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

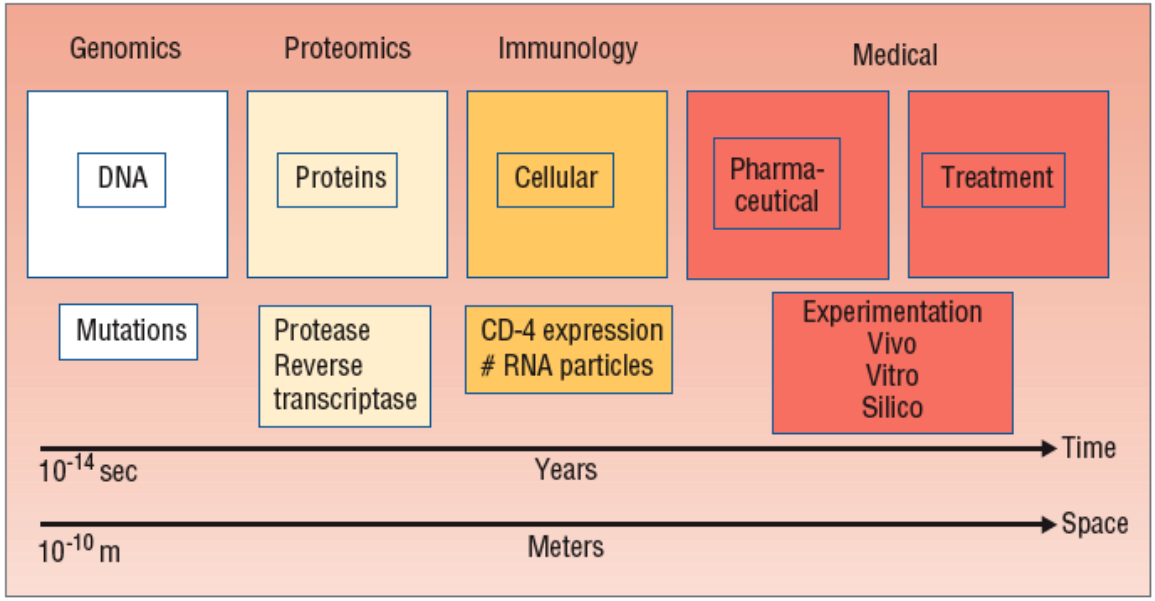
Alfredo

- **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

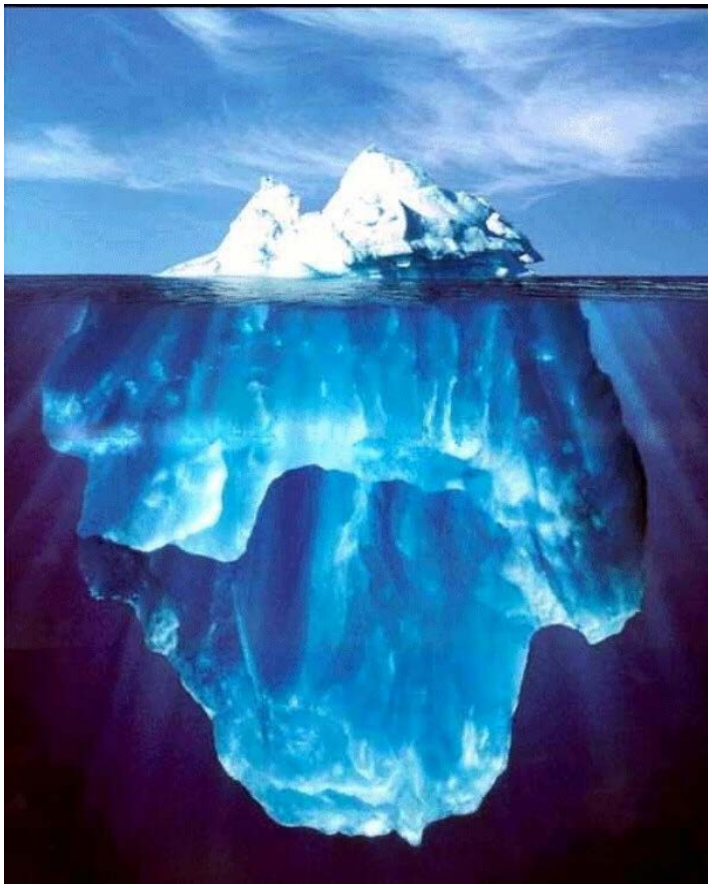
Current roadmap

- 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

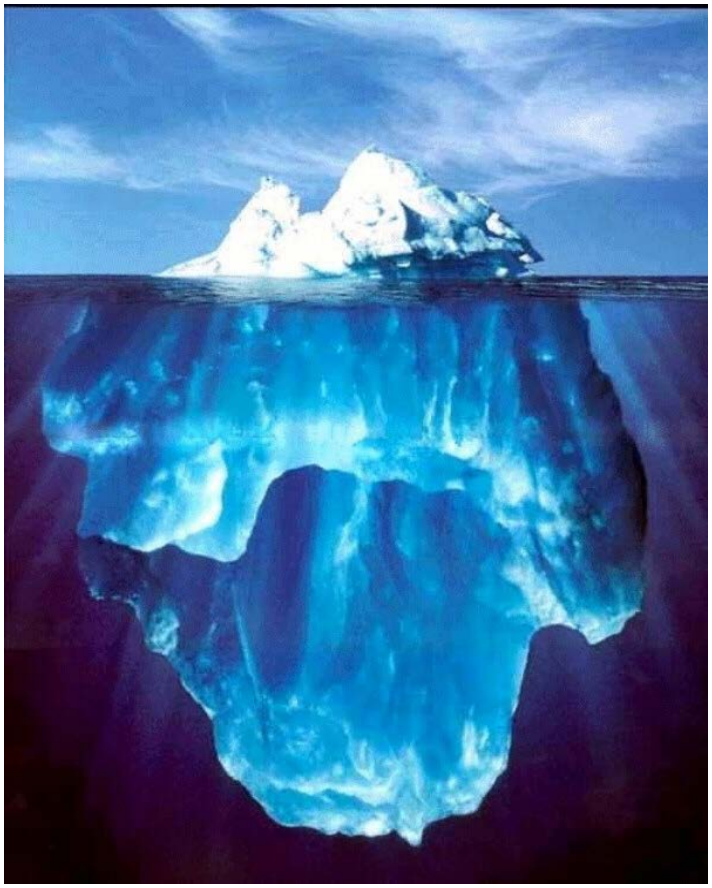
Big Data in Personalized Medicine



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

Biology: The big challenges of big data, ***Nature*** *498, 255–260, 2013*

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

Biology: The big challenges of big data, ***Nature*** *498, 255–260, 2013*

Computer science and informatics

- *System centric*: information and computation using applied mathematics, engineering techniques, etc.

Computer science and informatics

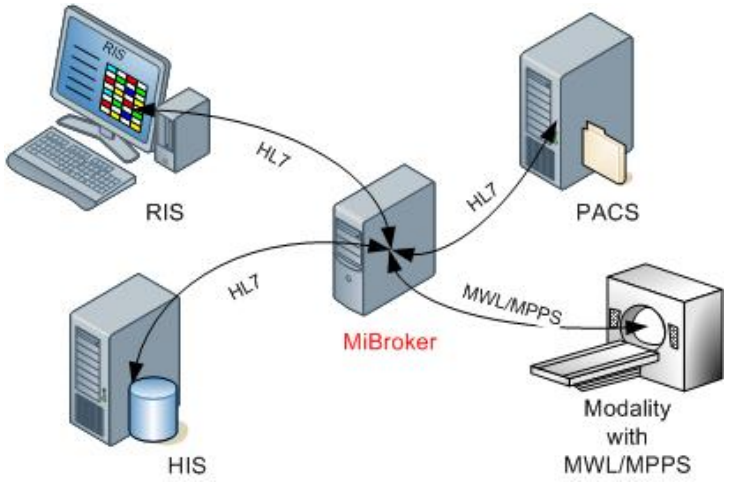
- *System centric*: information and computation using applied mathematics, engineering techniques, etc.
- *Data centric*: processing, management, and retrieval of information

Relevance of informatics

- Why is informatics important in biomedical science?
  - Assess information and knowledge needs
  - Characterize, evaluate, refine processes
  - Implement, refine support systems
  - 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)



MillenSys broker, http://www.millensys.com/

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.)

Computable Phenotyping and decision support

- 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

Computable Phenotyping and decision support

- 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

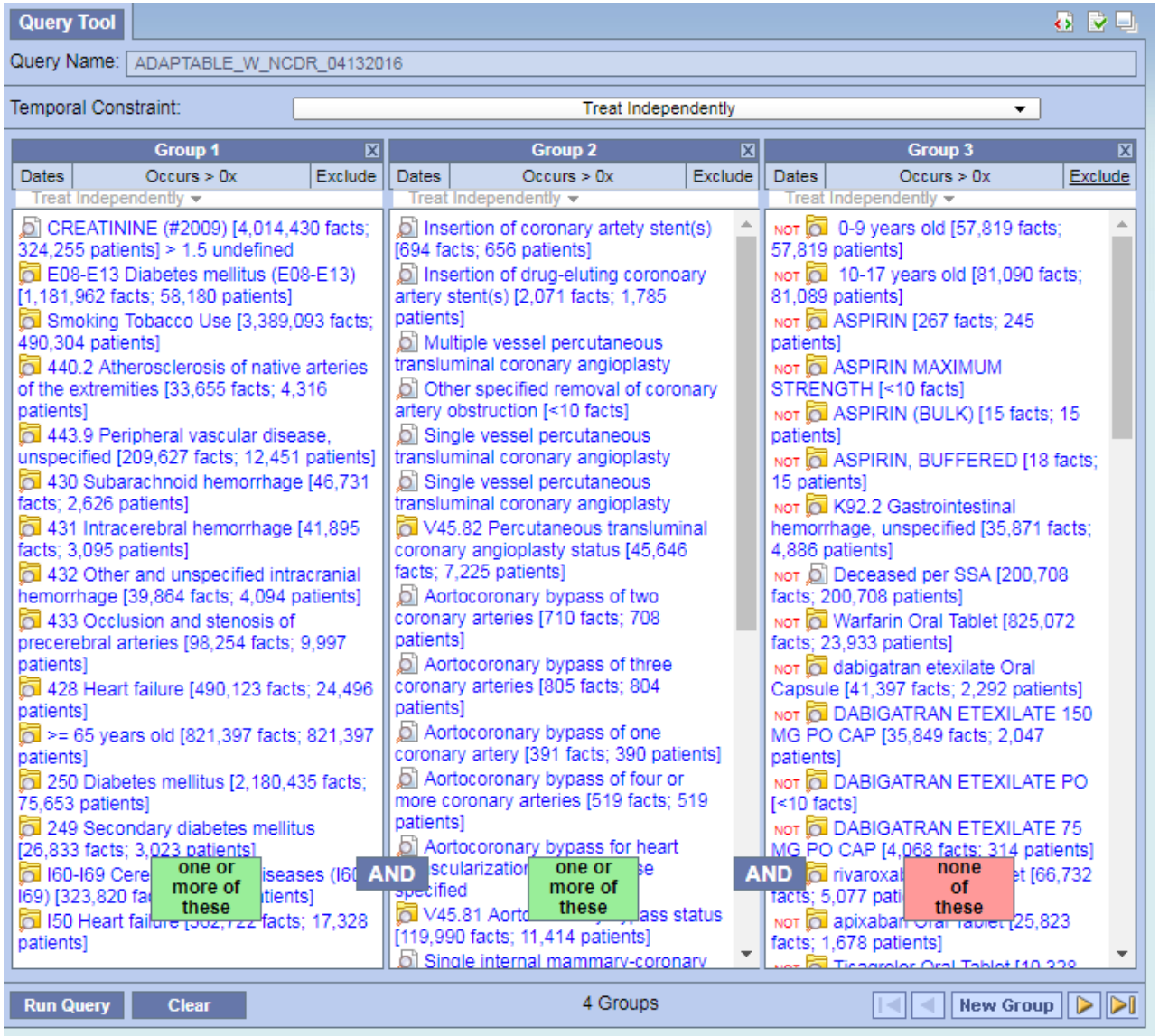
Models and tools

- 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

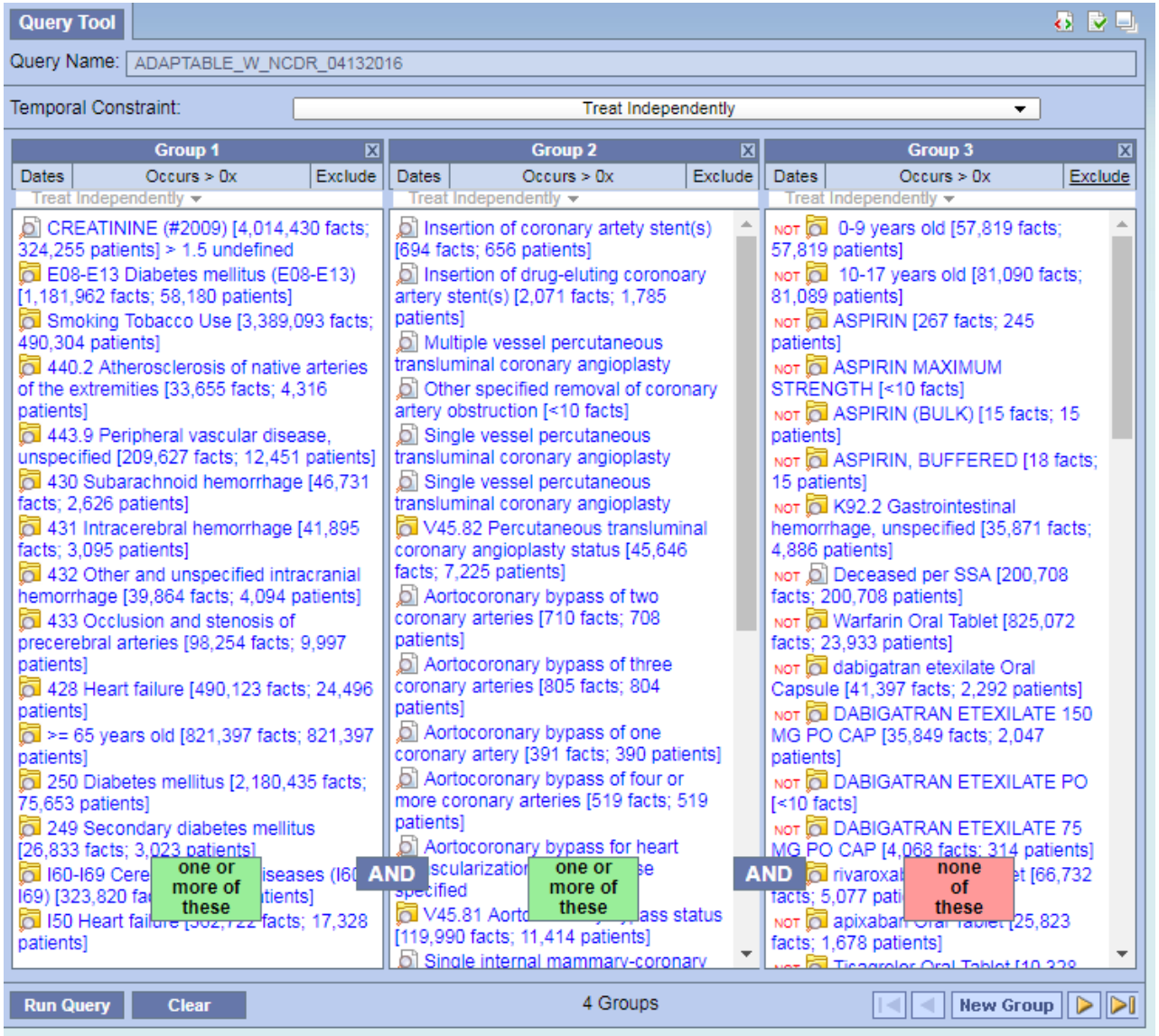
e.g. PCORnet's ADAPTABLE in i2b2



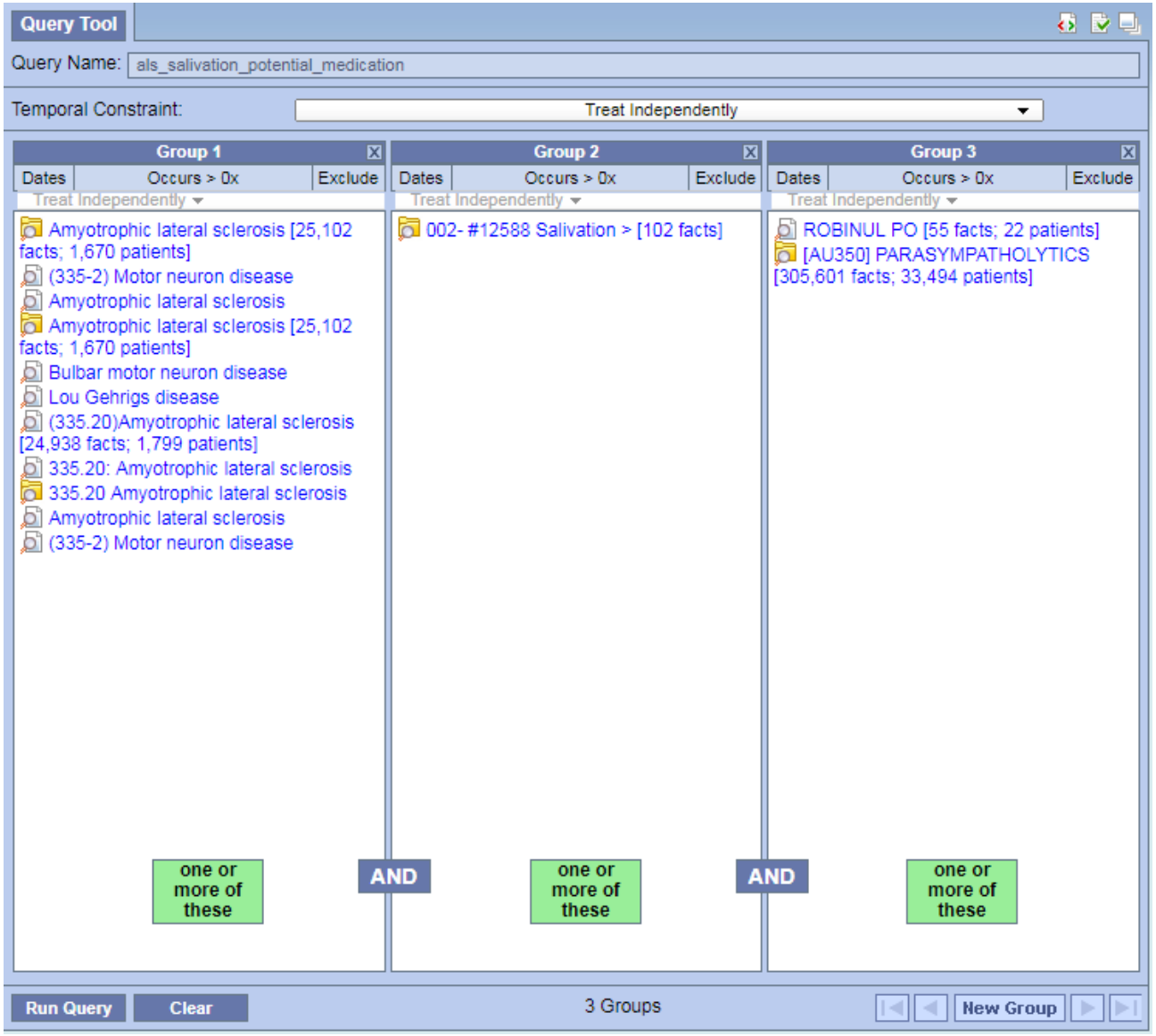
e.g. PCORnet's ADAPTABLE in i2b2

- 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



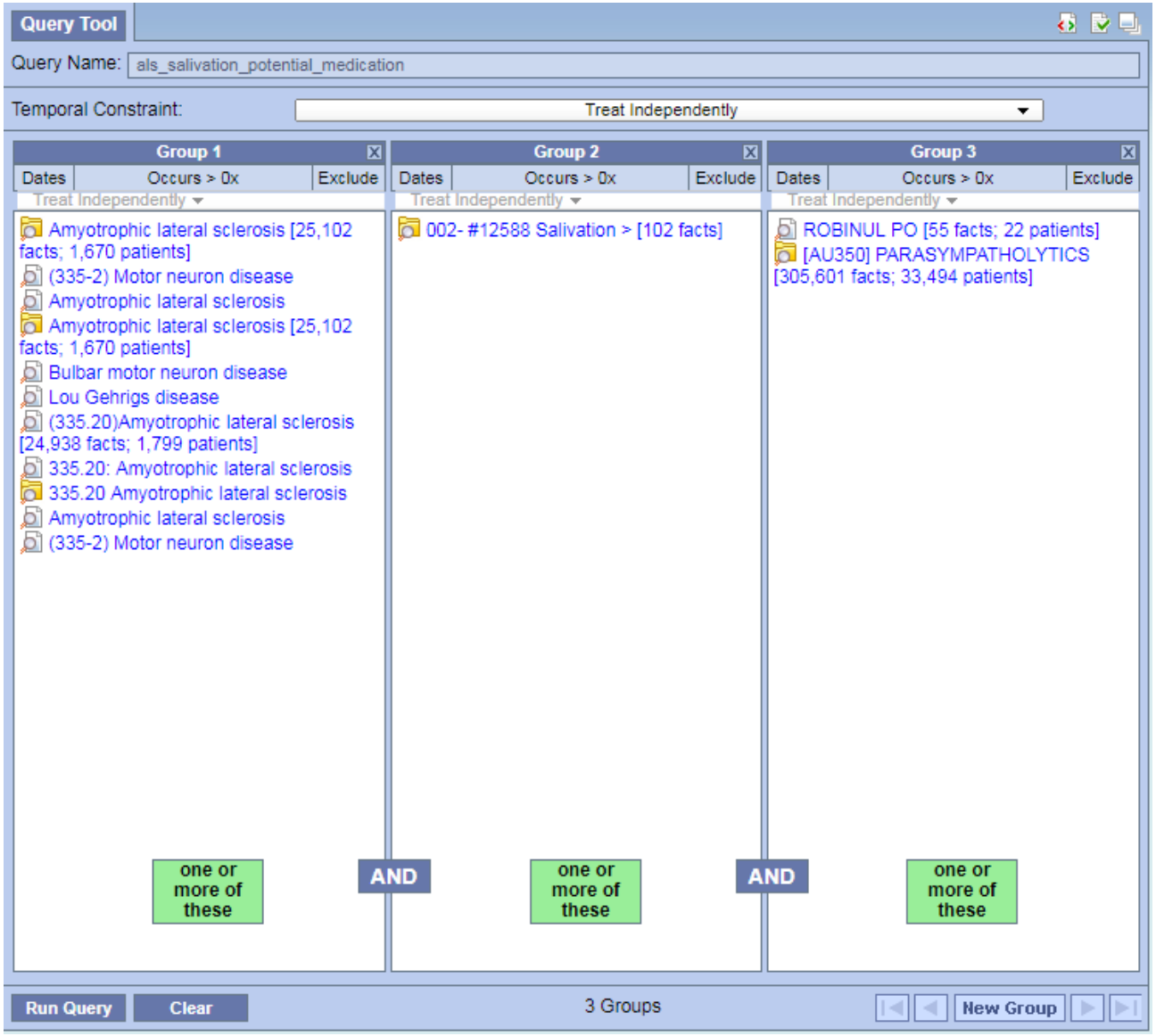
e.g. GPC's ALS cohort in i2b2



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

Group 1

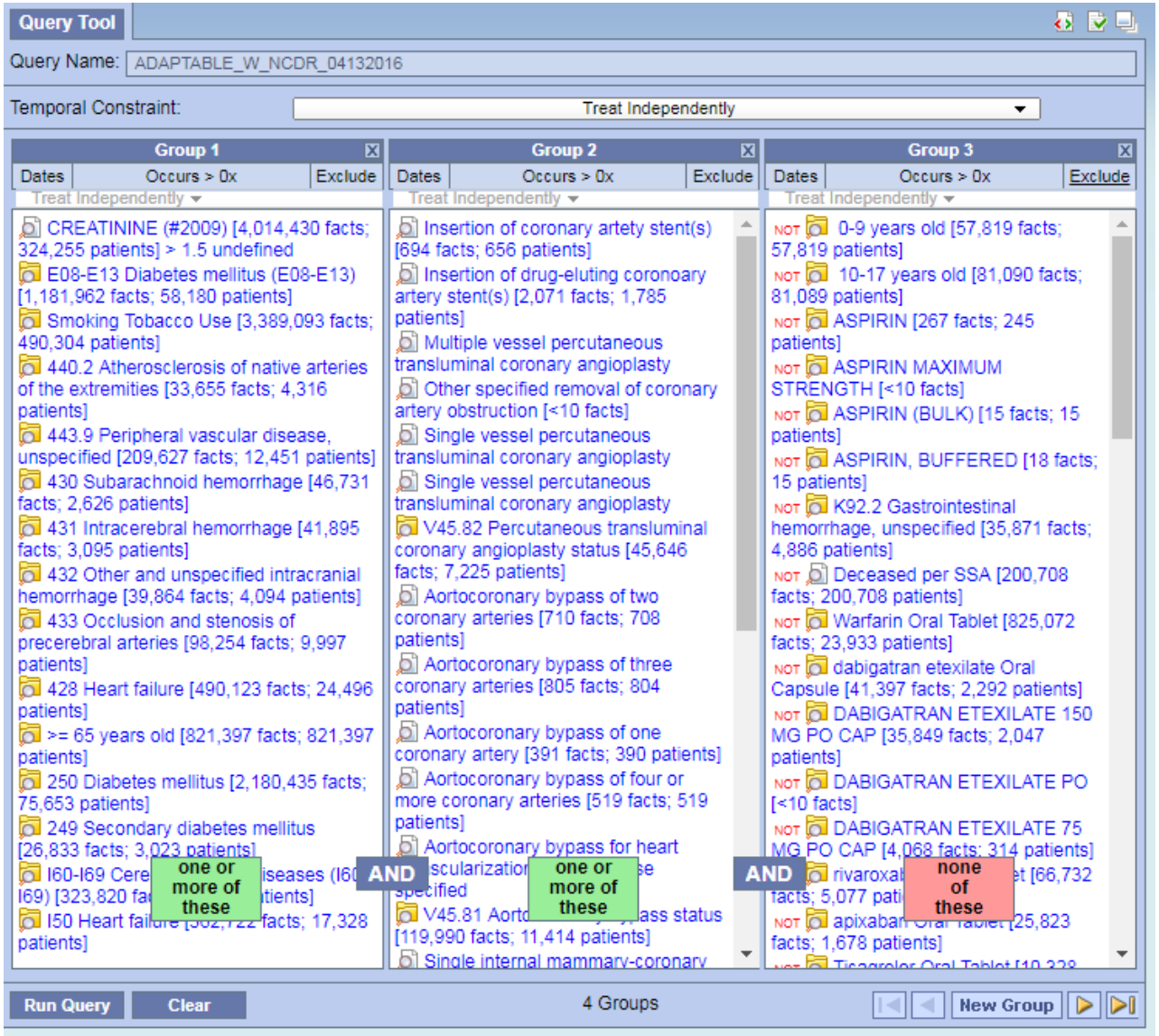
Date From: none **Date To:** none **Excluded?** false **Occurs X times:** > 0 **Relevance %:** 100 **Temporal Constraint:** Treat Independently

Path	Concept/Term	Other Information
Laboratory Tests \ CHEMISTRY (KUH) \ 198-GENERAL CHEMISTRY	CREATININE (#2009) [4,014,430 facts; 324,255 patients] > 1.5 undefined	GT : 1.5 undefined
Diagnoses \ ICD10 \ E00-E89 Endocrine, nutritional and metabolic diseases (E00-E89)	E08-E13 Diabetes mellitus (E08-E13) [1,181,962 facts; 58,180 patients]	
History \ Social History \ Tobacco Usage	Smoking Tobacco Use [3,389,093 facts; 490,304 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 440-449.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ 440 Atherosclerosis	440.2 Atherosclerosis of native arteries of the extremities [33,655 facts; 4,316 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 440-449.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ 443 Other peripheral vascular disease	443.9 Peripheral vascular disease, unspecified [209,627 facts; 12,451 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	430 Subarachnoid hemorrhage [46,731 facts; 2,626 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	431 Intracerebral hemorrhage [41,895 facts; 3,095 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	432 Other and unspecified intracranial hemorrhage [39,864 facts; 4,094 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	433 Occlusion and stenosis of precerebral arteries [98,254 facts; 9,997 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 420-429.99 OTHER FORMS OF HEART DISEASE	428 Heart failure [490,123 facts; 24,496 patients]	
Demographics \ Age	>= 65 years old [821,397 facts; 821,397 patients]	
Diagnoses \ ICD9 \ 240-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS \ 249-259.99 DISEASES OF OTHER ENDOCRINE GLANDS	250 Diabetes mellitus [2,180,435 facts; 75,653 patients]	
Diagnoses \ ICD9 \ 240-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS \ 249-259.99 DISEASES OF OTHER ENDOCRINE GLANDS	249 Secondary diabetes mellitus [26,833 facts; 3,023 patients]	
Diagnoses \ ICD10 \ I00-I99 Diseases of the circulatory system (I00-I99)	I60-I69 Cerebrovascular diseases (I60-I69) [323,820 facts; 24,686 patients]	
Diagnoses \ ICD10 \ I00-I99 Diseases of the circulatory system (I00-I99) \ I30-I52 Other forms of heart disease (I30-I52)	I50 Heart failure [362,722 facts; 17,328 patients]	

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

Group 1

Date From: none **Date To:** none **Excluded?** false **Occurs X times:** > 0 **Relevance %:** 100 **Temporal Constraint:** Treat Independently

Path	Concept/Term	Other Information
Laboratory Tests \ CHEMISTRY (KUH) \ 198-GENERAL CHEMISTRY	CREATININE (#2009) [4,014,430 facts; 324,255 patients] > 1.5 undefined	GT : 1.5 undefined
Diagnoses \ ICD10 \ E00-E89 Endocrine, nutritional and metabolic diseases (E00-E89)	E08-E13 Diabetes mellitus (E08-E13) [1,181,962 facts; 58,180 patients]	
History \ Social History \ Tobacco Usage	Smoking Tobacco Use [3,389,093 facts; 490,304 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 440-449.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ 440 Atherosclerosis	440.2 Atherosclerosis of native arteries of the extremities [33,655 facts; 4,316 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 440-449.99 DISEASES OF ARTERIES, ARTERIOLES, AND CAPILLARIES \ 443 Other peripheral vascular disease	443.9 Peripheral vascular disease, unspecified [209,627 facts; 12,451 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	430 Subarachnoid hemorrhage [46,731 facts; 2,626 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	431 Intracerebral hemorrhage [41,895 facts; 3,095 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	432 Other and unspecified intracranial hemorrhage [39,864 facts; 4,094 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 430-438.99 CEREBROVASCULAR DISEASE	433 Occlusion and stenosis of precerebral arteries [98,254 facts; 9,997 patients]	
Diagnoses \ ICD9 \ 390-459.99 DISEASES OF THE CIRCULATORY SYSTEM \ 420-429.99 OTHER FORMS OF HEART DISEASE	428 Heart failure [490,123 facts; 24,496 patients]	
Demographics \ Age	>= 65 years old [821,397 facts; 821,397 patients]	
Diagnoses \ ICD9 \ 240-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS \ 249-259.99 DISEASES OF OTHER ENDOCRINE GLANDS	250 Diabetes mellitus [2,180,435 facts; 75,653 patients]	
Diagnoses \ ICD9 \ 240-279.99 ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS \ 249-259.99 DISEASES OF OTHER ENDOCRINE GLANDS	249 Secondary diabetes mellitus [26,833 facts; 3,023 patients]	
Diagnoses \ ICD10 \ I00-I99 Diseases of the circulatory system (I00-I99)	I60-I69 Cerebrovascular diseases (I60-I69) [323,820 facts; 24,686 patients]	
Diagnoses \ ICD10 \ I00-I99 Diseases of the circulatory system (I00-I99) \ I30-I52 Other forms of heart disease (I30-I52)	I50 Heart failure [362,722 facts; 17,328 patients]	

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

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 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

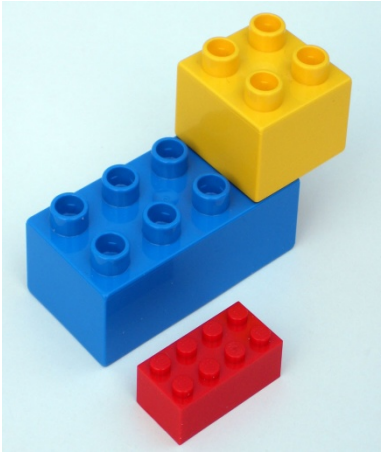
- Adapt the CDC idea of Centers of Excellence in Public Health in regards to informatics knowledge
  - ID key components
  - Create a cornerstone
- Create a baseline, think big
- Mind our context
- Initial focus on health disparities in Latino populations
  - Obesity and diabetes, cancer
- Develop, disseminate,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

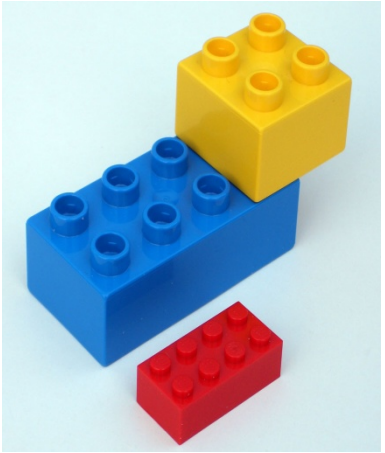
Components

- 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

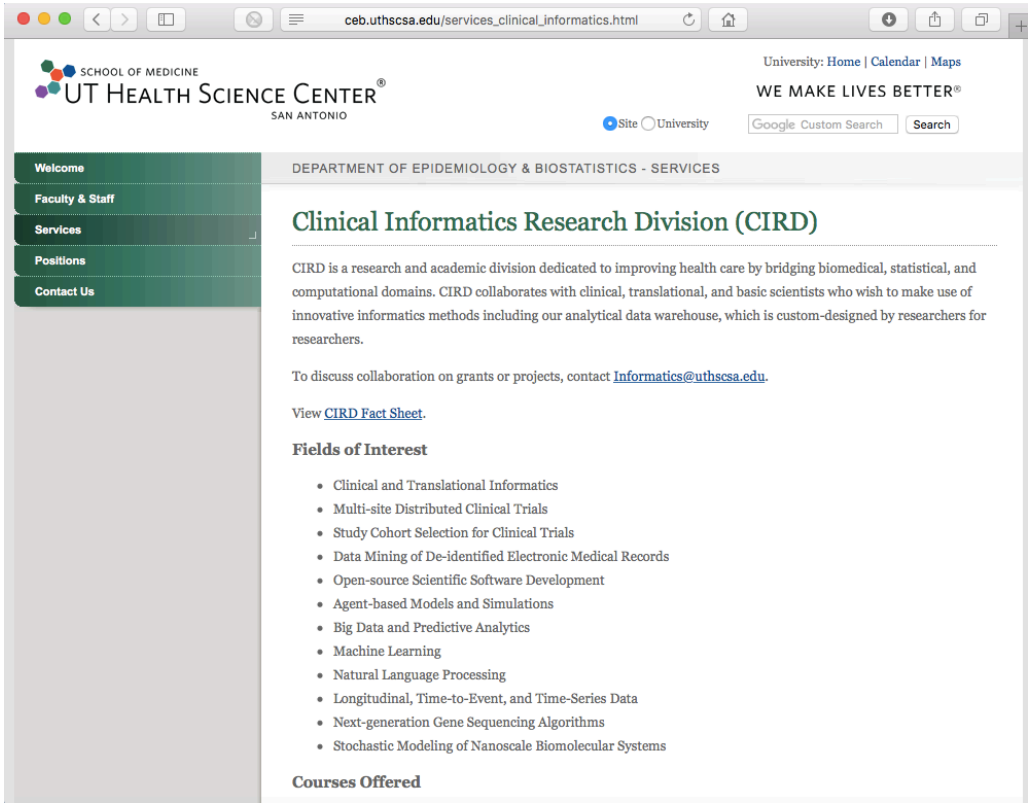


A BMI center of excellence in South Texas

- 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



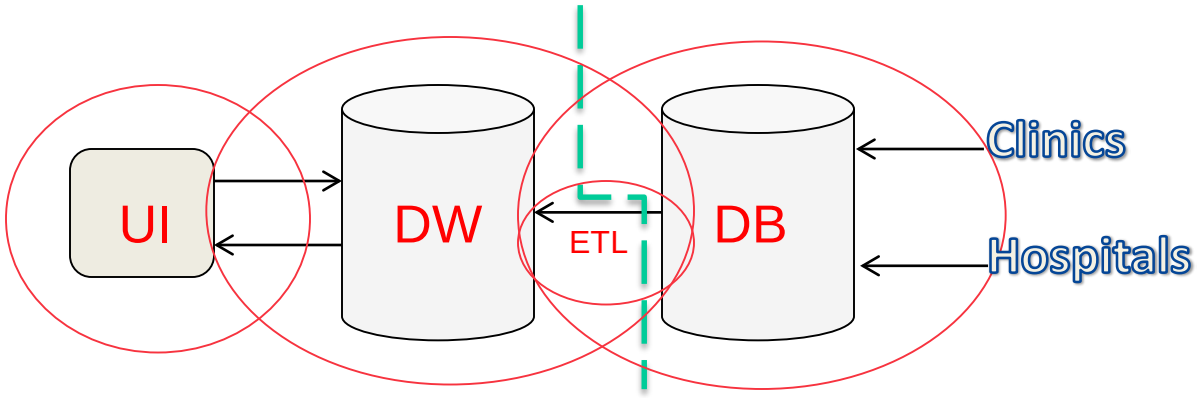
Solid external funding base

- 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

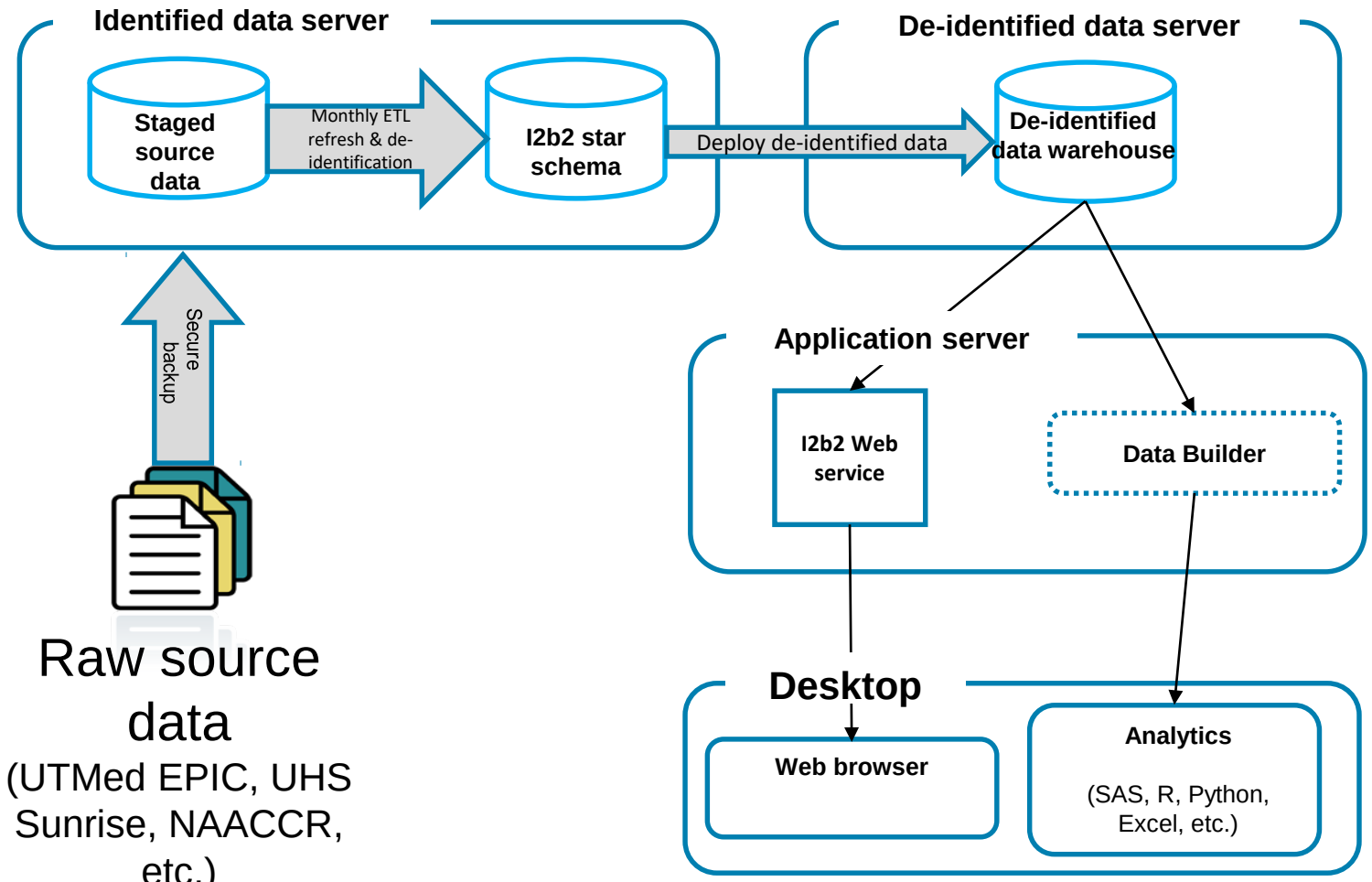
A research informatics baseline

- 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

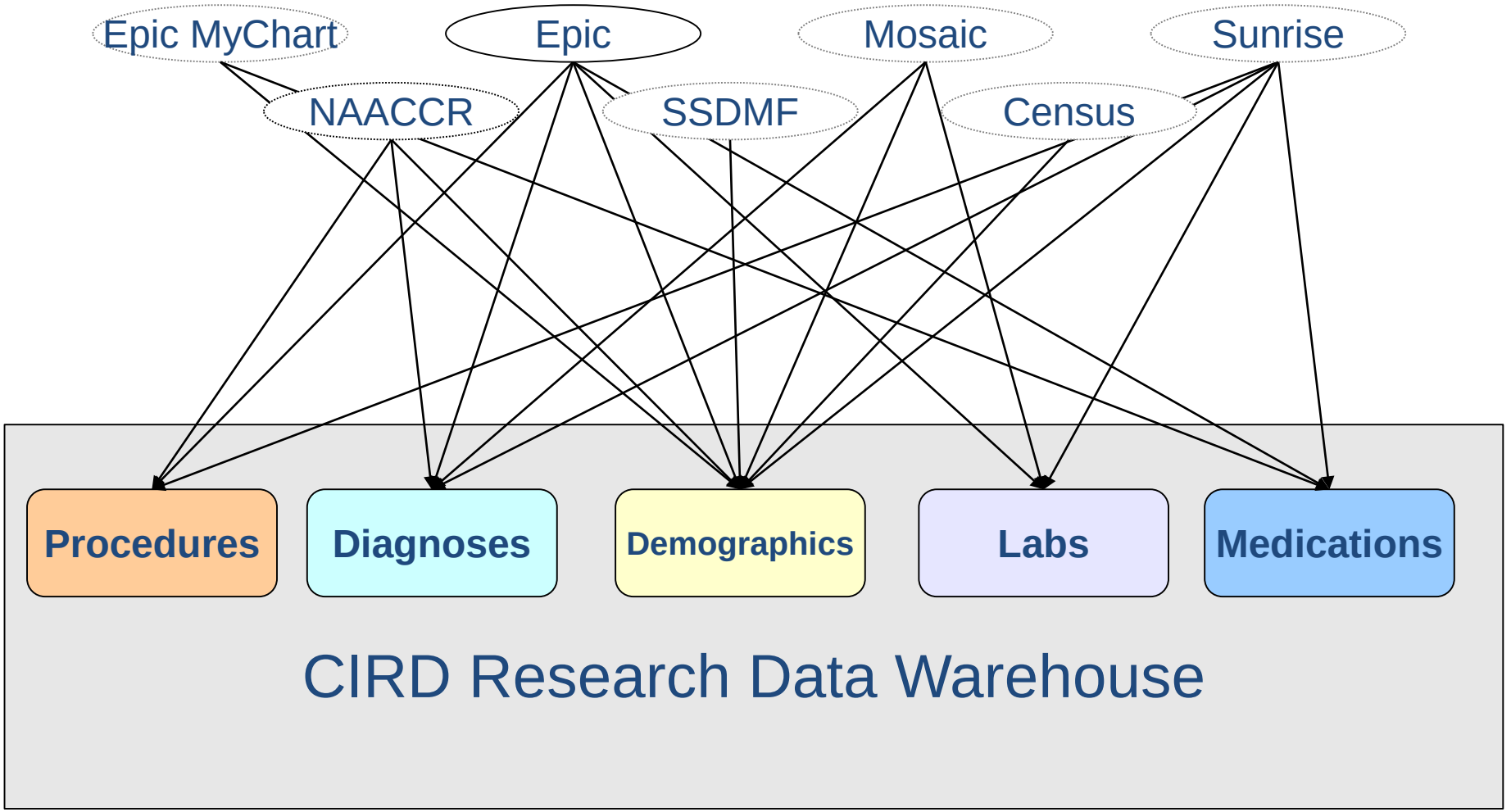


Sample data elements

Table 1: Data Elements Available in i2b2

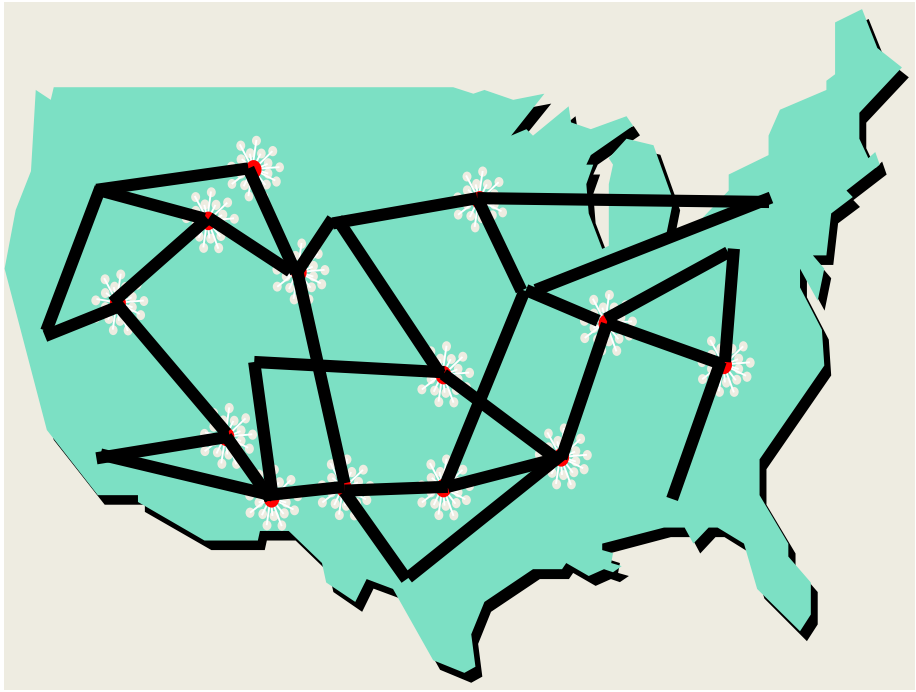
Data Element	Medical Terminology	Patients	Facts
Medications	RXNorm, VA drug classes	201,973	8,220,939
Diagnoses	ICD9, ICD10	318,849	17,772,539
Family History		156,419	5,969,790
Laboratory Values	LOINC	93,710	11,821,195
Procedures	CPT/HCPCS, ICD9, ICD10	309,988	3,884,754
Demographics		383,752	4,242,691
Visit Vitals		211,987	12,207,236
Flowsheets		213,044	11,656,060
NAACCR registry		14,129	2,703,751
Alerts		231,196	12,031,921
Allergies		74,611	144,516
Tobacco Use		235,828	7,460,275
Insurance Type		383,752	8,148,608
Provider Specialty		383,752	16,241,710
Visit Type		383,752	8,148,608

Sample use case: get EMR data from multiple sources and organize into a common data model

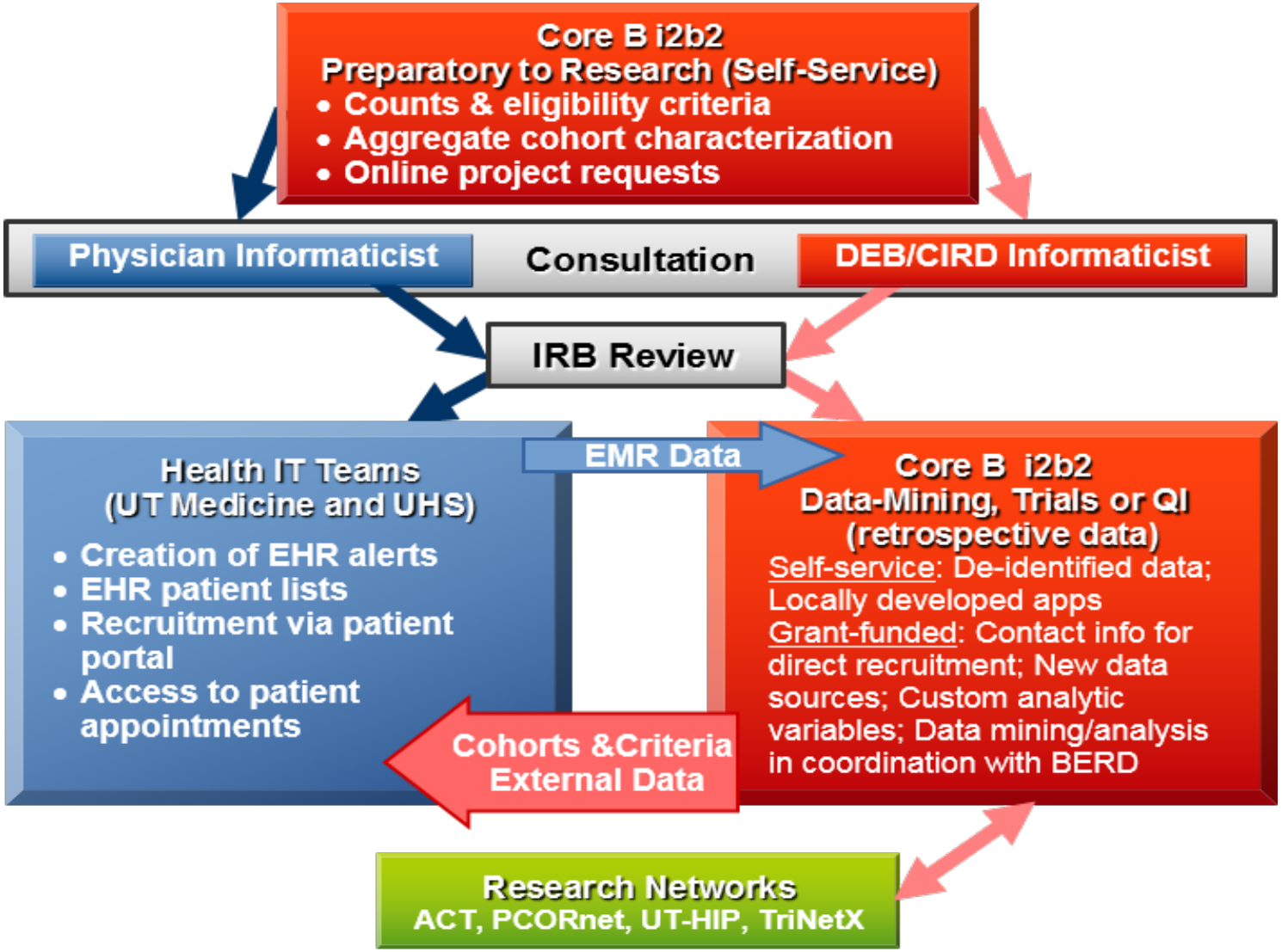


Think big

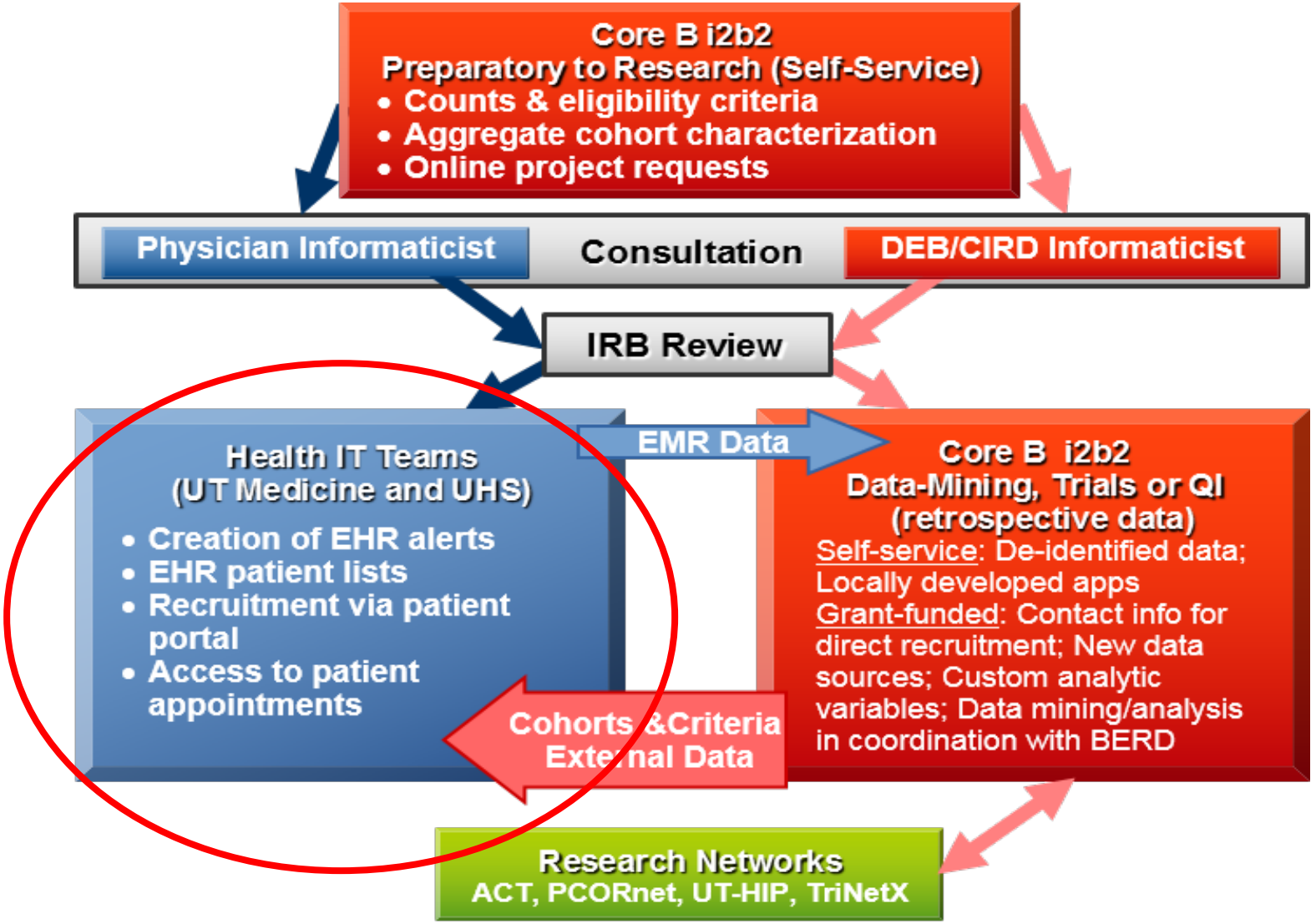
- IIMS Informatics
  - ACT SHRINE
- PCORI/PCORnet/GPC
  - SNOW SHRINE
- TriNetX



IIMS Informatics Core



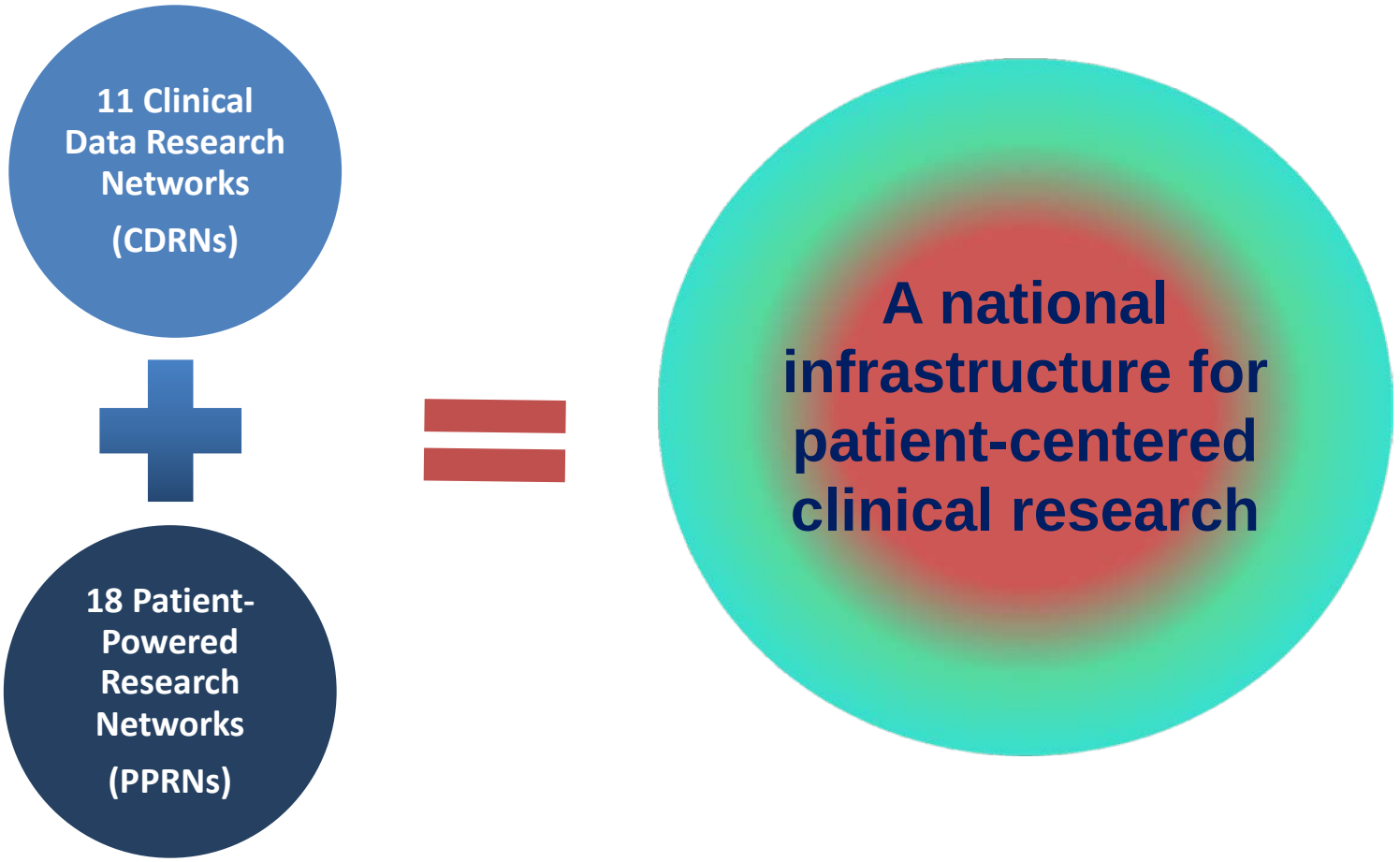
IIMS Informatics Core



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



The GPC

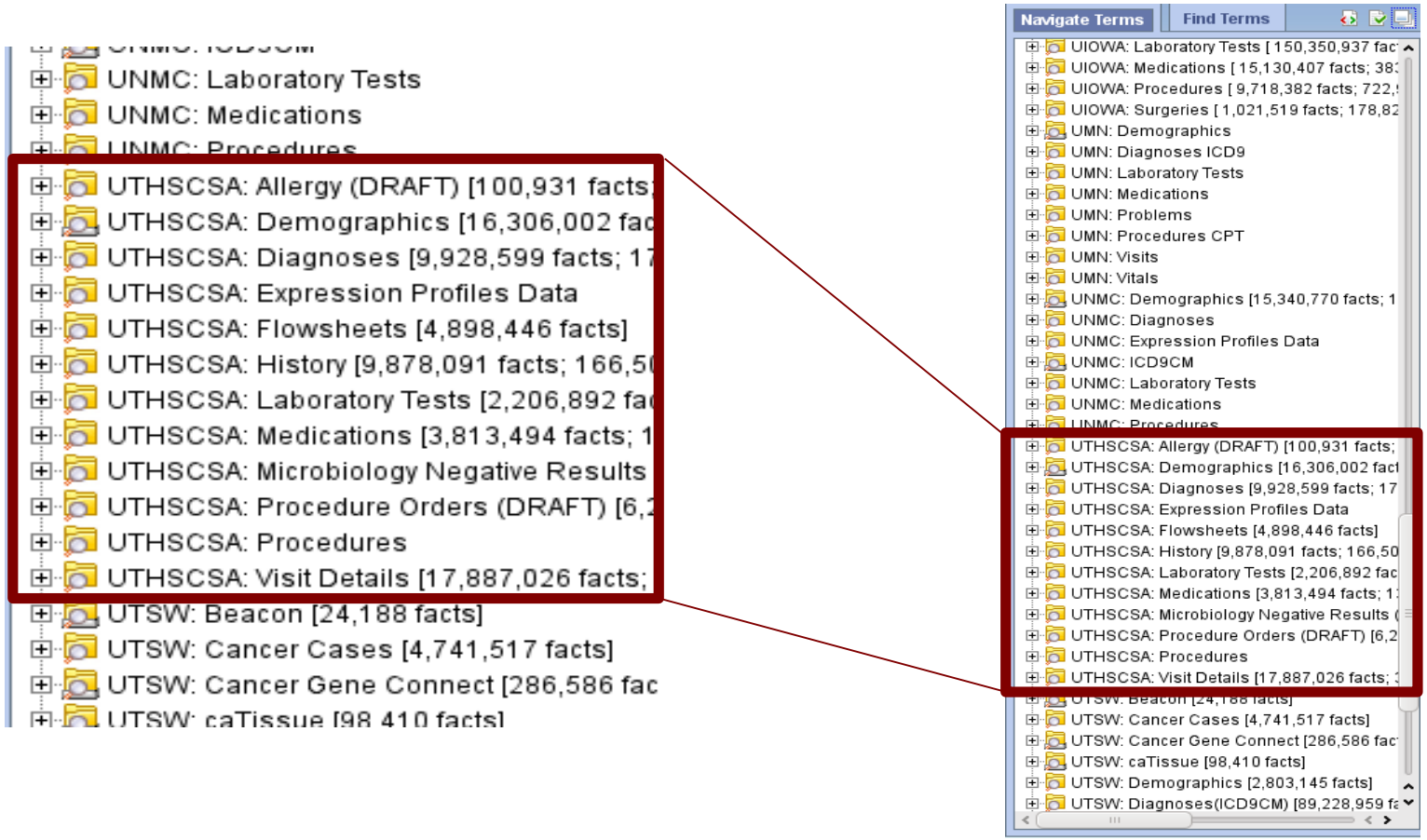
- The Great Plains Collaborative (GPC) grant
  - National Patient-Centered Outcomes Research Institute (PCORI), grant awarded 2013, renewed in 2015
- Research data network of networks (CDRNs)
  - 12 leading medical centers in 9 states
  - Breast Cancer, ALS and Obesity/Diabetes initial test cases
- Money, 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 Phase1
  - Lead in obesity
  - GPC representatives to the national PCORnet committee
- A number of PCORnet trials in Phase2
  - ADAPTABLE, ABX, ResDac, etc

Snapshot

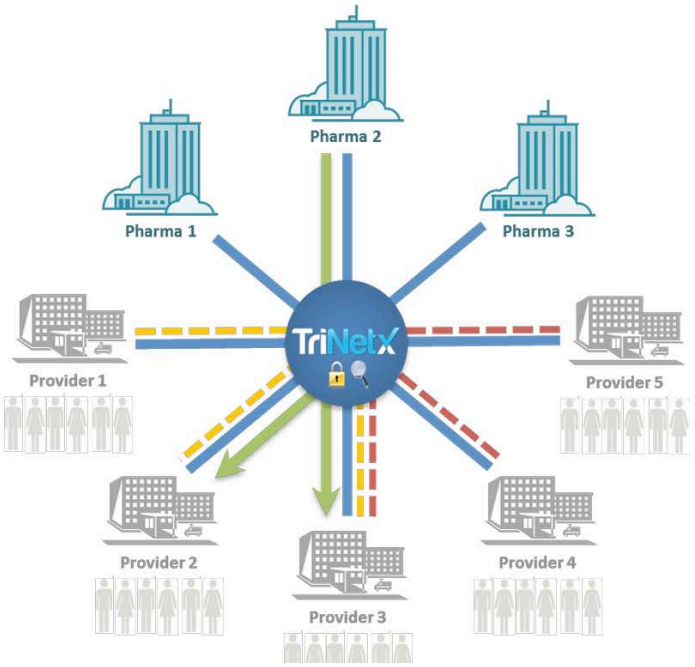

UNMC: Laboratory Tests
UNMC: Medications
UTHSCSA: Allergy (DRAFT) [100,931 facts;
UTHSCSA: Demographics [16,306,002 fac
UTHSCSA: Diagnoses [9,928,599 facts; 17
UTHSCSA: Expression Profiles Data
UTHSCSA: Flowsheets [4,898,446 facts]
UTHSCSA: History [9,878,091 facts; 166,5
UTHSCSA: Laboratory Tests [2,206,892 fa
UTHSCSA: Medications [3,813,494 facts; 1
UTHSCSA: Microbiology Negative Results
UTHSCSA: Procedure Orders (DRAFT) [6,2
UTHSCSA: Procedures
UTHSCSA: Visit Details [17,887,026 facts;
UTSW: Beacon [24,188 facts]
UTSW: Cancer Cases [4,741,517 facts]
UTSW: Cancer Gene Connect [286,586 fac
Navigate Terms
Find Terms
UIOWA: Laboratory Tests [150,350,937 fac
UIOWA: Medications [15,130,407 facts; 38
UIOWA: Procedures [9,718,382 facts; 722,
UIOWA: Surgeries [1,021,519 facts; 178,82
UMN: Demographics
UMN: Diagnoses ICD9
UMN: Laboratory Tests
UMN: Medications
UMN: Problems
UMN: Procedures CPT
UMN: Visits
UMN: Vitals
UNMC: Demographics [15,340,770 facts; 1
UNMC: Diagnoses
UNMC: Expression Profiles Data
UNMC: ICD9CM
UNMC: Laboratory Tests
UNMC: Medications
UTHSCSA: Allergy (DRAFT) [100,931 facts;
UTHSCSA: Demographics [16,306,002 fact
UTHSCSA: Diagnoses [9,928,599 facts; 17
UTHSCSA: Expression Profiles Data
UTHSCSA: Flowsheets [4,898,446 facts]
UTHSCSA: History [9,878,091 facts; 166,50
UTHSCSA: Laboratory Tests [2,206,892 fac
UTHSCSA: Medications [3,813,494 facts; 1
UTHSCSA: Microbiology Negative Results (
UTHSCSA: Procedure Orders (DRAFT) [6,2
UTHSCSA: Procedures
UTHSCSA: Visit Details [17,887,026 facts;
UTSW: Cancer Cases [4,741,517 facts]
UTSW: Cancer Gene Connect [286,586 fac
UTSW: caTissue [98,410 facts]
UTSW: Demographics [2,803,145 facts]
UTSW: Diagnoses(ICD9CM) [89,228,959 fa

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**

= Pharma Research Network
= Multi-provider Peer Network
= Multi-provider Peer Network
= Workflow



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 stategies

- 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.

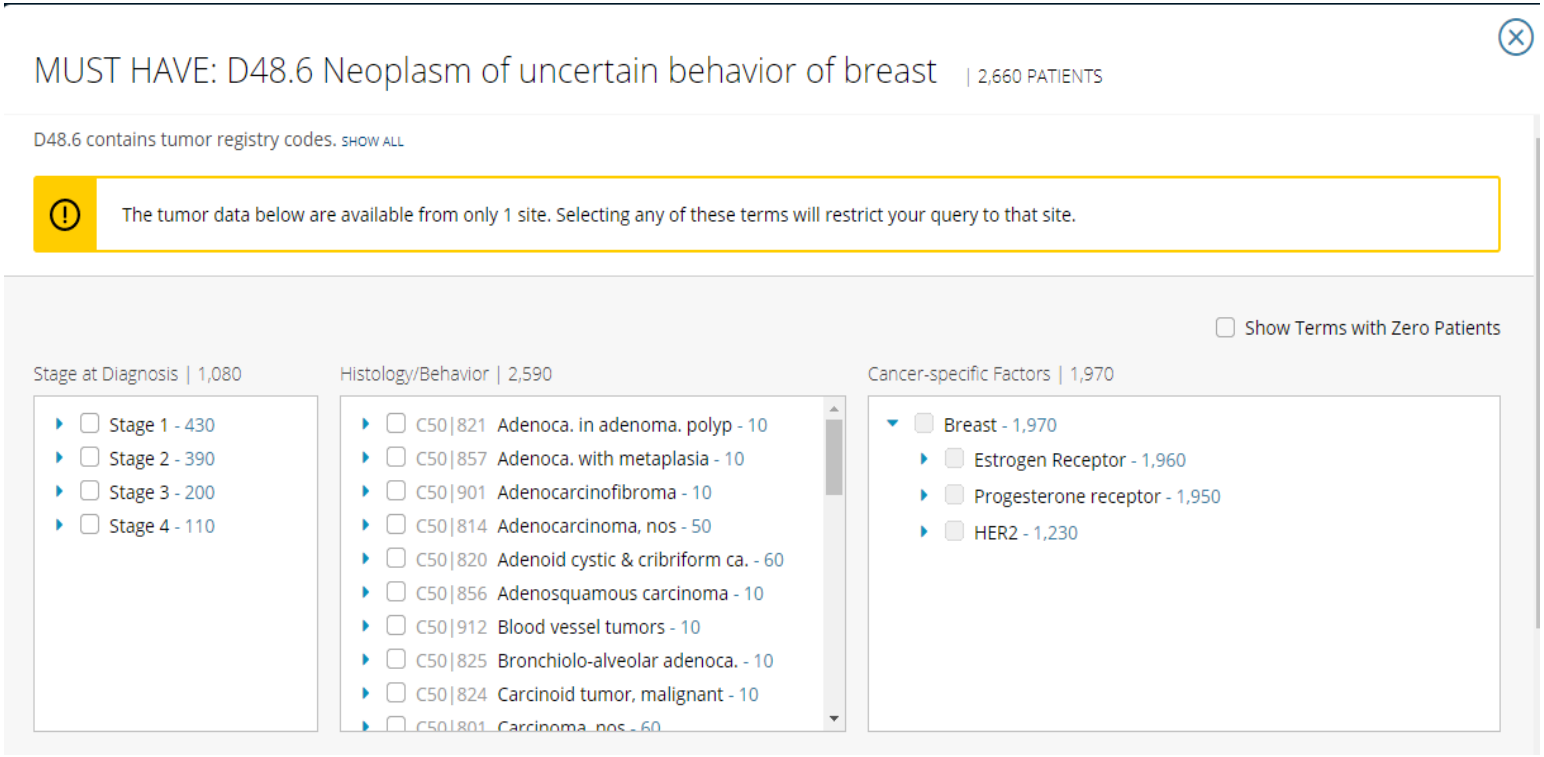
Use case: cancer data integration

- 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.

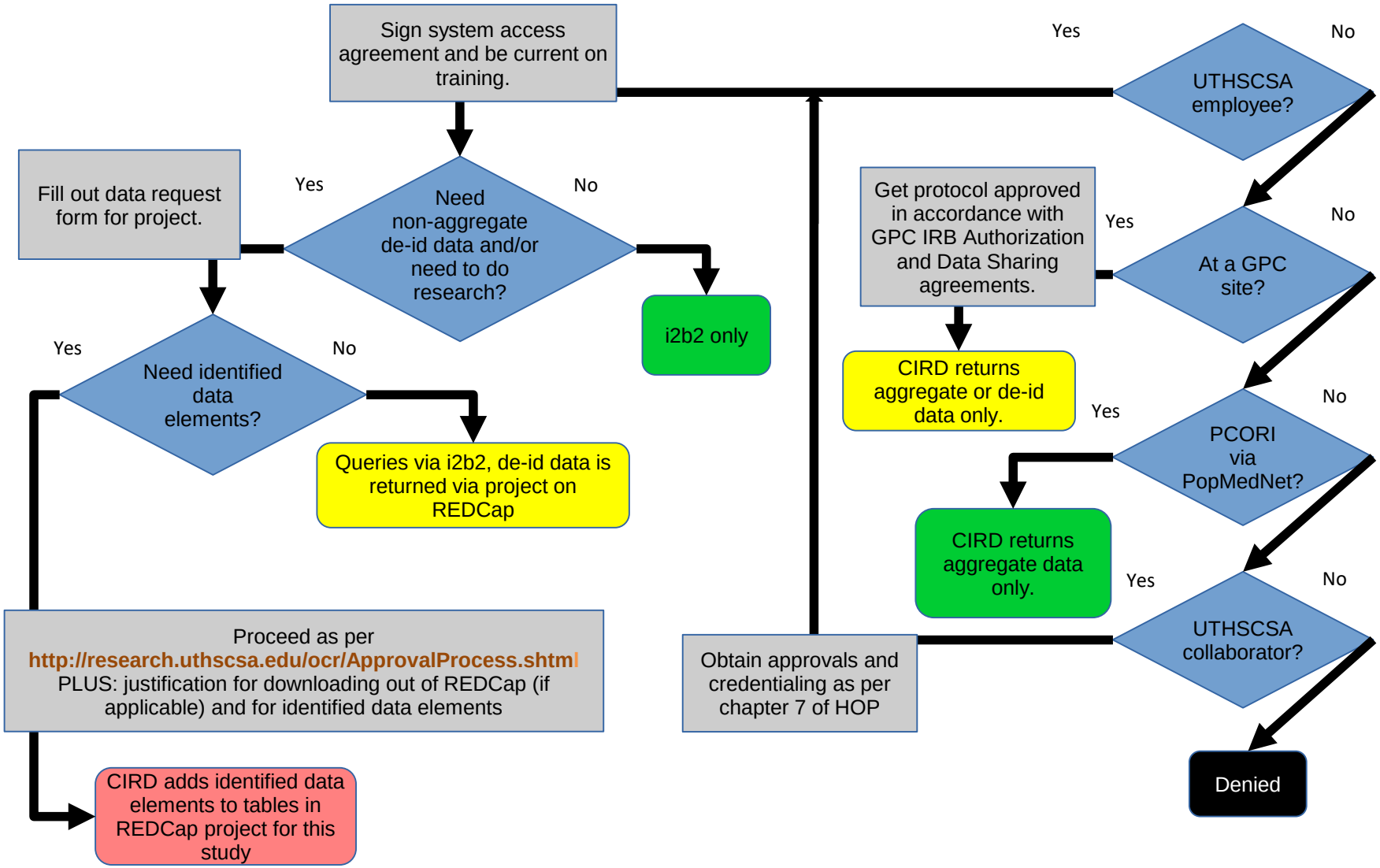
Use case: cancer data integration



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 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



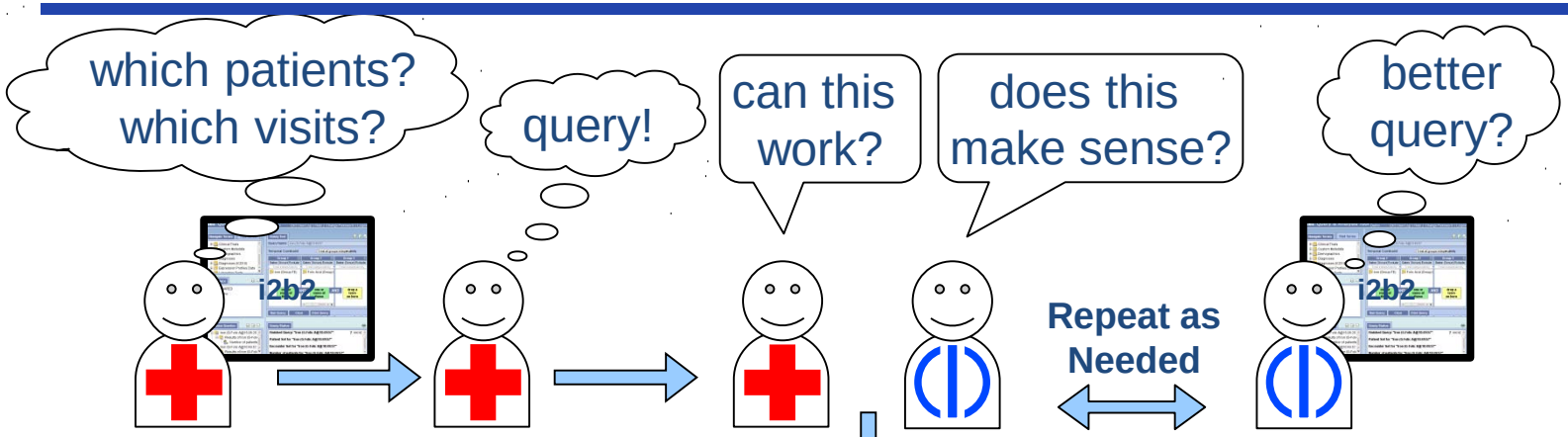
A BMI center of excellence in South Texas

- 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

Improving the way we work

- 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



research goals

appropriate statistical model, minimal sample size

output data in proper format

Researcher

Statistician

Informatician

Four Basic Formats

One visit/patient

or

Constant sequence of visits per patient

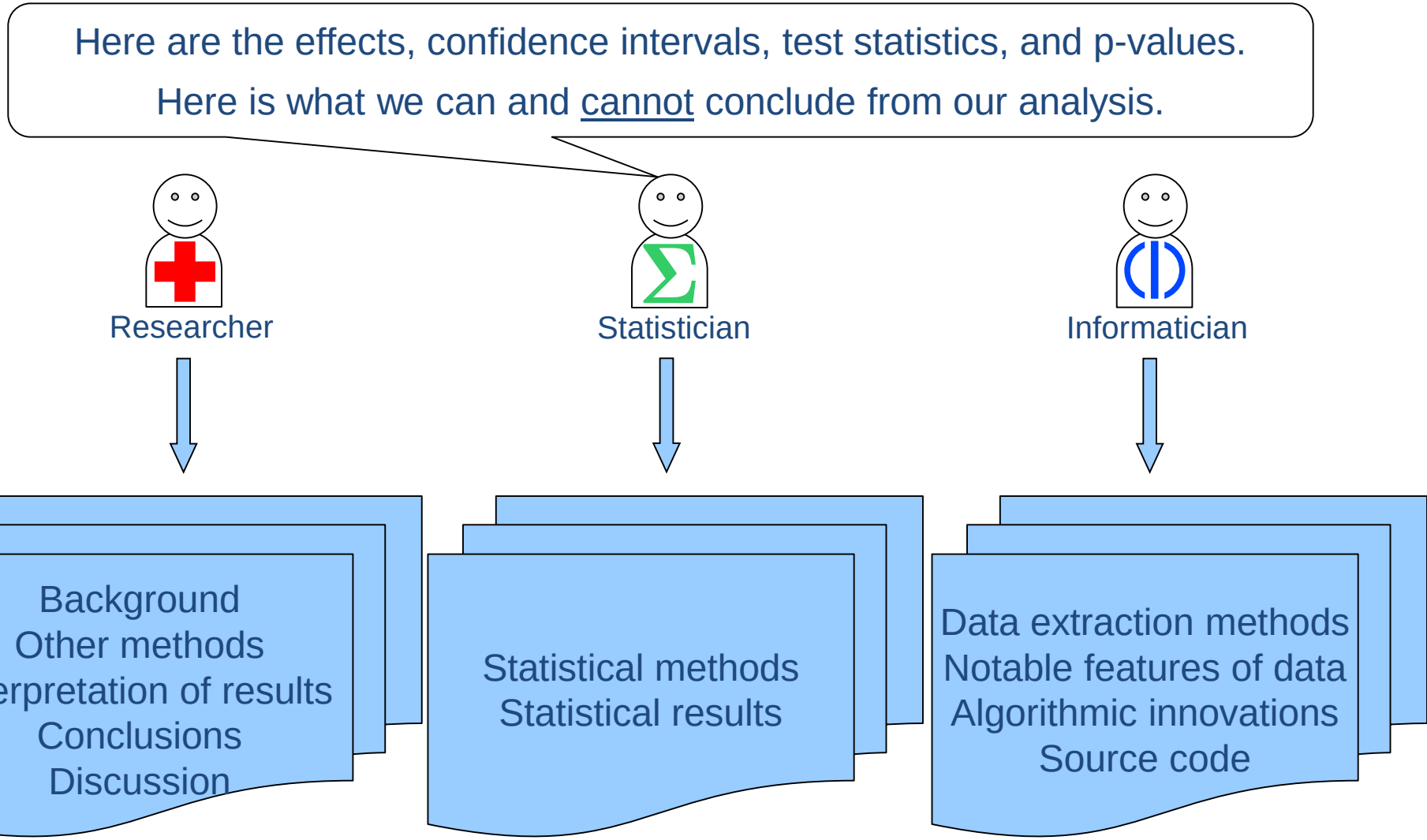
or

Time series

or

Frequency table

Model interpretation



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 importnat, 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)

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

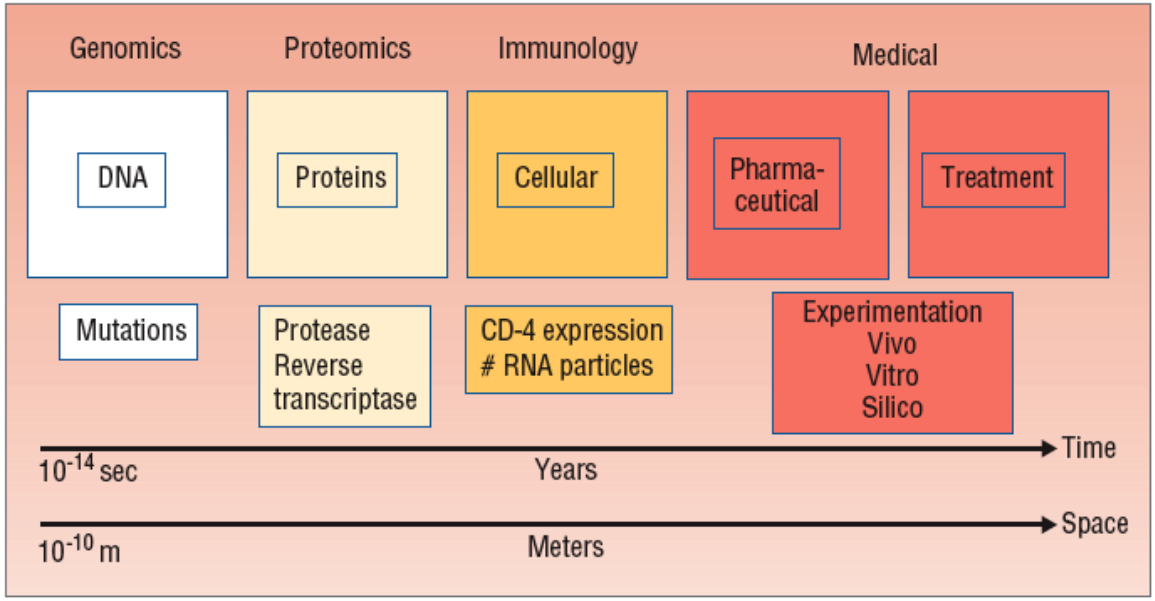
Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child 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 exquisite knowledge of 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 'team' 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 multiple clone tiling paths to capture all of the sequence variations. All of these deficiencies and others not cited provide a less-than-optimal alignment target for next-generation sequencing data and can confound the analytical validity of variants 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 augment 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.

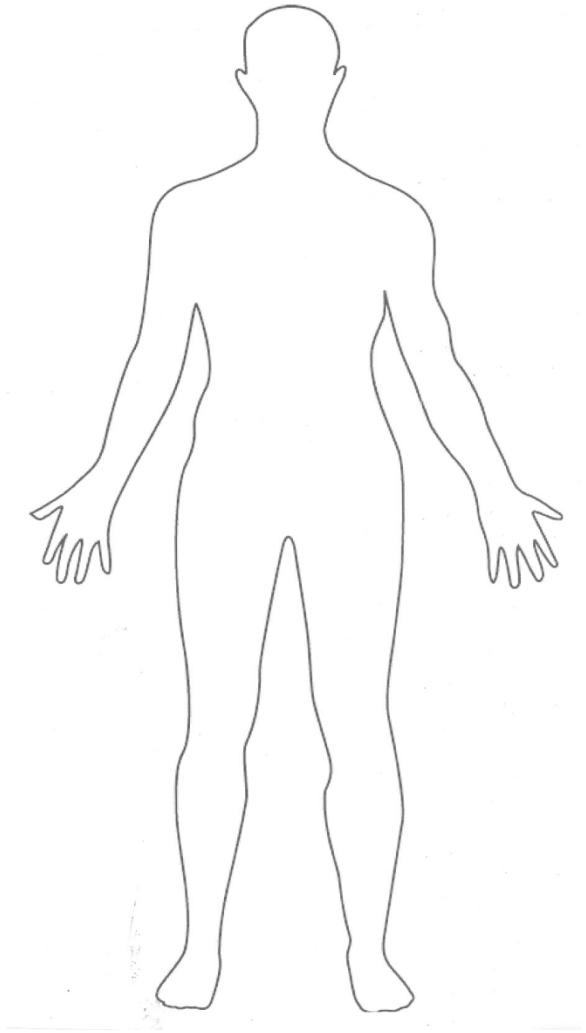
*Correspondence: emardis@wustl.edu
The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

Multi-scale complexity in BM science



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

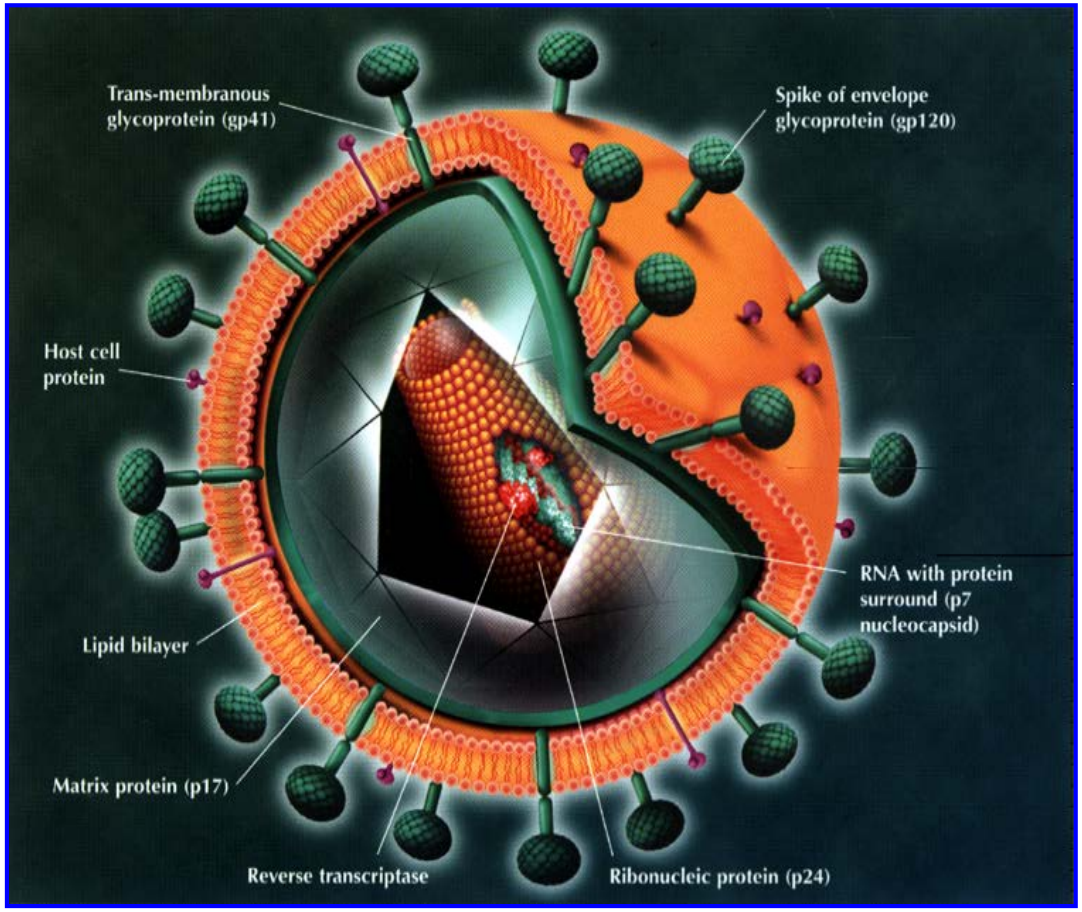
- Multi-cellular Organisms
- Social, Epidemiological Networks
- Earthquakes
- Traffic flow
- Immune System (HIV)

Complexity in HIV

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

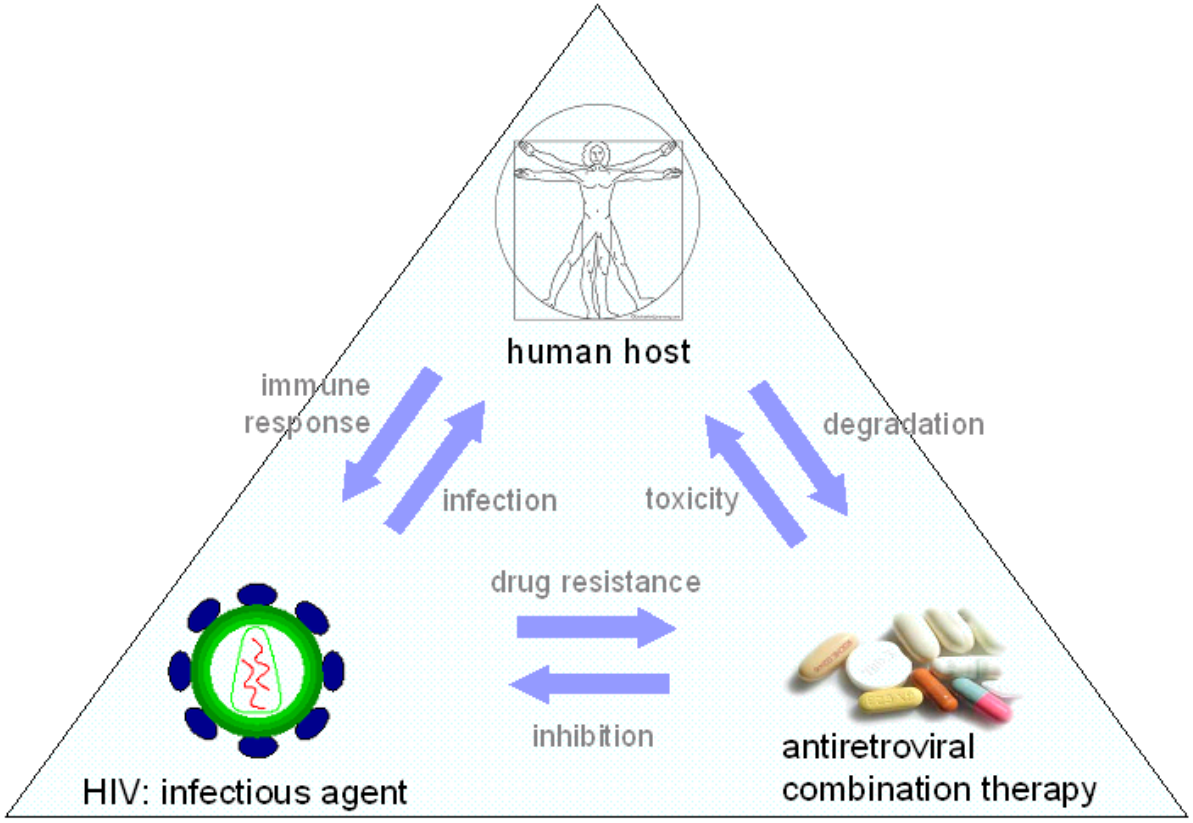
Neil Ferguson, Nature, 12 April 2007

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



Agent-based simulation models for HIV infection

- HIV is a unique problem to model in many ways
  - Does not fit traditional epidemiological models for disease, it's not transmitted by air or casual contact
  - Mainly a result of human behavior (with some exceptions like mother-to-child infection and blood transfusions)
  - 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

Activity vs Process

"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

	Population	Simulation Parameter Data Sources	Simulation Tool	Survey Significance
Heuveline et al. (2003)	Heterosexual, Eastern and southern Africa	Unspecified	Unspecified	Introduction of migration, marriage, and divorce
Teweldemedhin et al. (2005)	Heterosexual, South Africa	Department of Health, South Africa (2002); Rehle & Shisana (2003); Shisana & Simbayi (2002)	JADE	Decision based functions
Sumodhee et al. (2005)	MSM, Taiwan	Unspecified	Unspecified	Population groups based based on behavior patterns, behavior modification support
Rhee (2006)	Papa New Guinea	Gray et al. (2003) Nadler (2005) Vanhem et al.	AnyLogic	Preferential attachment behavior, Awareness
Tawfik, Farag (2008)	Heterosexual, South Africa	Statistics South Africa, UNESCO, BBC World Service Trust African Media Development Initiative, 2006; UNAIDS, Women and AIDS Fact Sheet, 2004.	Unspecified	Prevention strategy simulation
Alam et al. (2008)	Heterosexual, South Africa	South African National HIV Survey	Repast J	Simple Sexual Mixing Scheme
Mei et al. (2010)	MSM, Amsterdam	Amsterdam Cohort Study (ACS)	Luke et al. (2004); O'Madadhain et al. (2005)	Dynamic network
Knittel et al (2011)	Heterosexual	Gray et al. (2003), Public health studies; Evolutionary and social psychology; Theoretical work from psychology and public health	Repast J	Additional sexual behavior behaviors, multiple partners
Richardson and Grund (2012)	IDU, Bushwick, Brooklyn	Curtis et al. (1995); Friedman et al. (1997); Friedman et al. (1999); Kottiri et al. (2002)	NetLogo	Calibrated agent-based model

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

Prevention Strategy	*Description*
Behaviour change programs	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
Education and awareness programs	Identify the impact of awareness programs on high-risk groups
Predictors of condom use with steady and random partners	Identify strategies to promote condom use that increase awareness about their effectiveness against not only unwanted pregnancies but also HIV and other STDs
Psychosocial variables such as depression	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
Linking to appropriate care and prevention services	Programs that provide linking and close monitoring 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
Support of adherence to treatment regimens	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
Environmental–structural interventions targeting sex workers	Implementing standard routine programs which monitor indicators on service provision, service uptake, and community activities
Diagnosing HIV infections outside medical settings	Using widely available devices like OraQuick to increase access to early diagnosis and referral for treatment and prevention services in high-HIV prevalence settings, including correctional facilities
Making HIV testing a routine part of medical care	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
Harm Reduction Programs	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

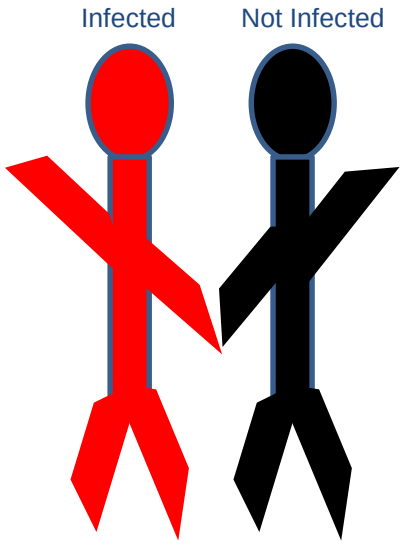
HIV infection AB-simulation

- 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



1. Nagelkerke NJD, Arora P, Jha P, Williams B, McKinnon L, de Vlas SJ. The Rise and Fall of HIV in High-Prevalence Countries: A Challenge for Mathematical Modeling. PLoS Comput Biol [Internet]. 2014 Mar 13 [cited 2016 Apr 15];10(3). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3952813/
2. Rothenberg R. HIV Transmission Networks. Curr Opin HIV AIDS. 2009 Jul;4(4):260–5.
3. Gopalappa C, Farnham PG, Chen Y-H, Sansom SL. Progression and Transmission of HIV/AIDS (PATH 2.0) A New Agent-Based Model to Estimate HIV Transmissions in the United States. Med Decis Making. 2016 Sep 19;0272989X16668509.
4. Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. Proc Natl Acad Sci U S A. 2002 May 14;99(Suppl 3):7280–7.

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

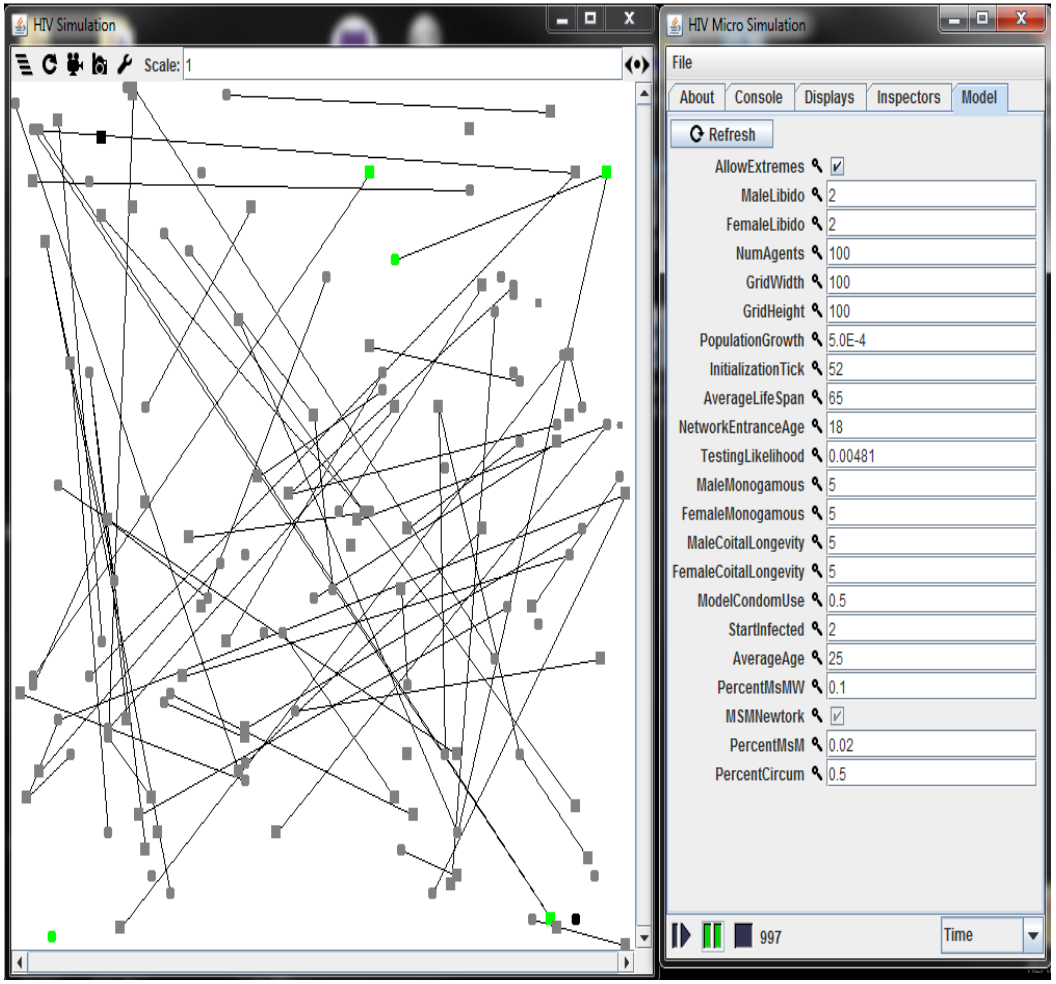
Approach

- 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



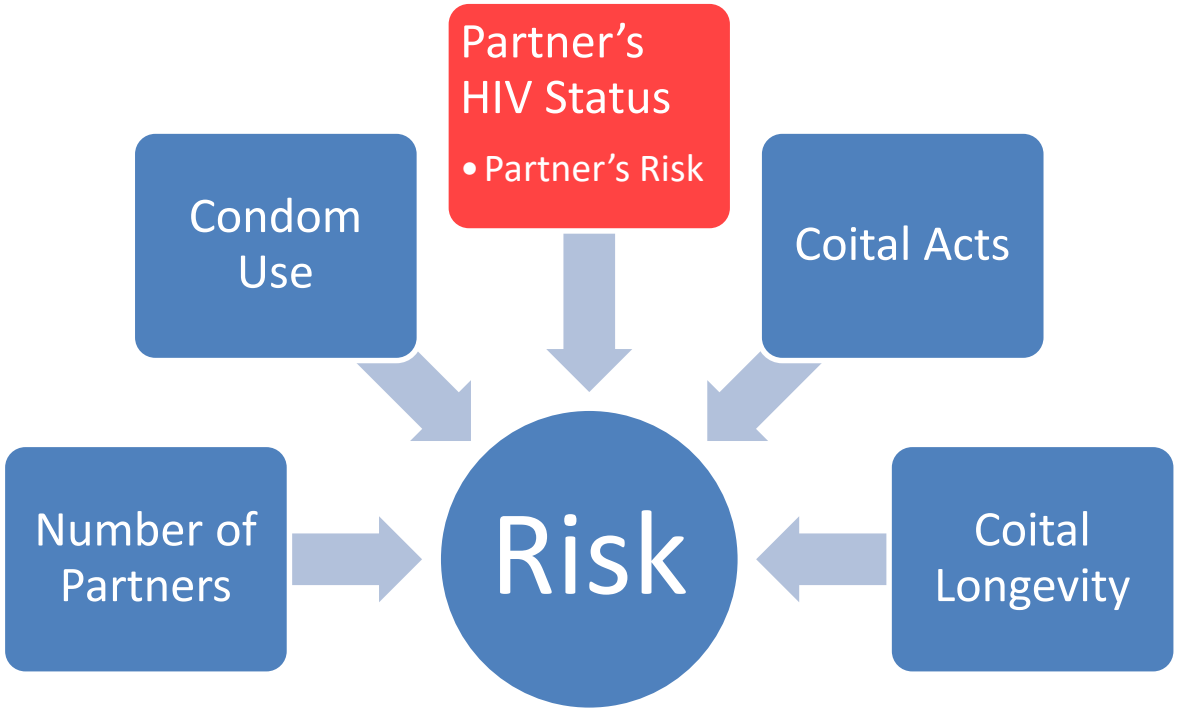
Model design

- 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

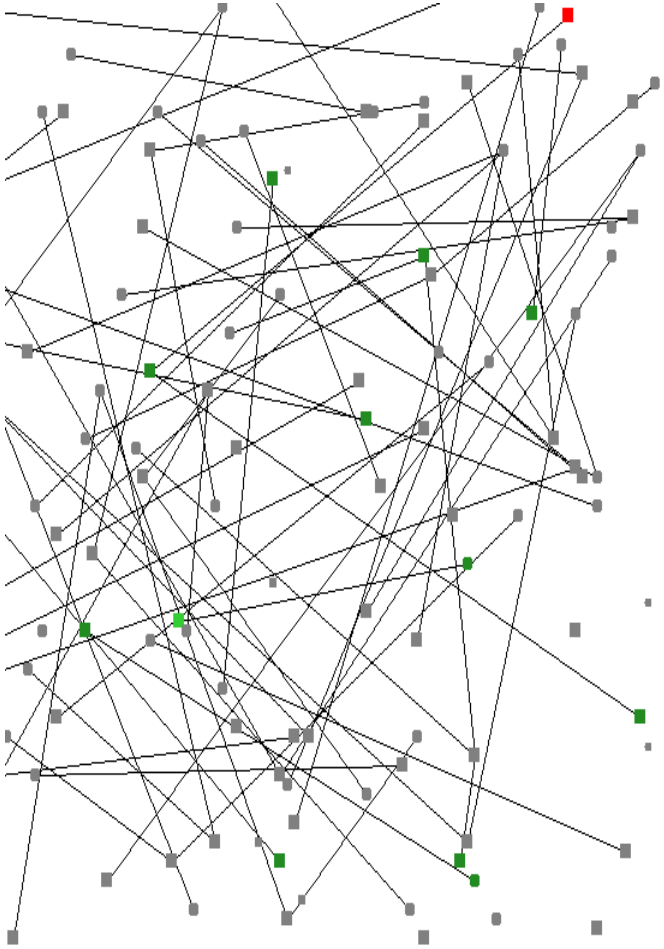


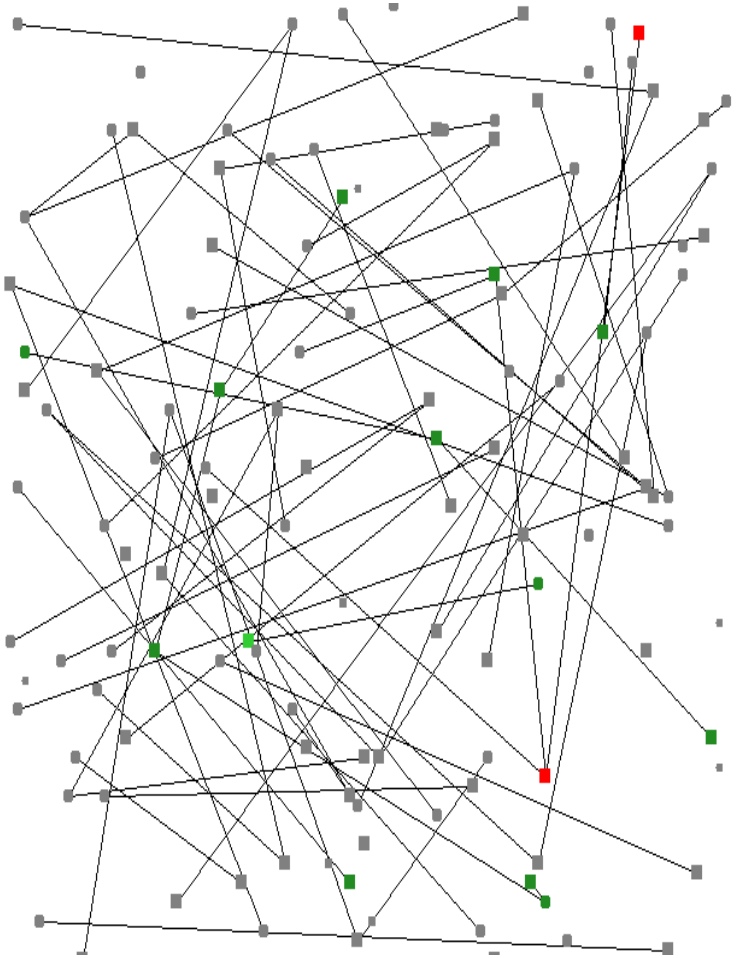
Population risk behavior

- 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 % [1]

1 Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. PLOS Med. 2005 Oct 25;2(11):e298.

Agent networks

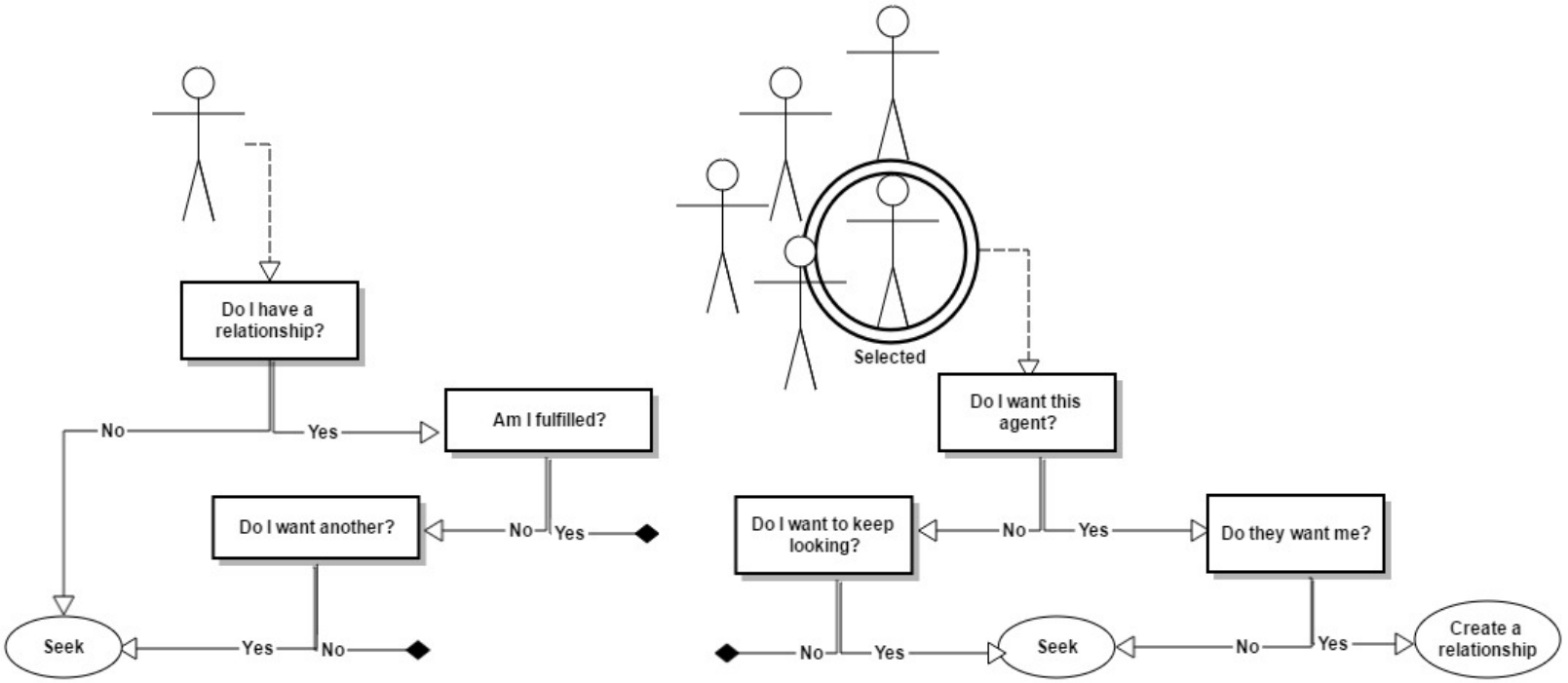




Agent networks

- 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.

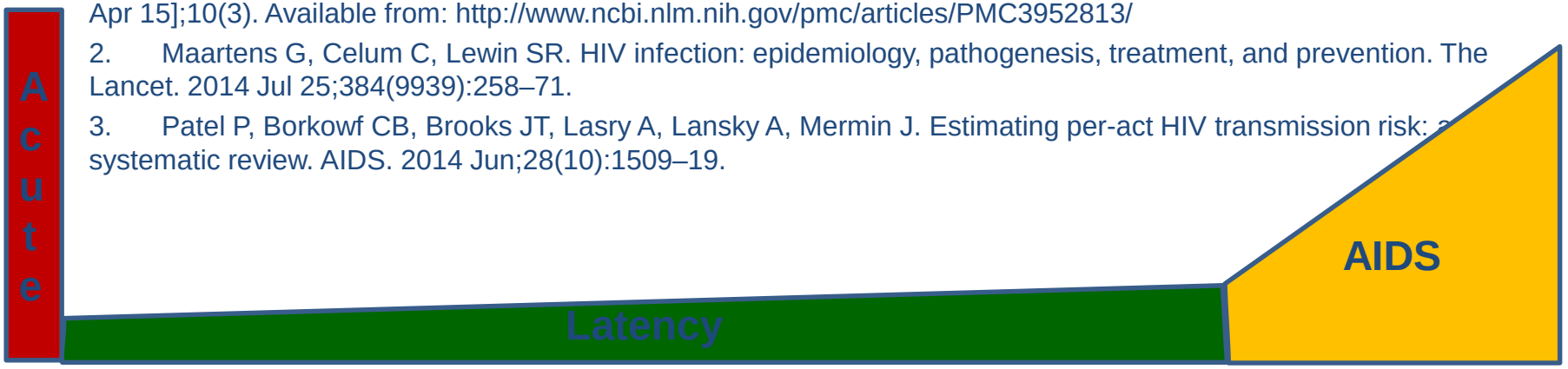
Agent wellness and treatment

- 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 [1]
- Viral suppression reduces infectiousness by 96% [2]
- Per Mode Risk based on CDC Per interaction estimates [3]
  - Anal vs Vaginal
  - Insertive vs Receptive

1. Nagelkerke NJD, Arora P, Jha P, Williams B, McKinnon L, de Vlas SJ. The Rise and Fall of HIV in High-Prevalence Countries: A Challenge for Mathematical Modeling. PLoS Comput Biol [Internet]. 2014 Mar 13 [cited 2016 Apr 15];10(3). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3952813/
2. Maartens G, Celum C, Lewin SR. HIV infection: epidemiology, pathogenesis, treatment, and prevention. The Lancet. 2014 Jul 25;384(9939):258–71.
3. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A, Mermin J. Estimating per-act HIV transmission risk: a systematic review. AIDS. 2014 Jun;28(10):1509–19.



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.