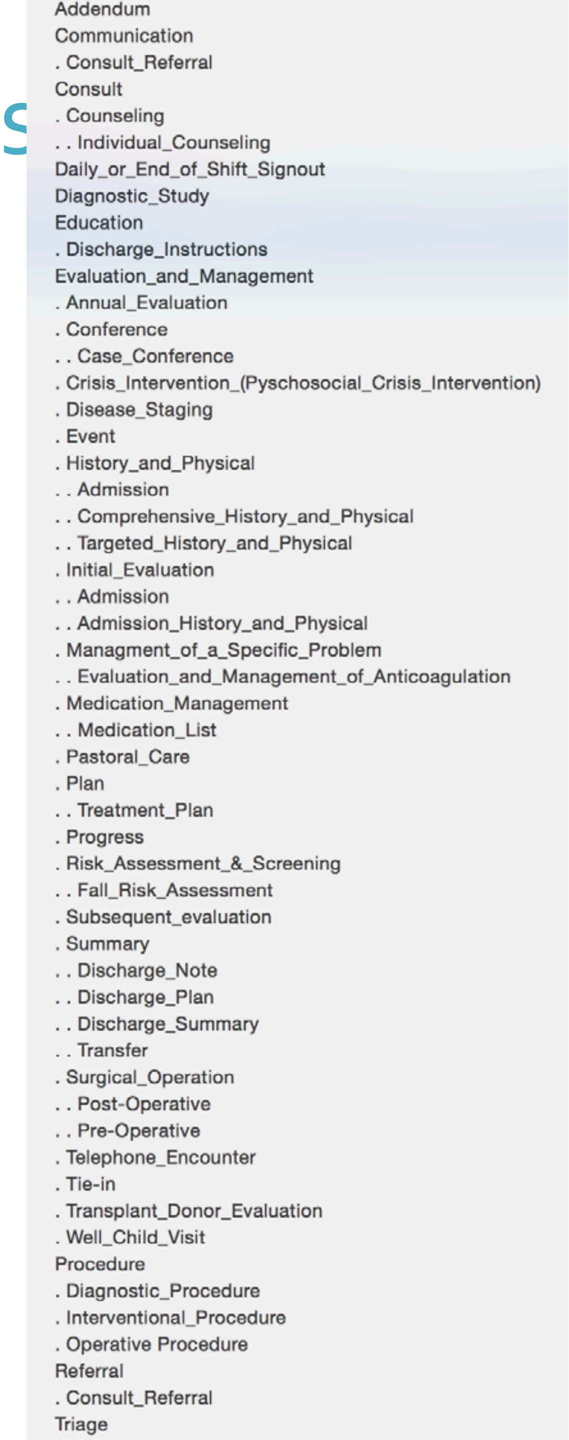
- At CUMC
  - Home
  - Inpatient
  - Outpatient
    - Rehab, ICU, ED
  - Telephone
- Propose to stick to original LOINC codes

Note – Type of Service propos

- High-level LOINC taxonomy of [type of service]

- At CUMC, modified mapping from LOINC

- Proposed: compare to at least one more institution



```
Addendum
Communication
. Consult_Referral
Consult
. Counseling
. . Individual_Counseling
Daily_or_End_of_Shift_Signout
Diagnostic_Study
Education
. Discharge_Instructions
Evaluation_and_Management
. Annual_Evaluation
. Conference
. . Case_Conference
. Crisis_Intervention_(Pyschosocial_Crisis_Intervention)
. Disease_Staging
. Event
. History_and_Physical
. . Admission
. . Comprehensive_History_and_Physical
. . Targeted_History_and_Physical
. Initial_Evaluation
. . Admission
. . Admission_History_and_Physical
. Managment_of_a_Specific_Problem
. . Evaluation_and_Management_of_Anticoagulation
. Medication_Management
. . Medication_List
. Pastoral_Care
. Plan
. . Treatment_Plan
. Progress
. Risk_Assessment_&_Screening
. . Fall_Risk_Assessment
. Subsequent_evaluation
. Summary
. . Discharge_Note
. . Discharge_Plan
. . Discharge_Summary
. . Transfer
. Surgical_Operation
. . Post-Operative
. . Pre-Operative
. Telephone_Encounter
. Tie-in
. Transplant_Donor_Evaluation
. Well_Child_Visit
Procedure
. Diagnostic_Procedure
. Interventional_Procedure
. Operative Procedure
Referral
. Consult_Referral
Triage
```

Note – Document Kind proposed

- High-level LOINC taxonomy of kind of document
- Example filtered based on CUMC note types

Note
Report
Letter
Instruction
Advanced Directive
Administrative Note

Proposed edits to CDM

- Edits to the Note table
- New table: Note_NLP

New table: Note_NLP

- New proposed table that stores output of NLP pipeline
- Note_NLP table that contains all the NLP extracted concepts, with a flexible structure wrt modifiers that can work for all types of concepts

- Keep data provenance at the concept level
- Similar to Condition_occurrence table in CDM
  - E.g. Condition_era contains more inferred information
  - Inferences about NLP outputs belong to a different table
    - Eg. "low sodium" → "hyponatremia"

Storing modifiers

- Use case: Phenotyping
- Most frequent NLP-derived queries
  - Mention of positive concept (not negated, attributed to the patient, and without any uncertainty, conditional, or general indicator)
  - Mention of negated concept
  - No mention of concept
  - Temporal mention (”history of”, “presents with”)
- Store modifiers in Note_NLP
  - Most frequent
  - Common to all semantic types

Additional table: Note_NLP

Note_NLP_id	Unique identifier for each concept extracted from NLP
note_id	Foreign key identifier to the note the concept was extracted from (Note table).
section_concept_id	Foreign key to predefined concept identifier in the Standardized Vocabularies (LOINC) reflecting the section the extracted concept belongs to.
snippet	Small window of text surrounding term mention
lexical_variant	Raw text extracted from NLP
Note_NLP_concept_id	Foreign key to concept id (Concept Table). Domain concept is provided as part of the Concept table.
NLP_system	String describing system and version used for NLP (data provenance)
NLP_date	Date describing date at which note was processed
Term_exists	Optional boolean; summary modifier that signifies presence or absence of a term for given patient (e.g., not negated, not conditional, not generic, not uncertain → termmention_ispresent=YES)
Value_as_concept_id	Optional foreign key to standard terminology (e.g., “high”); value of term
Value_as_number	Optional float; potential value of term
Unit_concept_id	Optional foreign key to unit concepts (e.g., “mg/ml”); unit of term value
Term_temporal	Optional time expression extracted associated to term, “past”, “present”

Other modifiers Note_NLP

- All other modifiers: two solutions discussed by NLP WG
  - All modifiers are stored as a string in Note_NLP
  - All modifiers are stored in a different table

Note_NLP_modifiers_id	Foreign key to term mention in Note_NLP
Modifier_concept_id	Foreign key to standard terminology (e.g., “negation_status”, “certainty”)
Value_as_concept_id	Foreign key to standard terminology (e.g., “high”)
Value_as_Number	Float Number (e.g., 30)
Unit_concept_id	Foreign key to unit concepts (e.g., “mg/ml”)

Questions / feedback / ideas…

- NLP Working group meetings
  Second Wednesday of the month, 2pm EST

- Thank you!