Patient-centered Computable Phenotyping in Health Disparities Research

Alfredo Tirado-Ramos, Ph.D.
Chief, Clinical Research Informatics Division
Department of Epidemiology and Biostatistics, School of Medicine
University of Texas Health at San Antonio, USA
Thank you CGW and AGH!

- Marian Bubak
- Mariusz Sterzel
- Karol Krawentek
Alfredo

• **University of Texas at San Antonio**
  – **Founding Director** of the Clinical Research Informatics Division, Long School of Medicine

• **Clinical and Translational Science Awards (CTSA)**
  – Informatics **Core Director**, IIMS (local)
  – Informatics Domain Task Force, (national)

• **Patient-centered Outcomes Research Institute (PCORI)**
  – **Principal Investigator** (local)
  – Obesity Task Force (national)

• **Pepper Older Americans Independence Center**
  – Clinical Informatics Core Director (local)
Our discussion’s thread

- Biomedical science and informatics
- Patient-centered Computable Phenotyping
- A BMI cluster of excellence in South Texas
- Discussion and future directions
Multi-scale complexity in BM science

From molecule to man: Decision support in individualized e-health, *Computer* 39 (11), 40-46, 2006
Current roadmap

- Big Data, Personalized Medicine, Personalized Medicine, etc… *making genomic information an integral part of clinical care*

- Current roadmap:
  - Structure of Genome
  - Biology of Genome
  - Biology of Disease
  - Medicine & Healthcare
Current roadmap

• Big Data, Personalized Medicine, Personalized Medicine, etc…
  making genomic information an integral part of clinical care

• Current roadmap:
  – Structure of Genome
  – Biology of Genome
  – Biology of Disease
  – Medicine & Healthcare

• Massive analytic databases
"The bottleneck in scientific productivity is increasingly moving from data production to the management, communication and interpretation of such data."

Big Data in Personalized Medicine

“The bottleneck in scientific productivity is increasingly moving from data production to the management, communication and interpretation of such data.”

BMI as a key component

Computer science and informatics

- System centric: information and computation using applied mathematics, engineering techniques, etc.
Computer science and informatics

- **System centric**: information and computation using applied mathematics, engineering techniques, etc.
- **Data centric**: processing, management, and retrieval of information
Relevance of informatics

• Why is informatics important in biomedical science?
  o Assess information and knowledge needs
  o Characterize, evaluate, refine processes
  o Implement, refine support systems
  o Continuous improvement
Data and information

• Quantification of information into workable knowledge
  – Messaging (e.g., HL7)

• Data as signals
  – Shannon’s work on signal processing, data compression/communication and entropy)
Data and decision support systems

- Decision support systems
  - Cognitive aspects of decision making
  - Knowledge based (inference engines) vs non-knowledge based (artificial intelligence)
  - Methodologies (modeling, data aggregation, simulation, etc.)
A computable phenotype is a machine-readable set of inclusion/exclusion criteria for a patient cohort.

In the context of EHRs, computable phenotypes reflect clinical conditions leveraging standardized medical terminology codes (e.g. ICD10, LOINC) and logical conditions (e.g. AND/OR/NOT).

Criteria must be specific enough so they can be turned into a computable query, yet generalized enough so they can be portable between different data sources.
A computable phenotype is a machine-readable set of inclusion/exclusion criteria for a patient cohort. In the context of EHRs, computable phenotypes reflect clinical conditions leveraging standardized medical terminology codes (e.g. ICD10, LOINC) and logical conditions (e.g. AND/OR/NOT). Criteria must be specific enough so they can be turned into a computable query, yet generalized enough so they can be portable between different data sources.
Models and tools

- There are different models for creating and consuming computable phenotypes, with different strengths and weaknesses, including OMOP, i2b2, SHRINE
- There are also tools to facilitate their construction, including GPC Babel, PCORnet Front Door, TriNetX
- Phenotype definition requires a multi-disciplinary team experienced with electronic medical records which can also bridge the expertise gap between developers, clinicians, nurses, governance officers and other support personnel
Models and tools

- There are different models for creating and consuming computable phenotypes, with different strengths and weaknesses, including OMOP, i2b2, SHRINE
- There are also tools to facilitate their construction, including GPC Babel, PCORnet Front Door, TriNetX
- Phenotype definition requires a multi-disciplinary team experienced with electronic medical records which can also bridge the expertise gap between developers, clinicians, nurses, governance officers and other support personnel
e.g. PCORnet’s ADAPTABLE in i2b2
A view of the ADAPTABLE phenotype in i2B2, consisting of 3 groups:

- Inclusion: Stroke, Cardiac events, diabetes, tobacco use; note that the instructions did not contain EXPLICIT details on what codes and diagnosis to include and thus a developer without training in healthcare erroneously included hemorrhagic stroke (intracerebral hemorrhage, subarachnoid hemorrhage), which would be a contradiction to the use of aspirin.
- Inclusion Procedures for heart disease or diagnosis of past procedures.
- Exclusion criteria age, bleeding, aspirin allergy, warfarin and other blood thinner use.
e.g. PCORnet’s ADAPTABLE in i2b2
e.g. GPC’s ALS cohort in i2b2
e.g. GPC’s ALS cohort in i2b2

- A view of the ALS cohort selection, including a mixture of text and ICD9 codes, a panel that looked at flowsheet data, data that may or may not be structured and is stored differently in most systems, and medications.
e.g. GPC’s ALS cohort in i2b2
e.g. PCORnet’s ADAPTABLE in Babel

<table>
<thead>
<tr>
<th>Date From: none</th>
<th>Date To: none</th>
<th>Excluded?</th>
<th>Occurs X times: &gt; 0</th>
<th>Relevance %: 100</th>
<th>Temporal Constraint: Treat Independently</th>
<th>Other Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Tests \ CHEMISTRY (KUH) \ 198-GENERAL CHEMISTRY</td>
<td>CREATININE (#2009)</td>
<td>4,014,430 facts; 324,255 patients</td>
<td>&gt; 1.5 undefined</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD10 \ E00-E69 Endocrine, nutritional and metabolic diseases (E00-E89)</td>
<td>E08-E13 Diabetes mellitus (E08-E13)</td>
<td>1,104,982 facts; 56,180 patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History \ Social History \ Tobacco Usage</td>
<td>Smoking Tobacco Use [3,356,993 facts; 490,304 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 390-459.99 DIABETES MELLITUS</td>
<td>440.2 Atherosclerosis of native arteries of the extremities [33,655 facts; 4,316 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 390-459.99 DIABETES MELLITUS</td>
<td>443.9 Peripheral vascular disease, unspecified [205,927 facts; 12,451 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 390-459.99 DIABETES MELLITUS</td>
<td>430 Subarachnoid hemorrhage [46,731 facts; 2,626 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 390-459.99 DIABETES MELLITUS</td>
<td>431 Intracerebral hemorrhage [41,965 facts; 3,995 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics \ Age</td>
<td>B = 65 years old [321,357 facts; 21,397 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 124-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS</td>
<td>Type 1 diabetes mellitus (E11) [2,180,430 facts; 75,653 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD9 \ 124-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS</td>
<td>240 Secondary diabetes mellitus [26,833 facts; 3,023 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD10 \ E00-199 Diseases of the circulatory system (E00-E59)</td>
<td>100-199 Cerebrovascular diseases (I00-I69) [223,320 facts; 24,060 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnoses ICD10 \ E00-199 Diseases of the circulatory system (E00-E59)</td>
<td>100-199 Other forms of heart disease (I00-I52)</td>
<td>150 Heart failure [362,722 facts; 17,320 patients]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
e.g. PCORnet’s ADAPTABLE in Babel

- Expansion of this phenotype extends the detailed listing out to 23 pages of inclusion and exclusion criteria, hidden from standard i2b2 view
e.g. PCORnet’s ADAPTABLE in i2b2
**e.g. PCORnet’s ADAPTABLE in Babel**

<table>
<thead>
<tr>
<th>Path</th>
<th>Concept/Term</th>
<th>Other Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory Tests \ CHEMISTRY (KUH) \ 198-GENERAL CHEMISTRY</td>
<td>CREATININE (#2009) [4,014,430 facts; 324,255 patients] &gt; 1.5 undefined</td>
<td>CT : 1.5 undefined</td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ E00-E08 Endocrine, nutritional and metabolic diseases \ E09-E89</td>
<td>E08-E13 Diabetes mellitus (E08-E13) [1,101,902 facts; 56,150 patients]</td>
<td></td>
</tr>
<tr>
<td>History \ Social History \ Tobacco Usage</td>
<td>Smoking Tobacco Use [3,366,093 facts; 490,304 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J40-44.99 Atherosclerosis</td>
<td>440.2 Atherosclerosis of native arteries of the extremities [33,655 facts; 4,316 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J44 Other peripheral vascular disease</td>
<td>443.9 Peripheral vascular disease, unspecified [205,627 facts; 12,451 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J43 Subarachnoid hemorrhage</td>
<td>430 Subarachnoid hemorrhage [46,731 facts; 2,626 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J43 Intracerebral hemorrhage</td>
<td>431 Intracerebral hemorrhage [41,665 facts; 3,995 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J43 Other and unspecified intracranial hemorrhage</td>
<td>432 Other and unspecified intracranial hemorrhage [39,064 facts; 4,094 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ J43 Occlusion and stenosis of precerebral arteries</td>
<td>433 Occlusion and stenosis of precerebral arteries [86,254 facts; 9,957 patients]</td>
<td></td>
</tr>
<tr>
<td>Diagnoses \ ICD-10 \ J30-55.99 DISEASES OF THE CIRCULATORY SYSTEM \ J40-44.99 OTHER FORMS OF HEART DISEASE</td>
<td>428 Heart failure [400,123 facts; 24,406 patients]</td>
<td></td>
</tr>
<tr>
<td>Demographics \ Age</td>
<td>&gt;= 65 years old [821,357 facts; 821,397 patients]</td>
<td></td>
</tr>
</tbody>
</table>
Timing

- Computable phenotypes can be condition-specific, design-specific, and protocol-specific.
- Different tactics may be optimal depending on whether the condition of interest is chronic, acute, or transient.
- Their successful use is sensitive to the timing of observations/measurements vs. inception of study and is often confounded by fragmentation of care and incomplete data.
Computable phenotypes can be condition-specific, design-specific, and protocol-specific.

Different tactics may be optimal depending on whether the condition of interest is chronic, acute, or transient.

Their successful use is sensitive to the timing of observations/measurements vs. inception of study and is often confounded by fragmentation of care and incomplete data.
Computable Phenotyping in Health Disparities

- South Texas has a diverse population that may reflect the future of the nation’s ethnic melting pot
- High proportion of Latinos along the Texas-Mexico border
- High incidence of diabetes, obesity, hypertension and liver disease
- Many other potential chronic conditions in this large uninsured and underinsured populations
A BMI center of excellence in South Texas

- Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  - ID key components
  - Create a cornerstone
- Create a baseline, think big
- Mind our context
- Initial focus on health disparities in Latino populations
  - Obesity and diabetes, cancer
- Develop, disseminate, translate
A BMI center of excellence in South Texas

- Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  - ID key components
  - Create a cornerstone
- Create a baseline, think big
- Mind our context
- Initial focus on health disparities in Latino populations
  - Obesity and diabetes, cancer
- Develop, disseminate, translate
Components

- Institute for the Integration of Medicine and Science
- Cancer Therapy & Research Center
- Greehey Children’s Cancer Research Institute
- Barshop Center for Longevity and Aging Studies
- Research to Advance Community Health Center
- Cameron County Hispanic Cohort
- Rio Grande health Systems
Components

• Institute for the Integration of Medicine and Science
• Cancer Therapy & Research Center
• Greehey Children’s Cancer Research Institute
• Barshop Center for Longevity and Aging Studies
• Research to Advance Community Health Center
• Cameron County Hispanic Cohort
• Rio Grande health Systems
• Cornerstone: Biomedical Informatics
A BMI center of excellence in South Texas

- Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  - ID key components
  - Create a cornerstone
- Create a baseline, think big
- Mind our context
- Initial focus on health disparities in Latino populations
  - Obesity and diabetes, cancer
- Develop, disseminate, translate
The Clinical Informatics Research Division

- Clinical informatics research and academic unit
- Created in 09/2013 with $2,000,000 SOM seed funding
- Strong research and teaching focus

Clinical Informatics Research Division (CIRD)

CIRD is a research and academic division dedicated to improving health care by bridging biomedical, statistical, and computational domains. CIRD collaborates with clinical, translational, and basic scientists who wish to make use of innovative informatics methods including our analytical data warehouse, which is custom-designed by researchers for researchers.

To discuss collaboration on grants or projects, contact informatics@uthscsa.edu.

View CIRD Fact Sheet.

Fields of Interest
- Clinical and Translational Informatics
- Multi-site Distributed Clinical Trials
- Study Cohort Selection for Clinical Trials
- Data Mining of De-identified Electronic Medical Records
- Open-source Scientific Software Development
- Agent-based Models and Simulations
- Big Data and Predictive Analytics
- Machine Learning
- Natural Language Processing
- Longitudinal, Time-to-Event, and Time-Series Data
- Next-generation Gene Sequencing Algorithms
- Stochastic Modeling of Nanoscale Biomolecular Systems

Courses Offered
Solid external funding base

- Accrual for Clinical Trials Network, NCATS, $200,000; 2017-2022
- San Antonio Claude D. Pepper Older Americans Independence Centers P30, NIH, $3,961,771; 2015-2020
- Great Plains Collaborative Clinical Data Research Network, Patient Centered Outcomes Research Institute, $8,637,161; 2015-2018
- Great Plains Collaborative Clinical Data Research Network, Patient Centered Outcomes Research Institute, $6,999,689; 2014-2015
A research informatics baseline

- i2b2-based Clinical Research Data Warehouse
- Inpatient (UT clinic) and outpatient (county hospital) data
- Standard health informatics terminologies (e.g. ICD9, ICD10, LOINC)
In more detail

- Identified data server
  - Staged source data
  - Monthly ETL refresh & de-identification
- I2b2 star schema
- De-identified data server
  - De-identified data warehouse
  - Deploy de-identified data

- Application server
  - I2b2 Web service
- Data Builder
  - Desktop
    - Web browser
  - Analytics
    - (SAS, R, Python, Excel, etc.)

Raw source data
(UTMed EPIC, UHS Sunrise, NAACCR, etc.)
Sample data elements

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Medical Terminology</th>
<th>Patients</th>
<th>Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medications</td>
<td>RXNorm, VA drug classes</td>
<td>201,973</td>
<td>8,220,939</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>ICD9, ICD10</td>
<td>318,849</td>
<td>17,772,539</td>
</tr>
<tr>
<td>Family History</td>
<td></td>
<td>156,419</td>
<td>5,969,790</td>
</tr>
<tr>
<td>Laboratory Values</td>
<td>LOINC</td>
<td>93,710</td>
<td>11,821,195</td>
</tr>
<tr>
<td>Procedures</td>
<td>CPT/HCPCS, ICD9, ICD10</td>
<td>309,988</td>
<td>3,884,754</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td>383,752</td>
<td>4,242,691</td>
</tr>
<tr>
<td>Visit Vitals</td>
<td></td>
<td>211,987</td>
<td>12,207,236</td>
</tr>
<tr>
<td>Flowsheets</td>
<td></td>
<td>213,044</td>
<td>11,656,060</td>
</tr>
<tr>
<td>NAACCR registry</td>
<td></td>
<td>14,129</td>
<td>2,703,751</td>
</tr>
<tr>
<td>Alerts</td>
<td></td>
<td>231,196</td>
<td>12,031,921</td>
</tr>
<tr>
<td>Allergies</td>
<td></td>
<td>74,611</td>
<td>144,516</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td></td>
<td>235,828</td>
<td>7,460,275</td>
</tr>
<tr>
<td>Insurance Type</td>
<td></td>
<td>383,752</td>
<td>8,148,608</td>
</tr>
<tr>
<td>Provider Specialty</td>
<td></td>
<td>383,752</td>
<td>16,241,710</td>
</tr>
<tr>
<td>Visit Type</td>
<td></td>
<td>383,752</td>
<td>8,148,608</td>
</tr>
</tbody>
</table>
Sample use case: get EMR data from multiple sources and organize into a common data model.
Think big

- IIMS Informatics
  - ACT SHRINE
- PCORI/PCORnet/GPC
  - SNOW SHRINE
- TriNetX
IIMS Informatics Core

Core B i2b2
Preparatory to Research (Self-Service)
- Counts & eligibility criteria
- Aggregate cohort characterization
- Online project requests

Physician Informaticist | Consultation | DEB/CIRD Informaticist

Health IT Teams (UT Medicine and UHS)
- Creation of EHR alerts
- EHR patient lists
- Recruitment via patient portal
- Access to patient appointments

EMR Data

Cohorts & Criteria
External Data

Core B i2b2
Data-Mining, Trials or QI (retrospective data)
Self-service: De-identified data; Locally developed apps
Grant-funded: Contact info for direct recruitment; New data sources; Custom analytic variables; Data mining/analysis in coordination with BERD

Research Networks
ACT, PCORnet, UT-HIP, TriNetX
IIMS Informatics Core

Core B i2b2
Preparatory to Research (Self-Service)
- Counts & eligibility criteria
- Aggregate cohort characterization
- Online project requests

Physician Informaticist | Consultation | DEB/CIRD Informaticist

IRB Review

Health IT Teams (UT Medicine and UHS)
- Creation of EHR alerts
- EHR patient lists
- Recruitment via patient portal
- Access to patient appointments

EMR Data

Cohorts & Criteria
External Data

Core B i2b2
Data-Mining, Trials or QI
(retrospective data)
Self-service: De-identified data;
Locally developed apps
Grant-funded: Contact info for
direct recruitment; New data
sources; Custom analytic
variables; Data mining/analysis
in coordination with BERD

Research Networks
ACT, PCORnet, UT-HIP, TriNetX
IIMS Informatics Core objectives

- Deploy an innovative and synergistic biomedical informatics program to meet investigator needs, while assuring data integrity, availability, and confidentiality
- Provide training and self-service tools to our translational science community
- Improve access to data for health services research, cohort discovery, and subject recruitment through use of our Data Warehouse and EHR systems
- Through data sharing and access to diverse data sources, serve as a national gateway to South Texas populations
PCORI’s PCORnet

- 11 Clinical Data Research Networks (CDRNs)
- 18 Patient-Powered Research Networks (PPRNs)

= A national infrastructure for patient-centered clinical research
The GPC

• The Great Plains Collaborative (GPC) grant
  – National Patient-Centered Outcomes Research Institute (PCORI), grant awarded 2013, renewed in 2015

• Research data network of networks (CDRNs)
  – 12 leading medical centers in 9 states
  – Breast Cancer, ALS and Obesity/Diabetes initial test cases

• Money, code, expertise to build an i2b2-based informatics research infrastructure
The PCORI GPC
A national collaboration network

- Inpatient and outpatient data
- Consortium-wide data governance framework
  - Paving the way to the future of data-centric research
- 2 GPC surveys in Phase 1
  - Lead in obesity
  - GPC representatives to the national PCORnet committee
- A number of PCORnet trials in Phase 2
  - ADAPTABLE, ABX, ResDac, etc
TriNetX

- Multi-channel research network used by pharma to support clinical research, trial design and the initiation of clinical trials
- A number of trials (MS – Novartis, Osteogenesis – ICON, etc)

- Multiple research networks across same platform
- Creation of virtual data marts to limit scope
- Governance to prevent statistical site and patient re-identification
- Single data mapping to enable cross site querying
- Hybrid cloud hosted architecture for easy deployment and access
TriNetX workflow

• A pharmaceutical company uses the TriNetX application to define the profile for candidate participants in their drug trial

• The Trinetx application searches the de-identified medical databases of clinical data provider organizations to locate patients that match the profile for candidate participants

• TriNetX contacts the health organizations with profile matches for the criteria

• The pharmaceutical company and health organization discuss participation in the drug trial
Different models, different strategies

• We have used our computable phenotype capability for clinical trials coming in from the TriNetX clinical trials network, several dozen PCORNet Front Door queries that leverage the Common Data Model, the Amyotrophic Lateral Sclerosis and Family Weight and Health Survey demonstration projects from our Clinical Data Research Network, the Greater Plains Collaborative, and the PCORNet Antibiotics and Childhood Obesity Project.
Standardization

• PCORNet, e.g. has identified a number of aggregate measures of data that can highlight quality issues or differential coding practices at different institutions which has resulted in improved quality within the data warehouses of participating institutions.

• Standardized computable phenotypes can enable large-scale pragmatic clinical trials across multiple health systems while ensuring reliability and reproducibility.
Use case: cancer data integration

- Through our partnership with GPC we have access to the Kansas University Medical Center (KUMC) ‘Heron’ ETL code which pulls data from Epic’s CLARITY database, standardizes it, de-identifies it, and then loads it into an i2b2 Star schema.
- We also have an on-site NAACCR Registrar who curates information to send to Texas State Cancer Registry.
- ETL code has a branch of logic that allows us to process our NAACCR information.
Use case: cancer data integration

- Once TriNetX wanted to add our cancer data do their system the process was as simple as giving them the ontology in CSV and later running tests to make sure that the counts were as expected.
- There was no legal/governance issue since they already had permission to access our i2b2.
Use case: cancer data integration

MUST HAVE: D48.6 Neoplasm of uncertain behavior of breast | 2,660 PATIENTS

D48.6 contains tumor registry codes: SHOW ALL

The tumor data below are available from only 1 site. Selecting any of these terms will restrict your query to that site.

<table>
<thead>
<tr>
<th>Stage at Diagnosis</th>
<th>1,080</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Stage 1 - 430</td>
<td></td>
</tr>
<tr>
<td>□ Stage 2 - 390</td>
<td></td>
</tr>
<tr>
<td>□ Stage 3 - 200</td>
<td></td>
</tr>
<tr>
<td>□ Stage 4 - 110</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Histology/Behavior</th>
<th>2,590</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ C50</td>
<td>821 Adenoc. in adenoma. poly. - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>857 Adenoc. with metaplasia - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>901 Adenocarcinofibroma - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>814 Adenocarcinoma. nos. - 50</td>
</tr>
<tr>
<td>□ C50</td>
<td>820 Adenoid cystic &amp; cribriform ca. - 60</td>
</tr>
<tr>
<td>□ C50</td>
<td>856 Adenosquamous carcinoma - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>912 Blood vessel tumors - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>825 Bronchiolo-alveolar adenoc. - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>824 Carcinoid tumor, malignant - 10</td>
</tr>
<tr>
<td>□ C50</td>
<td>801 Carcinoma, nos. - 60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cancer-specific Factors</th>
<th>1,970</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Breast - 1,970</td>
<td></td>
</tr>
<tr>
<td>□ Estrogen Receptor - 1,960</td>
<td></td>
</tr>
<tr>
<td>□ Progesterone receptor - 1,950</td>
<td></td>
</tr>
<tr>
<td>□ HER2 - 1,230</td>
<td></td>
</tr>
</tbody>
</table>
A note on Governance

• Technology creates challenges as its development and adoption frequently outpace policy.

• We ask organizations that have relied on traditional decision-making processes to move at a speed at which they are not accustomed while addressing problems never before considered.

• We work closely with stakeholders through governance and shared committee membership to inform data policy-based decision makers and employ standard practices fully vetted by CTSA, PCORI, and other relevant organizations.
Sample workflow

Fill out data request form for project.

Sign system access agreement and be current on training.

Need non-aggregate de-id data and/or need to do research?

Need identified data elements?

Queries via i2b2, de-id data is returned via project on REDCap

Proceed as per [http://research.uthscsa.edu/ocr/ApprovalProcess.shtml](http://research.uthscsa.edu/ocr/ApprovalProcess.shtml)

PLUS: justification for downloading out of REDCap (if applicable) and for identified data elements

CIRD returns aggregate or de-id data only.

Obtain approvals and credentialing as per chapter 7 of HOP

CIRD adds identified data elements to tables in REDCap project for this study

Get protocol approved in accordance with GPC IRB Authorization and Data Sharing agreements.

i2b2 only

CIRD returns aggregate data only.

Obtained approvals and credentialing as per chapter 7 of HOP

CIRD adds identified data elements to tables in REDCap project for this study

UTHSCSA employee?

At a GPC site?

PCORI via PopMedNet?

UTHSCSA collaborator?

Denied
A BMI center of excellence in South Texas

• Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  – ID key components
  – Create a cornerstone
• Create a baseline, think big
• Mind our context
• Initial focus on health disparities in Latino populations
  – Obesity and diabetes, cancer
• Develop, disseminate, translate
We are developing **a new workflow model** to do cohort-based biomedical science at UTHSCSA.

We are creating **an informatics platform** that uses this model to bridge the huge communication gap between our clinicians, statisticians and informaticians.

We are creating **the new governance infrastructure** to support this new way of doing research at UTHSCSA.
Cohort identification process

which patients? which visits?
query!

can this work?
does this make sense?
better query?

Four Basic Formats
- One visit/patient
- Constant sequence of visits per patient
- Time series
- Frequency table

research goals
appropriate statistical model, minimal sample size
output data in proper format

Researcher Statistician Informatician
Repeat as Needed
Here are the effects, confidence intervals, test statistics, and p-values. Here is what we can and cannot conclude from our analysis.
A BMI center of excellence in South Texas

- Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  - ID key components
  - Create a cornerstone
- Create a baseline, think big
- Mind our context
- **Initial focus on health disparities in Latino populations**
  - Obesity and diabetes, cancer
- Develop, disseminate, translate
A knowledge base for a South Texas Healthcare Learning System

- Building a large scale integration framework for diverse EMR data
- Linking to biorepositories and genomics facilities across UT and the Rio Grande Valley
- Initial focus on health disparities in Latino populations
DHR and the Rio Grande Valley

• One of the largest physician-owned facilities in the United States
• Located in one of the poorest areas in the country
• 530-bed general acute care hospital with over 265,000 patients annually
• One of the largest emergency rooms in Hidalgo County
• Women’s Hospital and a Level III-C Neonatal Intensive Care Unit
Lower Rio Grande Valley
Go BIG: UT Health Intelligence Platform

- UT System-wide action plan with an initial commitment ($12M), based within the UT Quantum Leap Initiative
- Governance and Architecture steering committees
- Clinical outcomes and research as first low hanging fruit
- Potentially a truly massive data baseline
Still many challenges

• The biggest challenge in applying a computable phenotype is probably in the process of terminology aligning

• Something as simple as a medication “Tylenol” should not be complicated, but acetaminophen is an ingredient in multiple drugs

• Defining whether you need to include a particular type of medication that contains a specific ingredient is important, as is a mapping tool (e.g. RXNav can pull up all of the medications and their associated codes)
Standards can make life difficult too

- Laboratory tests, diagnosis, and procedures all have similar issues, some of which can be solved through the use of international coding standards like ICD9 or ICD10 or proprietary coding schemes such as CPT.
- Code mapping also presents similar error issue (e.g. a researcher provides a phenotype with just ICD9 codes which do not always map exactly to ICD10 codes).
- An additional challenge is the difference in whether or not data is stored in the ehr and how easy the data can be accessed (e.g. is it stored in a structured format or as free text that requires translation?)
Expensive “dream teams” when moving into linking to biorepositories

• Great interest in using whole-genome information to reveal *genetic basis* of disease
  – Large number of people involved in the analysis (molecular and computational biologists, geneticists, pathologists, research nurses, IT and system support)

• Cost of these “dream teams” unlikely to follow data generation pattern
  – “The $1,000 genome, the $100,000 analysis?”, opinion piece (Genome Medicine 2010, 2:84)
The $1,000 genome, the $100,000 analysis?

Elaine R. Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller syndrome, and several with cancers of different types. Although each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets. The required expertise to ‘solve’ each case included molecular and computational biologists, geneticists, pathologists and physicians with expertise in the disease and of treatment modalities, research nurses, genetic counselors, and IT and systems support specialists, among others. While much of the attendant effort was focused on the absolute importance of obtaining the correct diagnosis, the large number of specialists was critical for the completion of the data analysis, the annotation of variants, the interpretive ‘filtering’ necessary to deduce the causative or ‘actionable’ variants, the clinical verification of these variants, and the communication of results and their ramifications to the treating physician, and ultimately to the patient. At the end of the day, although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing for these patients, one does wonder how, in the clinical translation required for this practice to become commonplace, such a ‘dream team’ of specialists would be assembled for each case. In other words, even if the cost and speed of generating sequencing data continue their precipitous decreases, the cost of ‘learn’ analysis seems unlikely to immediately follow suit. However, rather than predicting from this reasoning that widespread diagnosis by sequencing is unlikely to occur widely, it is perhaps more fruitful to predict, in my opinion, what is probably required for it to occur. I therefore offer the following as food for thought.

One source of difficulty in using resequencing approaches for diagnosis centers on the need to improve the quality and completeness of the human reference genome. In terms of quality, it is clear that the clone-based methods used to map, assign a minimal tiling path, and sequence the human reference genome did not yield a properly assembled or contiguous sequence equally across all loci. Lack of proper assembly is often due to collapsing of sequence within repetitive regions, such as segmental duplications, wherein genes can be found once, the correct clones are identified and sequenced. At some loci, the current reference contains a single nucleotide polymorphism (SNP) that occurs at the minor allele frequency rather than being the major allele. In addition, some loci cannot be represented by a single tiling path and require a multiple clone tiling path to capture all of the sequence variations. All of these deficiencies and others (such as providing a less-than-optimal alignment target for next-generation sequencing data and condensing all/sequence variations into a single variant) are necessary to properly interpret patient-derived data. Hence, although it is difficult work to perform, the ongoing efforts of the Genome Resource Consortium [1] to improve the overall completeness and correctness of the human reference genome should be enhanced.

Along these lines, although projects such as the early SNP Consortium [2], the subsequent HapMap projects [3-5], and more recently the 1,000 Genomes Project [6] have identified millions of SNPs in multiple ethnic groups, there is much more diversity to the human genome than single base differences. In some ways, the broader scope of ‘beyond SNP’ diversity of the genome across human populations remains mysterious, including common copy number polymorphisms, large insertions and deletions, and inversions. Mining the 1,000 Genomes data using methods to identify genome-wide structural variation should augments this considerably [7], with validation playing an important role, as many methods are still nascent. Lastly, devising clever ways to provide all such classes of variants as a searchable space for sequence data alignment remains a significant challenge, as does the development of sequence alignment algorithms that facilitate the analysis of structurally complex loci.
From molecule to man: Decision support in individualized e-health, *Computer* 39 (11), 40-46, 2006
Thanks for your attention!
Examples

- Multi-cellular Organisms
- Social, Epidemiological Networks
- Earthquakes
- Traffic flow
- Immune System
Examples

• Multi-cellular Organisms
• Social, Epidemiological Networks
• Earthquakes
• Traffic flow
• Immune System (HIV)
Complexity in HIV

‘Understanding the dynamics of infectious-disease transmission demands a holistic approach’

Complexity in HIV

- $10^9$ new viruses produced every day
- RT makes an error during each transcription
- Due to the high error rate, multiple mutations
Complexity in HIV

- HIV: infectious agent
- human host
- drug resistance
- antiretroviral combination therapy
- inhibition
- degradation
- toxicity
- infection
- immune response
Agent-based simulation models for HIV infection

• HIV is a unique problem to model in many ways
  o Does not fit traditional epidemiological models for disease, it’s not transmitted by air or casual contact
  o Mainly a result of human behavior (with some exceptions like mother-to-child infection and blood transfusions)
  o Infection typically occurs through behavior such as unprotected sexual intercourse or sharing intravenous drug needles

• Agent-based models are generally considered good candidates for simulating HIV transmission networks since they allow for complex behaviors of individuals
“Because sexual transmission of HIV is an activity rather than a process it may be more natural to define HIV transmission as a probability between individuals -as opposed to a specific rate of infection- as is often defined in Differential Equation models.”

Rhee (2006)
## Overview of Agent-based models

<table>
<thead>
<tr>
<th>Study</th>
<th>Population</th>
<th>Data Sources</th>
<th>Simulation Tool</th>
<th>Survey Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuveline et al. (2003)</td>
<td>Heterosexual, Eastern and</td>
<td>Unspecified</td>
<td>Unspecified</td>
<td>Introduction of migration, marriage, and divorce</td>
</tr>
<tr>
<td></td>
<td>southern Africa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teweldemedhin et al. (2005)</td>
<td>Heterosexual, South Africa</td>
<td>Department of Health, South Africa (2002); Rehle &amp; Shisana (2003);</td>
<td>JADE</td>
<td>Decision based functions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shisana &amp; Simbayi (2002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sumodhee et al. (2005)</td>
<td>MSM, Taiwan</td>
<td>Unspecified</td>
<td>Unspecified</td>
<td>Population groups based on behavior patterns, behavior</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>modification support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mei et al. (2010)</td>
<td>MSM, Amsterdam</td>
<td>Amsterdam Cohort Study (ACS); O'Madadhain et al. (2005)</td>
<td>Repast J</td>
<td>Dynamic network</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gray et al. (2003), Public health studies; Evolutionary and social psychology; Theoretical work from psychology and public health</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richardson and Grund (2012)</td>
<td>IDU, Bushwick, Brooklyn</td>
<td>Curtis et al. (1995); Friedman et al. (1997); Friedman et al. (1999); Koitiri et al. (2002)</td>
<td>NetLogo</td>
<td>Calibrated agent-based model</td>
</tr>
</tbody>
</table>

*International Journal of Agent Technologies and Systems, 5(1), 53-63*
Some initial findings

• Risk behavior is difficult to simulate using agent-based systems, thought it is a key research element

• Multi-Agent Simulation (MAS) methods seem to have been used mostly to simulate smaller populations

• Divide and conquer: the Mei et al. model leverages both Multi-Agent Simulation (MAS) and Complex Networks (CNs) to overcome both the lack of complex individual to individual interactions in the MAS, along with the lack of a complex representation of individuals in a CN
Prevention strategy simulation

- Addition of the effects of prevention strategies
- More sophisticated simulation systems
- **Combination** of the potential impact of multiple prevention strategies and/or a combination of prevention strategies + propagation patterns
## Prevention strategies

<table>
<thead>
<tr>
<th>Prevention Strategy</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behaviour change programs</td>
<td>Programs tailored to risk groups that seek to encourage individuals to adopt safer sexual behaviours. Risk groups include sex works and intravenous drug users. Multiple studies have shown behaviour change programs to be effective.</td>
</tr>
<tr>
<td>Education and awareness programs</td>
<td>Identify the impact of awareness programs on high-risk groups.</td>
</tr>
<tr>
<td>Predictors of condom use with steady and random partners</td>
<td>Identify strategies to promote condom use that increase awareness about their effectiveness against not only unwanted pregnancies but also HIV and other STDs.</td>
</tr>
<tr>
<td>Psychosocial variables such as depression</td>
<td>Identify potential impact that psychosocial variables have in the continued transmission of HIV. Rates of depression in people with HIV are as high as 60% compared to general population rates of around 10%; women with HIV are twice as likely as men to be depressed.</td>
</tr>
<tr>
<td>Linking to appropriate care and prevention services</td>
<td>Programs that provide linking and close monitoring of HIV-infected detainees to medical services during jail and after release; behaviour that sends a person into the criminal justice system, including injection drug use and commercial sex work, are the same activities that can increase the risk for HIV acquisition and creation of infection distribution networks.</td>
</tr>
<tr>
<td>Support of adherence to treatment regimens</td>
<td>Retention in care is important in promoting medication persistence, which can both improve the health of the individual and decrease transmission of HIV to others.</td>
</tr>
<tr>
<td>Environmental–structural interventions targeting sex workers</td>
<td>Implementing standard routine programs which monitor indicators on service provision, service uptake, and community activities.</td>
</tr>
<tr>
<td>Diagnosing HIV infections outside medical settings</td>
<td>Using widely available devices like OnQuick to increase access to early diagnosis and referral for treatment and prevention services in high-HIV prevalence settings, including correctional facilities.</td>
</tr>
<tr>
<td>Making HIV testing a routine part of medical care</td>
<td>Health-care providers including HIV testing as part of routine medical care on the same voluntary basis as other diagnostic and screening tests. Studies have shown that people who know their HIV status are more likely to protect themselves and others from infection.</td>
</tr>
<tr>
<td>Harm Reduction Programs</td>
<td>Programs that provide clean needles and syringes to intravenous drug users. These programs have been shown to be effective in reducing the risk of HIV transmission among injection drug users.</td>
</tr>
</tbody>
</table>
HIV infection AB-simulation

- Collaborators: Barbara Taylor, M.D. (HIV specialist, Department of Infectious Disease), Dr. Dante Suarez (social simulation specialist)

- Simulate the spread of HIV through a population allowing the population to be customized to fit desired population
State of the Art on AB simulations on HIV transmission

- Lacking
  - Heterogeneity of agent susceptibility and infectiousness \(^1,^2\)
  - MSM and MSMW\(^2,^3\)
  - Emergent Behavior \(^3,^4\)
  - Heterogeneity\(^2\) of
    - Male/Female behavior
    - Transmission mode risk
  - Testing and Treatment Effects \(^3\)
  - Limited Population Size

---


Many issues

- Looking at a subset of the problem can lead to erroneous guesses
- Small scale models can’t grasp small sub-populations
- Variations in infectivity should not be ignored
- Human behavior is hard to quantify
- Human diversity in susceptibility is the key to viral resistance and human survival
Approach

- Create a single cohesive model
- Create a scalable component-based tool
- Testing and treatment availability
- Mutable population (agent death/population growth)
- Combined MSM, MSMW, MSWO, and females on the same sexual network
- Heterogeneity in behavior, transmission risk by mode, and agent susceptibility/infectivity
Implementation

• MASON Java Class Library
  – Detach visualizations from simulation
  – Designed for large, bulk simulations
  – Utilize the portability and processing power of Java
  – MASON add-ons for future expansion
    • Fuzzy Logic
    • Geospatial analysis
Implementation
Model design

- Population dynamics
  - Starting population size
  - Population growth
  - Average age
  - Average life span
  - Age at entrance into sexual network
- Priming period to allow network to mature prior to viral release
- Initial # of infected agents
- Preferential infection of high risk agents
- Focus on risk
Model design

- Population dynamics
  - Starting population size
  - Population growth
  - Average age
  - Average life span
  - Age at entrance into sexual network
- Priming period to allow network to mature prior to viral release
- Initial # of infected agents
- Preferential infection of high risk agents
- Focus on risk
Population risk behavior

- Partner’s HIV Status
  - Partner’s Risk
- Condom Use
- Number of Partners
- Coital Acts
- Coital Longevity
Population risk behavior

• Population risk
  – Condom Use

• Heterogeneous Male Female settings
  – Coital Longevity (Relationship length)
  – Monogamy (Frequency of concurrent relationships)
  – Libido (Frequency of coital acts)

• Male Settings
  – Male % MSMO (when set to 100% females are not modeled)
  – Male % MSMW
  – MSM network (narrows the selection pool for MSM)
    • MSMW randomly select MF or MSMO network
  – Circumcised %

Agent networks
Agent networks

- Agents with unfulfilled libido have a chance to seek from their network(s)
- Monogamy is rolled before forming concurrent partnerships
- Agents seek randomly from their chosen network (M/F or MO)
- The number of seeking events is governed by unfulfilled libido over time
- Coital interactions average agent libido and coital longevity
Modeling relationship forming
Testing and treatment behavior

- Population Testing Likelihood
  - Known status behavioral changes (increased condom usage)
- Population Treatment Likelihood
- Treatment Enforced at AIDS onset (if not already started and infection is known)
- Virologic Suppression Likelihood
  - Likelihood of achieving viral suppression, use to model racial modifiers, adherence estimates, and ART therapy at modeled time period.
Agent wellness and treatment

- Agents gradually lose wellness when not on treatment.
  - Conversion to AIDs in approximately 10 years (5 - 16+ years)
  - Death from AIDS in approximately 3 years (9 mo. – 6+ years)
- Reduced wellness hinders the agent’s libido with significant hindrance in AIDS
- Treatment stops the degradation of health, however, agents may still decline and die.
- Viral suppression starts increasing agent health, gradually returning them to wellness.
- Agents on treatment can live normal lifespans.
Agent infectivity

- Increase infectiousness in Acute and AIDS stages
- Viral suppression reduces infectiousness by 96%
- Per Mode Risk based on CDC Per interaction estimates
  - Anal vs Vaginal
  - Insertive vs Receptive

Still a work in progress

- Add time dependent changes to simulation based on configuration file (testing after 10 years, treatment after 15, improved treatment results over time)
- IV Drug Use (~9% infections in the US)
- Commercial sex workers (vectors of transmission, or victims of clientele)
- Selectivity (desirability, racial and religious, behavioral)
- Advanced social dynamics (influence)
- Evolving agents (affected by age, geography, and network)
Still a work in progress

- Evolving Networks (agents join and leave networks)
- Prevention method deployment (circumcision, PrEP, and education)
- Stratified behavioral changes for newly diagnosed HIV+ agents
- Coinfection (HSV2, gonorrhea, syphilis – Increase infectivity and/or susceptibility by up to 12X)
- Host genetics, viral mutation, anti-viral resistance
- Importing patient cohort data.
Still a work in progress

- Evolving Networks (agents join and leave networks)
- Prevention method deployment (circumcision, PrEP, and education)
- Stratified behavioral changes for newly diagnosed HIV+ agents
- Coinfection (HSV2, gonorrhea, syphilis – Increase infectivity and/or susceptibility by up to 12X)
- Host genetics, viral mutation, anti-viral resistance
- Importing patient cohort data.