MAYO CLINIC

Strategic Health IT Advanced Research Projects (SHARP)
Area 4: Secondary Use of EHR Data
Project 3: High-Throughput Phenotyping

21st June, 2010

Jyoti Pathak, PhD
Assistant Professor of Biomedical Informatics
Department of Health Sciences Research

Project 3: Collaborators

- CDISC (Clinical Data Interchange Standards Consortium)
- Centerphase Solutions
- IBM T.J. Watson Research Labs
- Intermountain Healthcare
- Mayo Clinic
- University of Utah

Outline

- Background
- Cohort Amplification (Susan Welch, Utah)
- Concept Frequency-based Cohort Identification (Wei-Qi Wei, Mayo)
- Commercial Viability (Jeff Tarlowe, Centerphase Solutions)
- Proposed Projects for Year 1
- Q & A & Discussion

The Big Question…

- The era of Genome-Wide Association Studies (GWAS) has arrived
  - Genotyping cost is asymptoting to free [Altman et al.]
  - Most (all?) published GWAS are done on carefully selected and uniformly characterized patient populations
  - Time consuming
- Clinical Phenotyping, on the other hand, is lacking
  - Slow-throughput
  - Costly and time consuming
- How “good” are EMRs (with inconsistencies and biases) as a source for phenotypes?

Why is this important?

- Bio-repositories are becoming popular
  - Linking biospecimens to personal health data
- Population-based studies for genetic and environmental conditions and contributions to disease etiology
  - Often limited in scope or population diversity
- Clinical trials eligibility
  - Cohort identification is always a bottleneck
- Quality metrics and HITECH Act
- Large-scale prospective cohort studies *could* be facilitated by availability of *complete*, *standardized*, and *unbiased* data from EMRs

Pros and Cons of EMR Data for Phenotyping

- We have a LOT of information about subjects
  - Demographics, labs, meds, procedures…
  - Team diagnoses as opposed to a diagnoses based on a single person's opinion
  - Potential for more reliable diagnoses
  - Identification of otherwise latent population differences

- Possible issues with using EMR data for phenotyping
  - Non-standardized, heterogeneous, unstructured data
  - Measured (e.g., demographics) vs. un-measured (e.g., socio-economic status) population differences
  - Hospital specialization and coding practices
  - Population/regional market landscape

But…the challenges can be addressed…if we

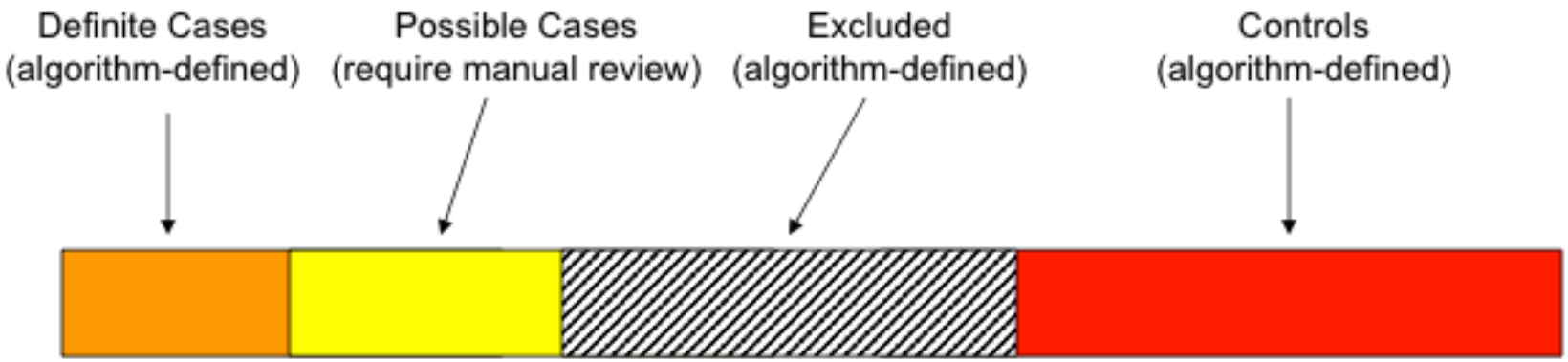
- Develop techniques for standardization and normalization of clinical data
- Develop techniques for transforming and managing unstructured clinical text into structured representations
- Develop techniques for resolving missing and inconsistent data
- Develop a scalable, robust and flexible framework for demonstrating all of the above in a “real-world setting”

SHARP Area 4 Project!

EMR-based Phenotype Algorithms

- Typical components
  - Billing and diagnoses codes
  - Procedure codes
  - Labs
  - Medications
  - Phenotype-specific co-variates (e.g., Demographics, Vitals, Smoking Status, CASI scores)
  - Pathology
  - Imaging?
- Organized into inclusion and exclusion criteria
- Experience from eMERGE (http://www.gwas.net)
  - Electronic Medical Records and Genomics Network

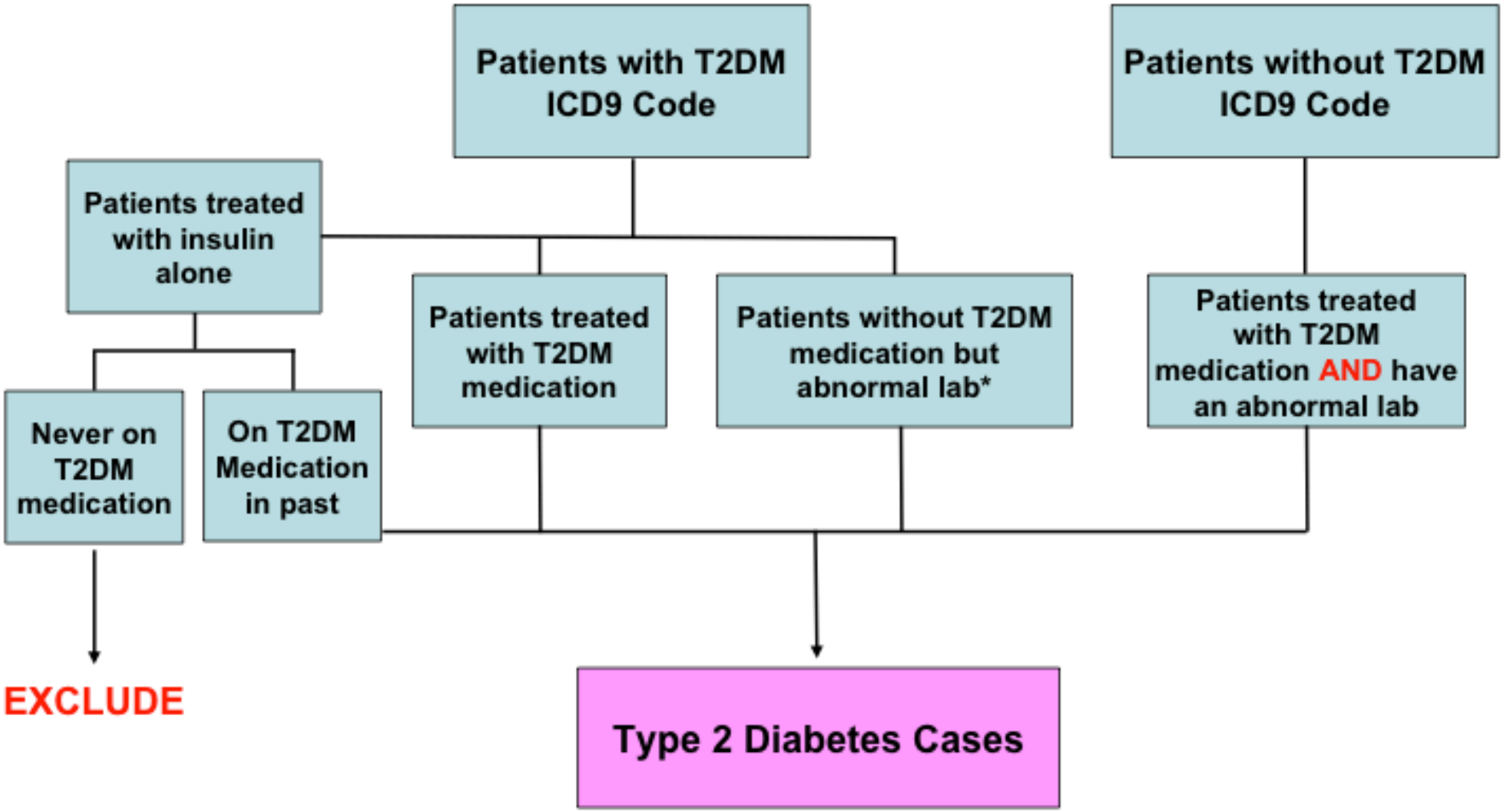
EMR-based Phenotype Algorithms



- Iteratively refine case definitions through partial manual review to achieve ~PPV ≥ 95%)

- For controls, exclude all potentially overlapping syndromes and possible matches; iteratively refine such that ~NPV ≥ 98%

Example: Type 2 Diabetes (cases)



*Random glucose > 200 mg/dl, Fasting glucose > 125 mg/dl, hemoglobin A1c > 6.5%

[eMERGE Consortium: Northwestern University]

ICD-9-CM codes for Type 2 Diabetes

Description	ICD9 Code
Diabetes with other coma	250.30
	250.32
Diabetes with hyperosmolarity	250.20
	250.22
Diabetes with unspecified complication	250.90
	250.92
Diabetes with other unspecified manifestation	250.80
	250.82
Diabetes with peripheral circulatory disorder	250.70
	250.72
Diabetes with neurological manifestations	250.60
	250.62
Diabetes with ophthalamic manifestations	250.50
	250.52
Diabetes with renal manifestations	250.40
	250.42
Diabetes without mention of complication	250.00
	250.02

[eMERGE Consortium: Northwestern University]

Prescribed Medications for Type 2 Diabetes

Drug class	Brand name
Sulfonylurea	Diabinese
Sulfonylurea	Glucotrol
Sulfonylurea	Glucotrol XL
Sulfonylurea	Micronase
Sulfonylurea	Glynase
Sulfonylurea	Diabeta
Sulfonylurea	Amaryl
Meglitinide	Prandin
Meglitinide	Starlix
Biguanide	Glucophage
Thiazolidinedione	Avandia
Thiazolidinedione	ACTOS
Alpha-glucosidase inhibitor	Precose
Alpha-glucosidase inhibitor	Glyset
DPPIV inhibitor	Januvia
GLP-1 analogue	Byetta

[eMERGE Consortium: Northwestern University]

Example: Type 2 Diabetes (Controls)

- Have not been assigned ICD-9 codes for diabetes or diabetes-related condition
- Not prescribed insulin, pramlintide, or any diabetic medications or supplies
- Has a reported glucose and it is <110 mg/dl
- No reported hemoglobin A1C ≥ 6.0%
- No reported family history of T2D

Challenges

- Algorithm design
  - Non-trivial; requires significant expert involvement
  - Highly iterative process
  - Time-consuming manual chart reviews
  - Representation of “phenotypic logic”
- Data access and representation
  - Lack of unified vocabularies, data elements, and value sets
  - Questionable reliability of ICD & CPT codes (e,g., omit codes that don’t pay well, billing the wrong code since it is easier to find)
  - Natural Language Processing needs
- And many more…

Outline

- Background
- Cohort Amplification (Susan Welch, Utah)
- Concept Frequency-based Cohort Identification (Wei-Qi Wei, Mayo)
- Commercial Viability (Jeff Tarlowe, Centerphase Solutions)
- Proposed Projects for Year 1
- Q & A

Cohort Amplification: An Associative Classification Framework for Identification of Disease Cohorts in the Electronic Health Record

Susan Rea Welch, M.S.,B.S.N.

Ph.D. Candidate, Dept. of Biomedical Informatics, U. of Utah

Fellow, Institute for Health Care Delivery Research, Intermountain Healthcare

Problems investigated:

- **Identify retrospective cohorts for research using EHR data**
- **Model to support multiple diseases out-of-the-box**
- **Generalized algorithms, national scope**

The University of Utah
Biomedical Informatics

Solution space

- Standarized EHR content (CCHIT) focused
- Process-of-care data
  - Consistent across diseases
  - Follow trail of clinical care
- Associative classification methods
  - Exhaustive search for relationships of actual data to disease
  - Supervised learning from any two ‘exemplar’ classes
  - Highly correlated attributes OK
  - Missing attributes OK
  - Induced associative “rules” in terms of EHR queries
  - Exposes EHR content
    - Comparison across exemplars presented: diseases, sites
    - Downstream specialization by disease
    - Concensus or throughput to other classification methods
    - Algorithms and methods for predictive rules

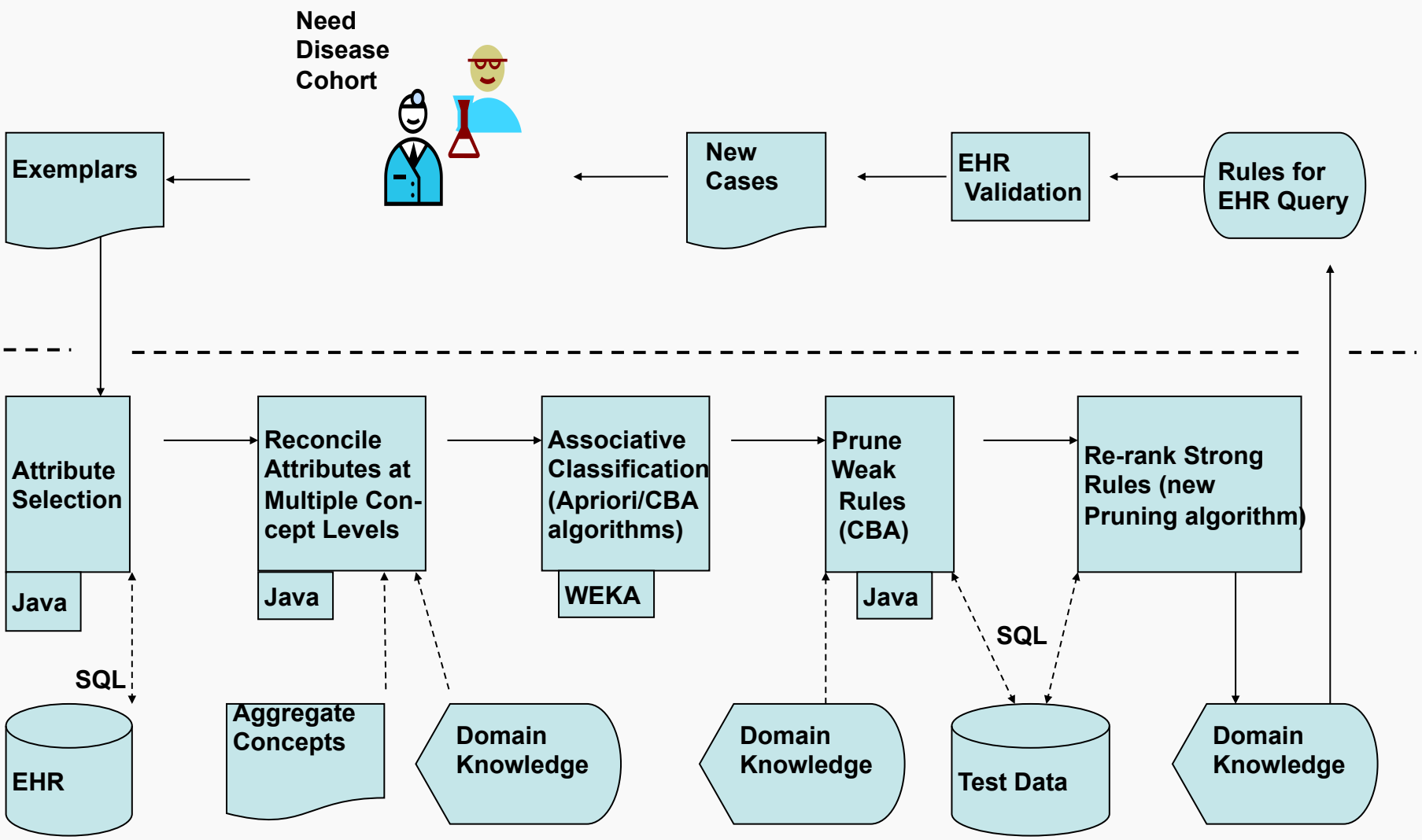
Candidate EHR Attributes

- Exhaustive search coded data for existence of code:
  - Diagnosis and procedure codes (ICD-9-CM codes)
  - Provider and ambulatory clinic procedure codes (CPT codes)
  - Provider specialty (local codes)
  - Lab observations (CPT codes)
  - Lab observations with results coded as 'Abnormal'
  - Imaging procedures (CPT codes)
  - Medication list (FirstDataBank pharm/chem, ingred)

- Plus DEFINE ANY condition: Exists or Not Exists
  - Inpatient admission, 1+ (Yes)
  - Emergency department visit, 1+ (Yes)
  - Greater than average number of ambulatory visits/year, 6+ (Yes)
  - Age > 64 (Yes)
  - Female gender (Yes)

  - HBA1C > 7 (Yes)
  - At least 1 named entity for Asthma in Clinical Documents (Yes)
  - Greater than 1 ICD-9-CM codes for Asthma (Yes)

PROCESS, COMPONENTS AND USE CASE



VALUE Added to Empiric Rules

- Quantitative evidence of frequency and "exclusivity" of known associated orders, medications, comorbidities
- Combinations of the above across data domains
- Exposes cases not meeting more common rules

Examples of Novel Induced Rules Associated with Diabetes in Order of Precedence

- Orders/tests for both HbA1C AND Urine Microalbumin
- Abnormal HbA1C by required Lab abnormal flag (CCHIT)
- Patient teaching CPT code and ICD-9-CM code for Hypertension
- Abnormal Urine Creatinine AND Abnormal Glucose professional glucometer
- Prescribed Loop Diuretics (rx hypertension if renal issues) and Order/test for Urine Microalbumin

Enhancements Envisioned

- Standardized terminologies
- Imbedded support for concept hierarchies (now manual)
- Sequential pattern association
- Consistent software platform
  - WEKA open source
  - Plug and play with other classifiers
- Improved algorithms for prediction

Opportunities in SHARP-N

- Combine machine-learning methods in sequential or consensus approaches
- Develop ontologies over rules, methods, data inputs, accuracy metrics by disease
  - Data domain coverage by methods/rules
  - Choices depending user's EHR constraints
  - Accuracy/technology trade-offs
- Develop methods for testing accuracy and generalizability of cohort identification methods
- Data quality analysis, especially for pooled cases across EHRs

Outline

- Background
- Cohort Amplification (Susan Welch, Utah)
- Concept Frequency-based Cohort Identification (Wei-Qi Wei, Mayo)
- Commercial Viability (Jeff Tarlowe, Centerphase Solutions)
- Proposed Projects for Year 1
- Q & A

MAYO CLINIC

A High Throughput Semantic Concept Frequency Based Approach for Patient Identification

Wei-Qi Wei, Cui Tao,
Guoqian Jiang, and Christopher G. Chute

Goals

- Propose a general automatic approach for patient identification or phenotype extraction.
- Investigate the benefits of involving existing SNOMED semantic knowledge in this task.

Creation of the Gold Standard

- Apply the NW algorithm on Mayo Clinic patient population
  - Potential subjects
- Cases/Controls
  - Randomly choose 1,600 T2DM cases
  - Choose 1,600 Controls matched cases by age and gender

Study Design

- Extraction of concept units
  - NLP techniques were applied on patients' clinical notes to extract SNOMED CT concept units
- Normalization of concept unit frequency

$$normalized\ frequency = \frac{\sum_{i=1}^{n} frequency\ in\ clinicl\ note\ i}{n}$$

- Machine learning algorithm
  - SVM algorithm was applied to obtain a predict model
  - Ten-fold cross validation

Results

	Precision	Recall	F-Score
Case	0.968±0.001	0.943±0.001	0.956±0.001
Control	0.945±0.001	0.969±0.001	0.957±0.001

Table 1: Performance by using all concept units as features.

Semantic Type	number of concept units	Case		
		Precision	Recall	F-Score
Disease or Syndrome	6457	0.969±0.002	0.935±0.001	0.952±0.001
Finding	4191	0.918±0.001	0.637±0.002	0.752±0.001
Body Part, Organ, or Organ Component	3644	0.730±0.003	0.580±0.003	0.646±0.002
Therapeutic or Preventive Procedure	3190	0.845±0.002	0.466±0.002	0.601±0.002
Sign or Symptom	2191	0.682±0.007	0.376±0.004	0.484±0.004
Neoplastic Process	1452	0.598±0.004	0.366±0.014	0.454±0.012
Pathologic Function	1185	0.703±0.006	0.487±0.008	0.575±0.004
Diagnostic Procedure	1114	0.584±0.002	0.813±0.008	0.680±0.003
Injury or Poisoning	1036	0.719±0.011	0.257±0.016	0.379±0.017
Body Location or Region	838	0.682±0.001	0.706±0.002	0.694±0.001
Mental or Behavioral Dysfunction	731	0.562±0.000	0.595±0.001	0.578±0.001
others	4061	0.861±0.003	0.601±0.002	0.708±0.002

Table 2: The performance of using various semantic type concept units.

Level	number of concept units	Case		
		Precision	Recall	F-Score
1	30	0.665±0.000	0.708±0.002	0.686±0.001
2	218	0.688±0.001	0.665±0.001	0.676±0.001
3	1467	0.843±0.001	0.820±0.001	0.831±0.001
4	6582	0.955±0.001	0.936±0.001	0.945±0.001
5	13116	0.958±0.001	0.938±0.001	0.948±0.001
6	21180	0.968±0.001	0.940±0.001	0.954±0.001
7	27461	0.966±0.002	0.945±0.001	0.956±0.001
8	31307	0.967±0.001	0.945±0.001	0.956±0.001
9	32785	0.967±0.001	0.944±0.001	0.956±0.001
10	33084	0.970±0.001	0.939±0.000	0.954±0.000
11	32662	0.969±0.000	0.944±0.001	0.956±0.001
12	31953	0.969±0.001	0.943±0.001	0.956±0.001

Table 3: Results of node collapse to various levels.

Next steps

- Expand the approach
  - Non-chronic diseases and old clinical notes
  - more features, e.g. lab tests and medications etc
  - Clinical notes from outside Mayo
- Other directions
  - Beyond name entities
    - Archetypes to present some phenotypes
  - Semantic relationships to help annotation and reasoning

Outline

- Background
- Cohort Amplification (Susan Welch, Utah)
- Concept Frequency-based Cohort Identification (Wei-Qi Wei, Mayo)
- Commercial Viability (Jeff Tarlowe, Centerphase Solutions)
- Proposed Projects for Year 1
- Q & A

2010 Strategic Health IT Advanced Research Projects
SHARP Face-to-Face Meeting
Project 3: High-Throughput Phenotyping

Introduction to Centerphase Solutions, Inc.

June 21, 2010

Centerphase Overview

Centerphase offers technology-enabled services to address critical needs of the biopharmaceutical industry through collaborations with premier academic medical centers and healthcare systems, starting with the Mayo Clinic

- ✓ **Clinical trial design and execution: initial focus**
- ✓ **Comparative effectiveness, post-marketing studies**
- ✓ **Safety surveillance, early detection**
- ✓ **Pharmacoeconomics, health outcomes, meaningful use**

Centerphase Team

Management

- Gary Lubin
  - Chief Executive Officer
- Jeff Tarlowe
  - Chief Operations Officer
- Beth Harper
  - Chief Clinical Officer

Experience

- More than 60 years of industry experience
- Focus on operational efficiencies in healthcare
- Expertise in starting and managing healthcare technology companies (co-founded Merck Capital Ventures)
- Passion for optimizing clinical trials performance

Board & Strategic Advisors

•Per Lofberg (Chairman), CEO Caremark

•Ken Getz, Tufts Center for Drug Development

•David Hardison, SAIC & SHARP advisor

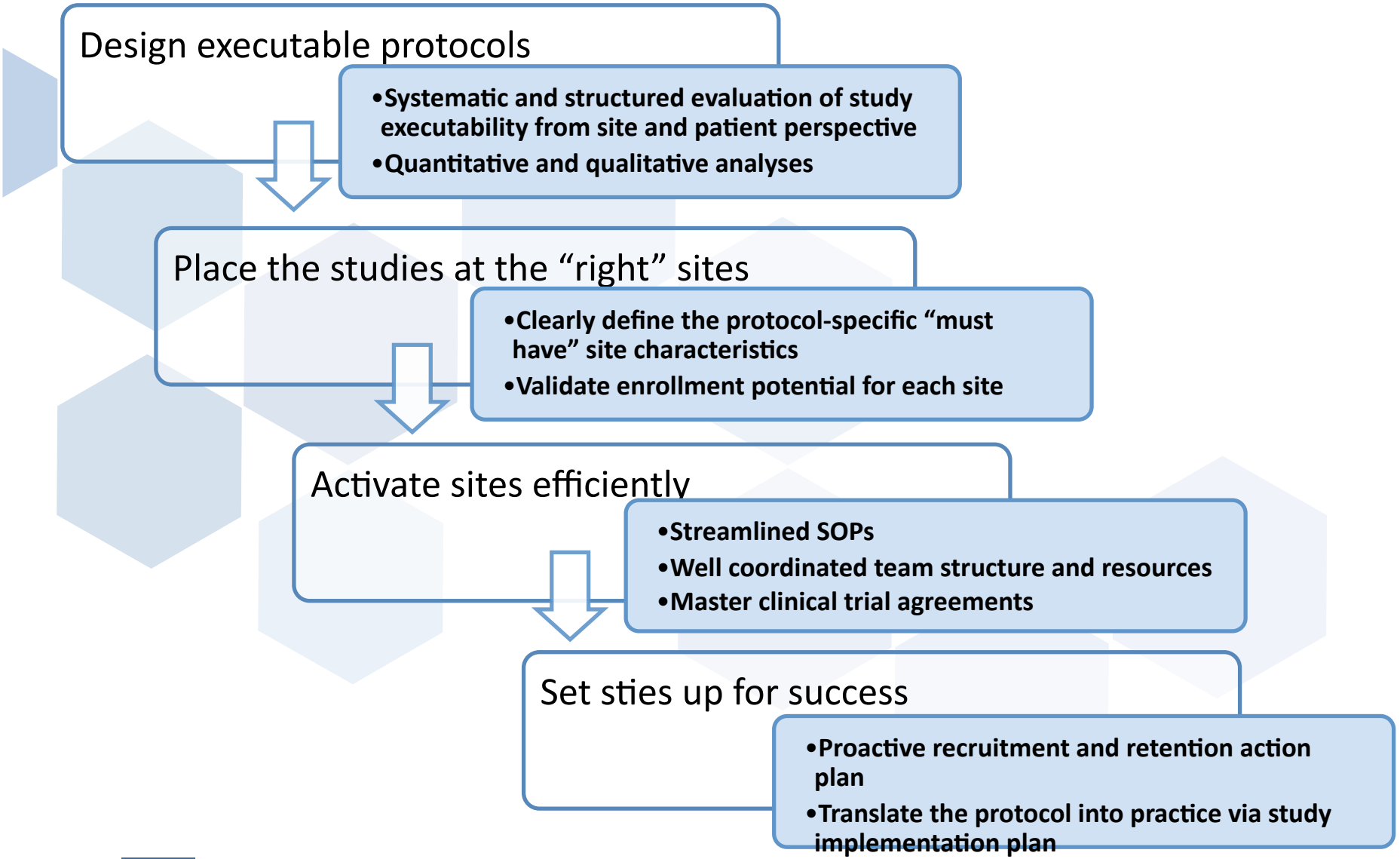
Centerphase Background

- Five years ago, Merck Capital Ventures explored creation of joint venture with a leading medical records company to develop secondary EHR applications
- Determined that market was not ready to take advantage of these applications: technology immature, data was under-scaled and deficient, and customer support was limited
- Initiated discussions with Mayo Clinic in early 2009 regarding development of market offerings around Mayo's longitudinal patient data repository
- Mayo's goals included access to improve medical care, generate new revenue streams, realize cost efficiencies and strengthen relationship with the biopharma industry
- Successfully launched company with Mayo at the beginning of 2010

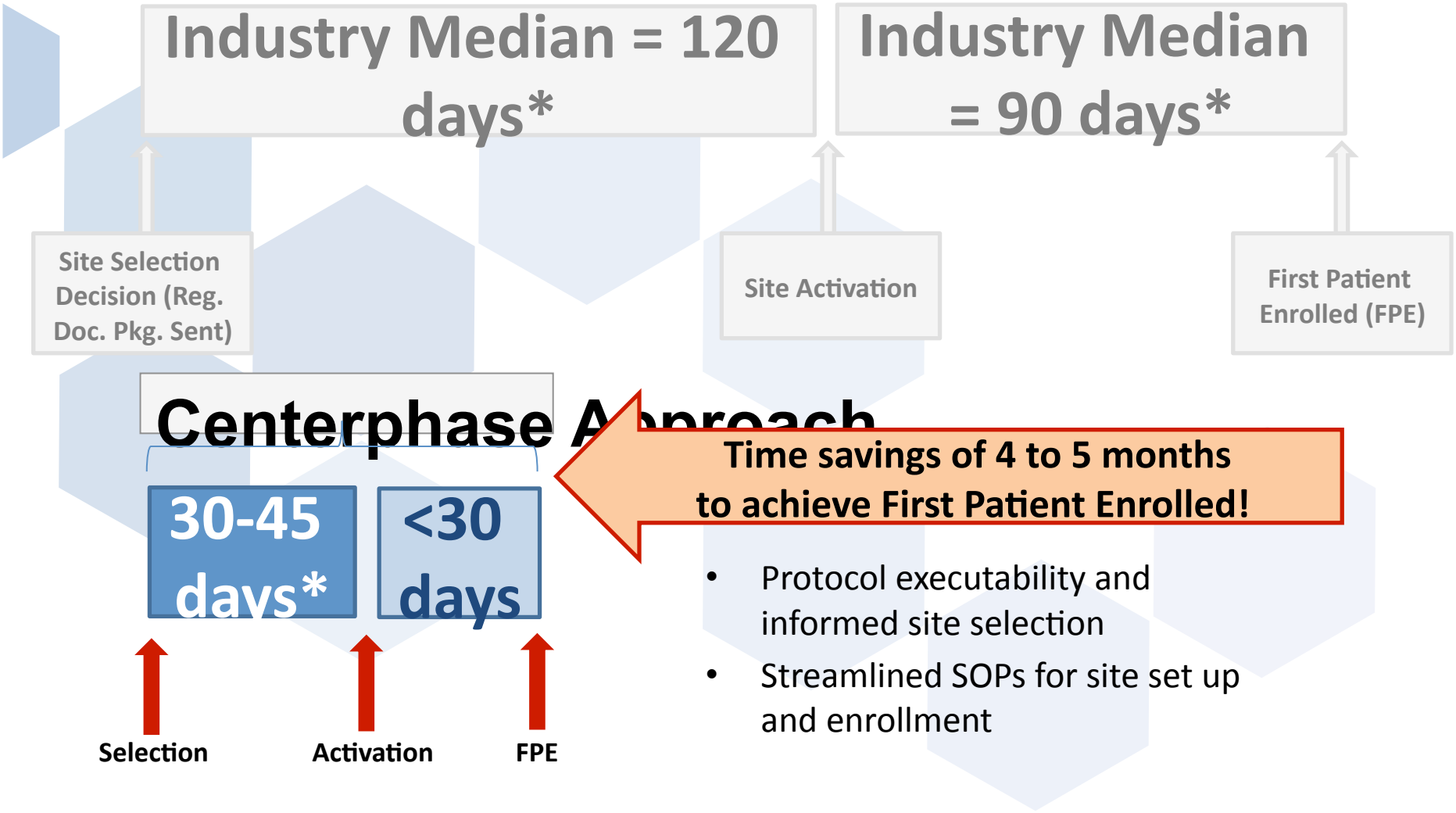
Centerphase Collaboration with Mayo

- ✓ Founding investor and clinical services provider
- ✓ Proprietary access to de-identified patient information from the electronic data trust
- ✓ Priority status given to Centerphase studies - faster IRB, contracting and budgeting review and approval processes
- ✓ Access to its network of main sites, community clinics, patients, physicians, researchers and scientists
- ✓ Dedicated staff to support Centerphase, e.g., CMO, Project Management, Legal, IRB, Finance, Biostatistics and IT

What Does Centerphase Deliver Today?
Initial Focus Is on Clinical Trials



What Does Centerphase Deliver Today?



* Varies depending on sponsor turnaround time

Centerphase's Vision

Establish hub for connecting the biopharmaceutical industry with premier health systems to perform high quality, predictable, timely clinical services

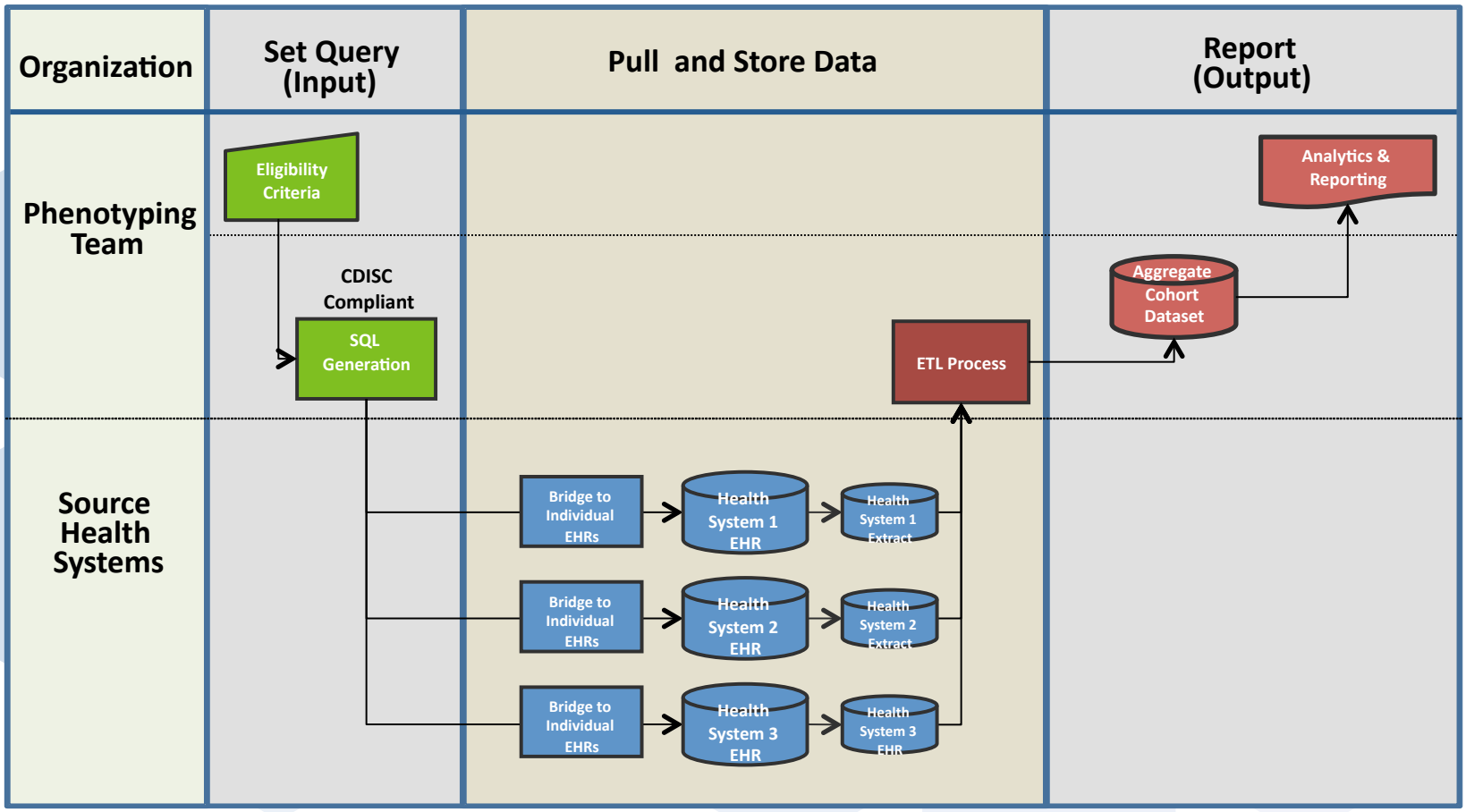
✓Centers of excellence / super-sites in key therapeutic areas – special focus on phase I / phase II studies

✓Large pool of diverse patient/candidates to recruit for trial work

✓Broad clinical capabilities to develop, evaluate and execute clinical trial

Building a worldwide network of clinical information sources and providing extensive analytical services for both prospective and retrospective comparative effectiveness and health outcome studies

High Throughput Phenotyping Team



The project team deliverables will be open-source tools for generation of CDISC compliant queries and cohort datasets.

Centerphase will serve to validate the commercial viability of these core tools in the market and supplement them with proprietary products and services (acknowledging the SHARP contributions). Health systems can decide at their option to avail themselves of these products and services through separate commercial discussions with Centerphase.

We Are Looking Forward to Working Together

Centerphase Solutions, Inc.
600 E. Crescent Avenue - Suite 205
Upper Saddle River, NJ 07458
(973) 629-3777

Gary Lubin
gary@centerphasesolutions.com:
Mobile: (973) 615-1679

Jeff Tarlowe
jeff@centerphasesolutions.com
Mobile: (201) 213-2900

Beth Harper
beth@centerphasesolutions.com:
Mobile: (817) 946-4782

Outline

- Background
- Cohort Amplification (Susan Welch, Utah)
- Concept Frequency-based Cohort Identification (Wei-Qi Wei, Mayo)
- Commercial Viability (Jeff Tarlowe, Centerphase Solutions)
- Proposed Projects for Year 1
- Q & A

Project 1: Machine Learning and Phenotyping

- EMR-derived phenotyping algorithm development is tedious, and time-consuming
  - Based on our eMERGE experience
- Research Question: To leverage machine learning methods for rule/algorithm development, and validate against expert developed ones
  - Use eMERGE library of phenotype algorithms for validation
  - Asthma and Diabetes as initial use-cases
- Preliminary work by Susan and Wei-Qi
  - Work with data normalization and NLP teams

Project 2: Phenotyping Logic Representation

- Phenotype algorithms are represented using “ad-hoc” pseudo code
  - Based on eMERGE experience
  - Mostly reside in MS Word files
- Research Question: To develop an implementation independent, phenotyping logic representation template
  - Existing work on rules languages (RuleML, RIF)
  - GroovyRules for UIMA (as an annotator)
- Leverage Centerphase’s expertise in trial eligibility criteria representation and querying

Project 3: Phenotype Data Heterogeneity

- Phenotype algorithms rely on billing data, meds etc., which *typically* have biases and are heterogeneous
- E.g., ICD-9 codes gives both false negatives/positives
  - Outpatient billing done by physicians (takes too long to find an unknown ICD-9 code)
  - Inpatient billing done by professional coders (omit codes that don't pay well)
  - Diagnoses evolve over time (initially billed/suspected diagnoses later determined to be incorrect)
- Research Question: To evaluate (and adjust for) the role of regional biases, practice, and population in phenotyping

[eMERGE Consortium: Vanderbilt University]

Additional Project Ideas

- Create a national library for clinical phenotyping algorithms
  - MS Word files do *not* scale
  - An FTP server will *not* work either
  - We need…programmatic access, querying, navigation
  - Promote re-use (where applicable)
- Probabilistic Cohort Representation
  - Binary Classification is the norm
    - "Yay"=case; "Nay"=control
  - What about the "Gray" ? (Not quite a case, but not a control either in a *stricter sense*)
    - Should we assign probabilities to such classifications (instead of "throwing away" subjects from the cohort)?

Deliverables and Timelines

- Tooling and software
  - Machine learning-based UIMA components for phenotyping
  - Rules-based template for phenotype logic representation
  - Tentatively Q3 Y1
- Manuscripts
  - Approximate ~3 (one per project)
  - Tentatively Q3-Q4 Y1
  - Approval via the publications committee
- Whitepaper and documentation (where applicable)
  - Tentatively Q4 Y1

Thank You!

Discussion Points

- Data access
  - De-identification vs. anonymous?
- What phenotypes to analyze first?
  - Asthma, T2D…(align with BEACON)
  - Other eMERGE phenotypes
- How to determine goodness of a phenotype algorithm?
  - PPV/NPV and chart reviews
  - But..some may be willing to accept more errors
- Account for flexibility, and validation re-entry points
  - ICD-9 coding conventions will change
  - New types of diagnoses introduced
- Other data/procedure sources
  - Pathology, Imaging, Spirometry, Sleep Lab….