Strategic Health IT Advanced Research Projects (SHARP)
Area 4: Secondary Use if EHR Data

Project Initiation

Thursday April 29th

PI: Christopher G Chute, MD DrPH

Agenda

- Introductions - 10 min
- Overview of Grant process & scope - 10 min
- Administrative Issues – 10 min
- Project Logistics – 15 min
- Project by Project – 40min
- Within Area 4 Integration– 10 min
- Cross-Sharp Program Integration – 10 min
- Questions – 15 min

Introductions

- Agilex Technologies
- CDISC (Clinical Data Interchange Standards Consortium)
- Centerphase Solutions
- Deloitte
- Group Health, Seattle
- IBM Watson Research Labs
- University of Utah
- Harvard University
- Intermountain Healthcare
- Mayo Clinic
- Minnesota HIE
- MIT and i2b2
- SUNY and i2b2
- University of Pittsburgh
- University of Colorado

Program Advisory Committee

Suzanne Bakken, RN DNSc, Columbia University
C. David Hardison, PhD, VP SAIC
Barbara A. Koenig, PhD, Bioethics, Mayo Clinic
Issac Kohane, MD PhD, i2b2 Director, Harvard
Marty LaVenture, PhD MPH, Minnesota Department of Health
Dan Masys, MD, Chair, Biomedical Informatics, Vanderbilt University
Mark A. Musen, MD PhD, Division Head BMIR, Stanford University
Robert A. Rizza, MD, Executive Dean for Research, Mayo Clinic
Nina Schwenk, MD, Vice Chair Board of Governors, Mayo Clinic
Kent A. Spackman, MD PhD, Chief Terminologist, IHTSDO
Tevfik Bedirhan Üstün, MD, Coordinator Classifications, WHO

SHARP Program: Background

- Funded by the Office of National Coordinator
- Support improvements in the quality, safety, and efficiency of healthcare
- Focus on solving current and future challenges that represent barriers to adoption and “meaningful use” of health IT
- Collaborative agreement between researchers, industry, healthcare providers, and other health IT stakeholders

SHARP Program: Focus Areas

- Area 1: Security of Health IT
  - Univ. of Illinois, Urbana Champaign
- Area 2: Patient-Centered Cognitive Support
  - Univ. of Texas Health Sciences Center
- Area 3: Healthcare Application & Network Platform Architectures
  - Harvard University
- Area 4: Secondary Use of EHR Data
  - Mayo Clinic
- Approx. $15 million per award (4 years)

Secondary Use of EHR Data: Research Areas

- Retrospectively and prospectively creating "in silico" cohorts of study controls
  - Approaches for the implementation of study and measures inclusion and exclusion criteria
- Methods for stratifying patients across categories of risk, demographics and care treatments
- Strategies, heuristics and methods to compensate for inconsistent and incomplete data
- Creating structured data from unstructured data using NLP to identify outcomes

04/28/2010 © 2010 Mayo Clinic 7

Secondary Use of EHR Data: Themes & Projects

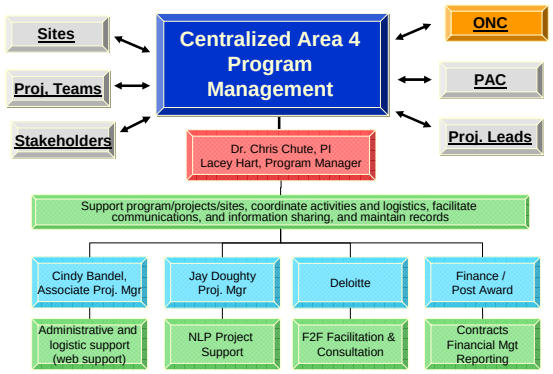
Themes	Projects	Players
Data Normalization / Phenotype Recognition / Data Quality and Evaluation Frameworks	Clinical Data Normalization	IBM, Mayo, Utah, Agilex
	Natural Language Processing (NLP)	Harvard, Group Health, IBM, Utah ,Mayo, MIT, SUNY, i2b2, Pittsburgh, Colorado
	High-Throughput Phenotyping	CDISC, Centerphase, Mayo, Utah
	UIMA and Scaling Capacity	IBM, Mayo
	Data Quality	Mayo, Utah
	Evaluation Framework	Agilex, MN HIE, Mayo, Utah

04/28/2010 © 2010 Mayo Clinic 8

Administrative Issues

- Budget Closures
  - Final Budget & Budget Justifications due TODAY to Michelle Kvall/Jeremy Eckhoff
- Contract Process
  - Starting May 3, contract officer will be in contact
- Project Management Roles
  - Mayo Clinic – serve as coordinating center & project specific task/resource management
  - Deloitte – face-to-face facilitation

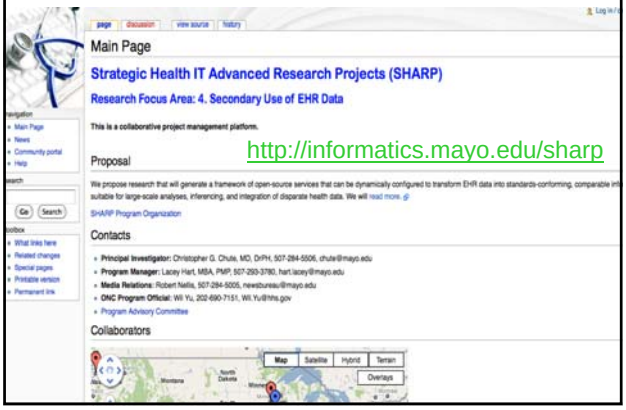
Program/Project Management



Logistics

- June 21/22 F2F
  - Technology Infrastructure
- Project Lead Calls
- Project Team Telecons
- PAC role/schedule
- Quarterly Reports
- Semi-annual Reports
- Science & Substance

Area 4: More information...



Clinical Data Normalization
Dr. Chute

Aims:

- Build generalizable data normalization pipeline
- Semantic normalization annotators involving LexEVS
- Establish a globally available resource for health terminologies and value sets
- Establish and expand modular library of normalization algorithms

Project 2: Clinical Natural Language Processing (cNLP)

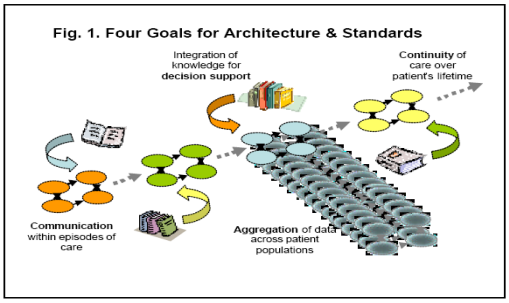
29th April, 2010

Guergana Savova, PhD

Project II: Clinical Natural Language Processing (cNLP)

- Overarching goal
  - High-throughput phenotype extraction from clinical free text based on standards and the principle of interoperability
- Focus
  - Information extraction (IE): transformation of unstructured text into structured representations
  - Merging clinical data extracted from free text with structured data

Integration of Information



National Health Information Infrastructure meeting, 2003

Data Normalization

- Informed by Project I
- University of Utah's models for episodes of care (www.clinicalelement.com)
  - Series of encounters between patient and health care system during which a problem is addressed (complaints, diagnoses, lab results, chronic medical problems, associated symptoms, physical examination findings, treatment plans).
  - Detailed clinical data for each episode

Data Normalization (cntd.)

- College of American Pathologists (CAP) cancer protocols
  - Example: colon cancer template – procedure, tumor site, size, histology, grade, tumor extension, margins, lymph nodes
- Medication profile (RxNORM)
  - Medication, dosage, route, frequency, form, strength
- Other standards: LOINC, SNOMED-CT, NDF-RT, CPT-4

Phenotyping

- Project III
  - Common grammar that can represent the formal syntax and semantics of the phenotype extraction algorithms in the form of constraint statements with appropriate boolean and logic operations
    - "operation to remove an ovary using a laser:" 83152002|oophorectomy|: 260686004|method|=257820006|laser excision-action|, where 83152002, 260686004, and 257820006 are SNOMED-CT concept identifiers.

cNLP Specific Aim 1

- Clinical concept and event discovery from the clinical narrative
  - (1) defining a set of clinical events and a set of attributes to be discovered
  - (2) identifying standards to serve as templates for attribute/value pairs
  - (3) creating a "gold standard" through the development of annotation schema, guidelines, and annotation flow, and evaluating the quality of the gold standard
  - (4) identifying relevant controlled vocabularies and ontologies for broad clinical event coverage
  - (5) methodological support for a broad array of clinical event discovery and template population
  - (6) extending Mayo Clinic's clinical Text Analysis and Knowledge Extraction System (cTAKES) information model, and implementing best-practice solutions for clinical event discovery.

cNLP Specific Aim 2

- Relation discovery among the clinical events discovered in Aim 1
  - (1) defining a set of relevant relations
  - (2) identifying standards-based information models for templated normalization
  - (3) creating a gold standard through the development of an annotation schema, guidelines, and annotation flow, and evaluating the quality of the gold standard
  - (4) developing and evaluating methods for relation discovery and template population
  - (5) implementing high-throughput scalable phenotype extraction solutions as annotators in cTAKES and UIMA-AS, either within an institution's local network or as a cloud-based deployment integrated with the institution's virtual private network.

Project II Investigators

- David Carrell, Seattle Group Health
- Wendy Chapman, University of Pittsburgh
- Peter Haug, University of Utah
- Jim Martin, University of Colorado
- Martha Palmer, University of Colorado
- Guergana Savova, Childrens Hospital Boston
- Peter Szolovits, MIT
- Wayne Ward, University of Colorado
- Ozlem Uzuner, University of Albany

Project 3: High-Throughput Phenotyping

29th April, 2010

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

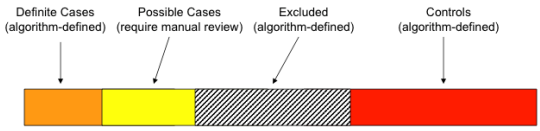
The Big Question…

- 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
- How "good" are EMRs (with inconsistencies and biases) as a source of phenotype?

EMR-based Phenotype Algorithms

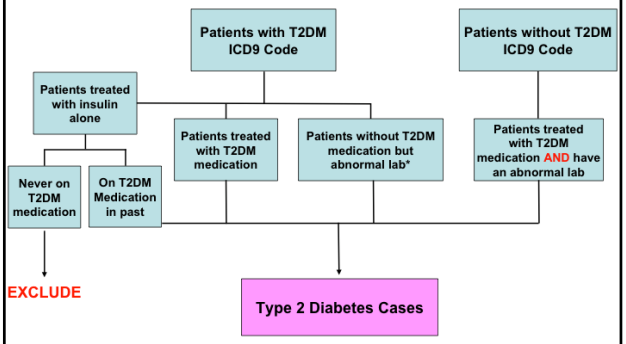
- Typical components
  - Billing and diagnoses codes
  - Procedure codes
  - Labs
  - Medications
  - Phenotype-specific co-variates (e.g., Demographics, Vitals, Smoking Status, CASI scores)
- Organized into inclusion and exclusion criteria

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%

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

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

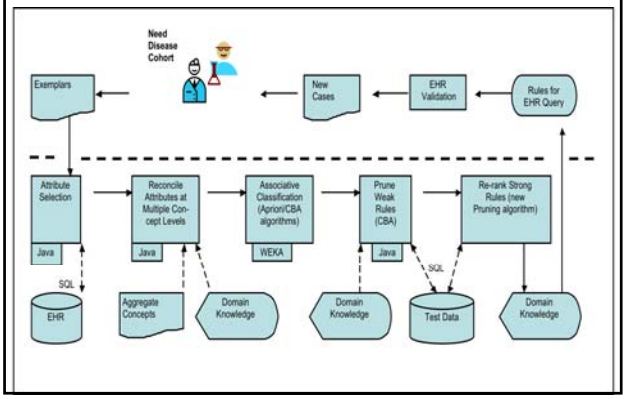
Example: Type 2 Diabetes (Controls)

- 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…

Semi-automatic Cohort Identification



Project 3: Collaborators

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

UIMA exploitation
Marshall Schor – IBM Research

- Use UIMA as a unifying framework, leveraging ecosystem
  - Work with team leads to identify "fit" (or not) of UIMA into subprojects
    - Phenotyping and Data Quality, especially
- Support UIMA and UIMA-AS use
  - Do UIMA-101 webinar or ?? for other teams
  - Consult on pipe line design / architectures / configuration
- Support scaling, capacity flexibility
  - Develop and deploy virtual machine images that can dynamically scale in cloud computing environments
  - Develop integration / deployment tooling with goal of simplicity
    - Enabling widespread adoption of POC

Data Quality
Dr. Bailey

Aims:

- Refine metrics for data consistency
- Deploy methods for missing or conflicting data resolution
- Integrate methods into UIMA pipelines
- Refine and enhance methods

Real-world evaluation framework
Dr. Huff

- We will iteratively test our normalization pipelines, including NLP where appropriate, against these normalized forms, and tabulate discordance.
  - Normalize retrospective data from the EMRs and compare it to normalized data that already exists in our data warehouses (Mayo Enterprise Data Trust).
- Use cohort identification algorithms in both EMR data and EDW data.
  - Normalize the data against CEMs.

Real-world evaluation framework

- Integrating normalization and phenotyping algorithms into HIE data flows and NHIN Connect linkages;
  - Validate data sent to or received from the UHIN network against CEM models
  - Use CEM models as the definition of payloads within NHIN Connect service calls
  - Use of NLP on document payloads that are already in use?
  - Questions
    - Data is not actually flowing in Utah yet. What is the status in Minnesota?
    - Who is communicating? Where should we try this out?
    - Is NHIN Connect in actual use in Minnesota's HIE?

Real-world evaluation framework

- Cohort identification for translational science protocols;
  - Data that is submitted to the FURTHeR database would be verified against CEM definitions.
  - Can we use NHIN Connect as the mechanism for querying data in FURTHeR? If so, we can use CEMs as the logical definition of data being addressed in the query.
  - Can we execute Cohort Amplification against the FURTHeR database?
  - Accuracy will be measured against the original EHR data
- Other questions
  - What disease cohort(s) should we use?
  - What database exists in Mayo's CTSA?
  - Timing: Is data actually flowing in Utah's & Mayo's dbs?

Area 4 Integration

- Project Lead Teleconferences
- Face to Face
- Transparent / centralized documentation
- Project management support

Cross-Sharp Program Integration

- PI Face to Face
- Yearly Jamboree w/Area Leads (rotating host)
- Potential for cross integration telecons
- Documentation transparency
  - Sharps.org – Area 1
  - Sharpc.org – Area 2
  - TBD – Area 3
  - Informatics.mayo.edu/sharp – Area 4 (sharp'n)

Questions