Phenotyping WG update

Nigam Shah



Electronic Phenotyping



Keyword queries for "noisy labeling"



tid	cui	str	Note freq	syn	Medline freq	% noun
2933	C0020255	hydrocephalus	29,634	NNS	19,541	64.61
42612	C0020255	hydrocephaly	113	NN	275	49.81
90773	C0020255	water on the brain	8	ROOT	1	50

Assumption: "long mention" is a reliable indicator of presence

Electronic Phenotyping



Error rate in labeling	Sample size
10 %	1.56 x
20 %	2.77 x
30 %	6.25 x
40 %	25 x



XPRESS- Extraction of Phenotypes from clinical **Re**cords using Silver Standards



Input: buildModel.R -- config.R, feature_vectors.Rda Output: model.Rda

APHRODITE

Automated PHenotype Routine for Observational Definition, Identification, Training and Evaluation



Phenotyping or Prediction



The source of features

person	
person_id	int
gender_concept_id	int
year_of_birth	int
month_of_birth	int
day_of_birth	int
time_of_birth	varchar(10)
race_concept_id	int
ethnicity_concept_id	int
location_id	int
provider_id	int
care_site_id	int
person_source_value	varchar(50)
gender_source_value	varchar(50)
gender_source_concept_id	int
race_source_value	varchar(50)
race_source_concept_id	int
ethnicity_source_value	varchar(50)
ethnicity_source_concept_id	int

measurement			
measurement_id	bigint unsigned		
person_id	mediumint(8) unsigned		
measurement_concept_id	int unsigned		
measurement_type_concept_id	int(10) unsigned		
measurement_date	date		
value_as_number	varchar(10)		
value_as_concept	int(10) unsigned		
range_low	varchar(10)		
range_high	varchar(10)		
unit_source_value	varchar(20)		

concept			
CONCEPT_ID	int(10) unsigned		
CONCEPT_NAME	varchar(255)		
DOMAIN_ID	varchar(255)		
VOCABULARY_ID	varchar(255)		
CONCEPT_CLASS_ID	varchar(255)		
STANDARD_CONCEPT	varchar(255)		
CONCEPT_CODE	varchar(255)		
VALID_START_DATE	varchar(255)		
VALID_END_DATE	varchar(255)		
INVALID_REASON	varchar(255)		

observation		
observation_id	bigint unsigned	
person_id	mediumint(8) unsigned	
observation_concept_id	int(10) unsigned	
observation_date	date	
observation_type_concept_id	int(8)	
observation_source_value	mediumint(8) unsigned	
qualifier_concept_id	int(8)	

drug_exposure			
bigint unsigned			
mediumint(8) unsigned			
int(10) unsigned			
date			
int(10) unsigned			
mediumint(8)			
varchar(50)			

condition_ocurrence			
condition_occurrence_id	bigint unsigned		
person_id	int(10) unsigned		
condition_concept_id	int(10) unsigned		
condition_type_concept_id	int(10) unsigned		
visit_occurrence_id	int(10) unsigned		
condition_source_value	text		
condition_start_date	date		

condition_era			
condition_era_id	bigint unsigned		
person_id	int(10) unsigned		
condition_concept_id	int(10) unsigned		
condition_era_start_date	datetime		
condition_era_end_date	datetime		
condition_occurrence_count	mediumint(8) unsigned		

drug_era			
drug_era_id	bigint unsigned		
person_id	int(10) unsigned		
drug_concept_id	int(10) unsigned		
drug_era_start_date	datetime		
drug_era_end_date	datetime		
drug_exposure_count	mediumint(8) unsigned		

relationship		
relationship_id	varchar(20)	
relationship_name	varchar(255	
is_hierarchical	varchar(1)	
defines_ancestry	varchar(1)	
reverse_relationship_id	varchar(20)	
relationship_concept_id	int	

death			
person_id	int		
death_date	date		
death_type_concept_id	int		
cause_concept_id	int		
cause_source_value	varchar(50)		
cause_source_concept_id	int		

concept_relationship		
concept_id_1	int	
concept_id_2	int	
relationship_id varchar(20)		
valid_start_date	date	
valid_end_date	date	
invalid_reason	varchar(1)	

observation_period		
observation_period_id	mediumint(8) unsigned	
person_id	mediumint(8) unsigned	
observation_period_start_date	date	
observation_period_end_date	date	
period_type_concept_id	int(8)	

note		
note_id	int unsigned	
person_id	mediumint(8) unsigne	
note_date	date	
note_concept_id	int(10) unsigned	
free_text	varchar(1)	

vocabulary			
vocabulary_id	varchar(20)		
vocabulary_name	varchar(255		
vocabulary_reference	varchar(255		
vocabulary_version	varchar(255		
vocabulary concept id	int		

concept_ancestor ancestor_concept_id

int

descendant_concept_id	int
min_levels_of_separation	int
max_levels_of_separation	int

concept_class		
concept_class_id	varchar(20	
concept_class_name	varchar(25	
concept_class_concept_id	int	

concept_synonym concept_id concept_synonym_name varchar(1000) language_concept_id int

domain

domain_id varchar(20) domain_name varchar(255) domain_concept_id int

outlier_patients pid mediumint(8) unsigned

visit_occurrence



Models built using APHRODITE

- Diabetes Mellitus
- Myocardial Infarction
- Familial Hyperlipidemia
- Celiac disease

Multi-class learning

- Diabetes Mellitus
- Myocardial Infarction

AUC	Sens.	Spec.	PPV
0.95	91 %	83 %	83 %
0.91	89 %	91 %	91 %
0.90	76.5%	93.6%	~20%
0.75	40 %	90 %	~4 %

AUC	Sens.	Spec.	PPV
0.96	52 %	99 %	99 %
0.97	94 %	94 %	90 %

Learning multiple models using neural nets

[very preliminary results – see slide notes]

- 1. Diabetes Mellitus
- 2. Myocardial Infarction
- 3. Familial Hyperlipidemia
- 4. Celiac disease
- 5. Acute liver injury
- 6. Acute renal injury
- 7. Congestive heart failure
- 8. Gastrointestinal complic.
- 9. Hepatitis C
- 10. Peripheral artery disease
- 11. Pancreatitis
- 12. Seizures

AUC	Sens.	Spec.	PPV
0.97	93 %	92 %	93 %
0.96	89 %	93 %	89 %
0.83	7 %	99 %	3 %
0.90	28 %	99 %	28 %
0.98	6 %	99 %	5 %
0.97	25 %	99 %	25 %
0.95	48 %	99 %	48 %
0.90	10 %	99 %	9 %
0.90	36 %	99 %	36 %
0.95	25 %	99 %	25 %
0.91	23 %	99 %	23 %
0.91	58 %	97 %	58 %

Discussion items

- Do we share models or the model building workflow (and retrain at each site)?
- Very few sites have data in CDM v5
- Storage of processed clinical notes is not standardized
- Incorporating 'Anchor' based learning (can help build predictive models)
- How do we use such models
 - cohort building
 - outcome ascertainment
 - clinician alerting (FIND FH example)
- How do such classifiers relate to consensus definition building?