University at Buffalo
Clinical and Translational Science Institute

UNSTRUCTURED DATA UNIVERSITY AT BUFFALO

Peter L. Elkin, MD, MACP, FACMI, FNYAM
Director, Informatics Core of the UB CTSA
Professor and Chair, Department of Biomedical Informatics

Professor of Internal Medicine

Professor of Surgery

Professor of Pathology and Anatomical Sciences

Advancing research discoveries to improve health for all

Sources of Unstructured Data

- Documents
- Reports
- Legions of Figures
- Tabular data names
- Field names in databases

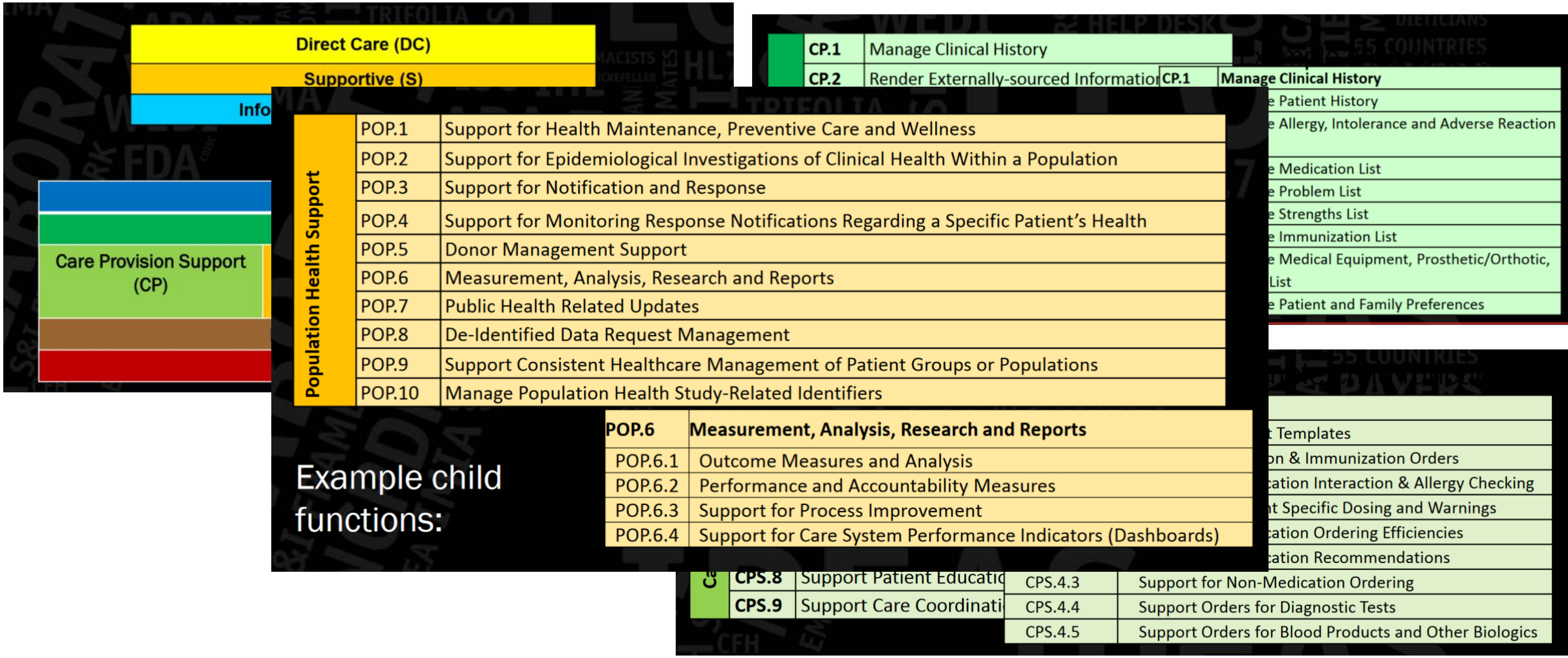
Some Datatypes are Only accessible from Unstructured Data

- Social Determinants of Health
- Signs and Symptoms
- Physical Exam findings
- Counseling
- Quality of Life
- Behavioral Data
- Street drug use
- Opinions

Electronic Health records

- Began in the 1960’s
  - HELP – Utah
  - CoSTAR – MGH
- Commercial Systems
  - Technicon – from Lockheed 1963 developed for El Camino Hosopital used NIH clinical center – and later become TDS (Han Article)
  - Meditech – 1969
  - 1977 MUM{PS was developed as a standard
  - 1979 – Epic started as an outpatient system
  - 1979 – Cerner which started as a lab system
  - 1980s – Boston Beth Israel System
  - 1980 – Regenstrief Institute of Indiana University
  - 1981 – VA Distributed Hospital Computing Program
  - 1994 – DHCP became VistA
  - 1994 – CPRS
  - 2009 – ARRA EHR Adoption

Electronic Health records Functional Specification from HL7



Direct Care (DC)

Supportive (S)

Info

Care Provision Support (CP)

Population Health Support		
	POP.1	Support for Health Maintenance, Preventive Care and Wellness
	POP.2	Support for Epidemiological Investigations of Clinical Health Within a Population
	POP.3	Support for Notification and Response
	POP.4	Support for Monitoring Response Notifications Regarding a Specific Patient's Health
	POP.5	Donor Management Support
	POP.6	Measurement, Analysis, Research and Reports
	POP.7	Public Health Related Updates
	POP.8	De-Identified Data Request Management
	POP.9	Support Consistent Healthcare Management of Patient Groups or Populations
	POP.10	Manage Population Health Study-Related Identifiers

Example child functions:

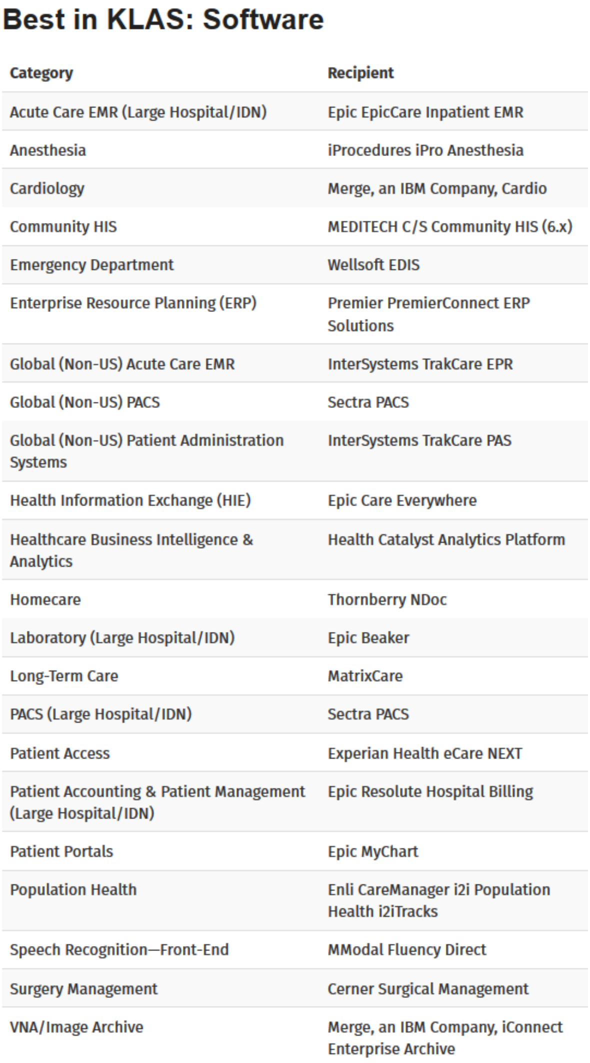
POP.6	Measurement, Analysis, Research and Reports
POP.6.1	Outcome Measures and Analysis
POP.6.2	Performance and Accountability Measures
POP.6.3	Support for Process Improvement
POP.6.4	Support for Care System Performance Indicators (Dashboards)

CP.1	Manage Clinical History
CP.2	Render Externally-sourced Informatio

CP.1	Manage Clinical History
	e Patient History
	e Allergy, Intolerance and Adverse Reaction
	e Medication List
	e Problem List
	e Strengths List
	e Immunization List
	e Medical Equipment, Prosthetic/Orthotic, List
	e Patient and Family Preferences

	t Templates
	on & Immunization Orders
	cation Interaction & Allergy Checking
	t Specific Dosing and Warnings
	cation Ordering Efficiencies
	cation Recommendations
CPS.4.3	Support for Non-Medication Ordering
CPS.4.4	Support Orders for Diagnostic Tests
CPS.4.5	Support Orders for Blood Products and Other Biologics

Ca	CPS.8	Support Patient Educatio
	CPS.9	Support Care Coordinati



Best in KLAS: Software

Category	Recipient
Acute Care EMR (Large Hospital/IDN)	Epic EpicCare Inpatient EMR
Anesthesia	iProcedures iPro Anesthesia
Cardiology	Merge, an IBM Company, Cardio
Community HIS	MEDITECH C/S Community HIS (6.x)
Emergency Department	Wellsoft EDIS
Enterprise Resource Planning (ERP)	Premier PremierConnect ERP Solutions
Global (Non-US) Acute Care EMR	InterSystems TrakCare EPR
Global (Non-US) PACS	Sectra PACS
Global (Non-US) Patient Administration Systems	InterSystems TrakCare PAS
Health Information Exchange (HIE)	Epic Care Everywhere
Healthcare Business Intelligence & Analytics	Health Catalyst Analytics Platform
Homecare	Thornberry NDoc
Laboratory (Large Hospital/IDN)	Epic Beaker
Long-Term Care	MatrixCare
PACS (Large Hospital/IDN)	Sectra PACS
Patient Access	Experian Health eCare NEXT
Patient Accounting & Patient Management (Large Hospital/IDN)	Epic Resolute Hospital Billing
Patient Portals	Epic MyChart
Population Health	Enli CareManager i2i Population Health i2iTracks
Speech Recognition—Front-End	MModal Fluency Direct
Surgery Management	Cerner Surgical Management
VNA/Image Archive	Merge, an IBM Company, iConnect Enterprise Archive

University at Buffalo
Clinical and Translational Science Institute

University at Buffalo
The State University of New York

High Performance Computing and Natural Language Understanding

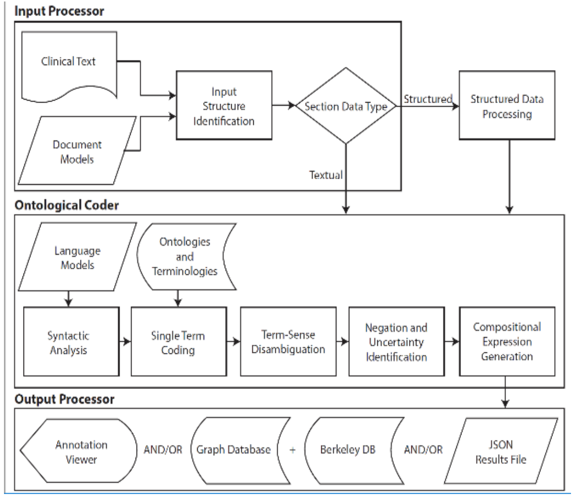
Peter L. Elkin[1], Daniel R Schlegel[2], Christopher Crowner[1], Frank LeHoullier[1]
[1]*Department of Biomedical Informatics, University at Buffalo, SUNY, Buffalo ,NY USA*
[2]*SUNY Oswego, New York USA*

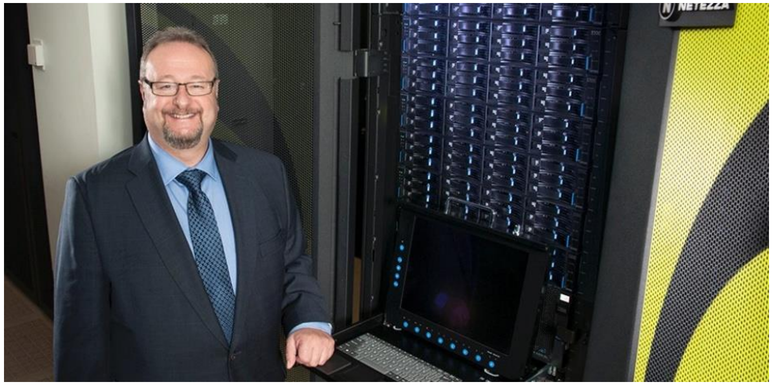
Introduction

Big data is expanding exponentially. We are looking at housing, processing, analyzing and retrieving Petabytes of data every day. With the advent of Genomic and Proteomic data we are increasingly challenged with understanding the patient’s phenotype with greater specificity and detail. This is going to require developing and applying ontology at a more granular and consistent fashion.

Methods

The UB Center for Computational Research (CCR) is an NSF sponsored supercomputing facility where we can scale to 16,000 nodes. We have a large number of high memory (>64GB) nodes. We installed a script to access the CCR scheduling application and deployed our HTP application (See Figure 1).

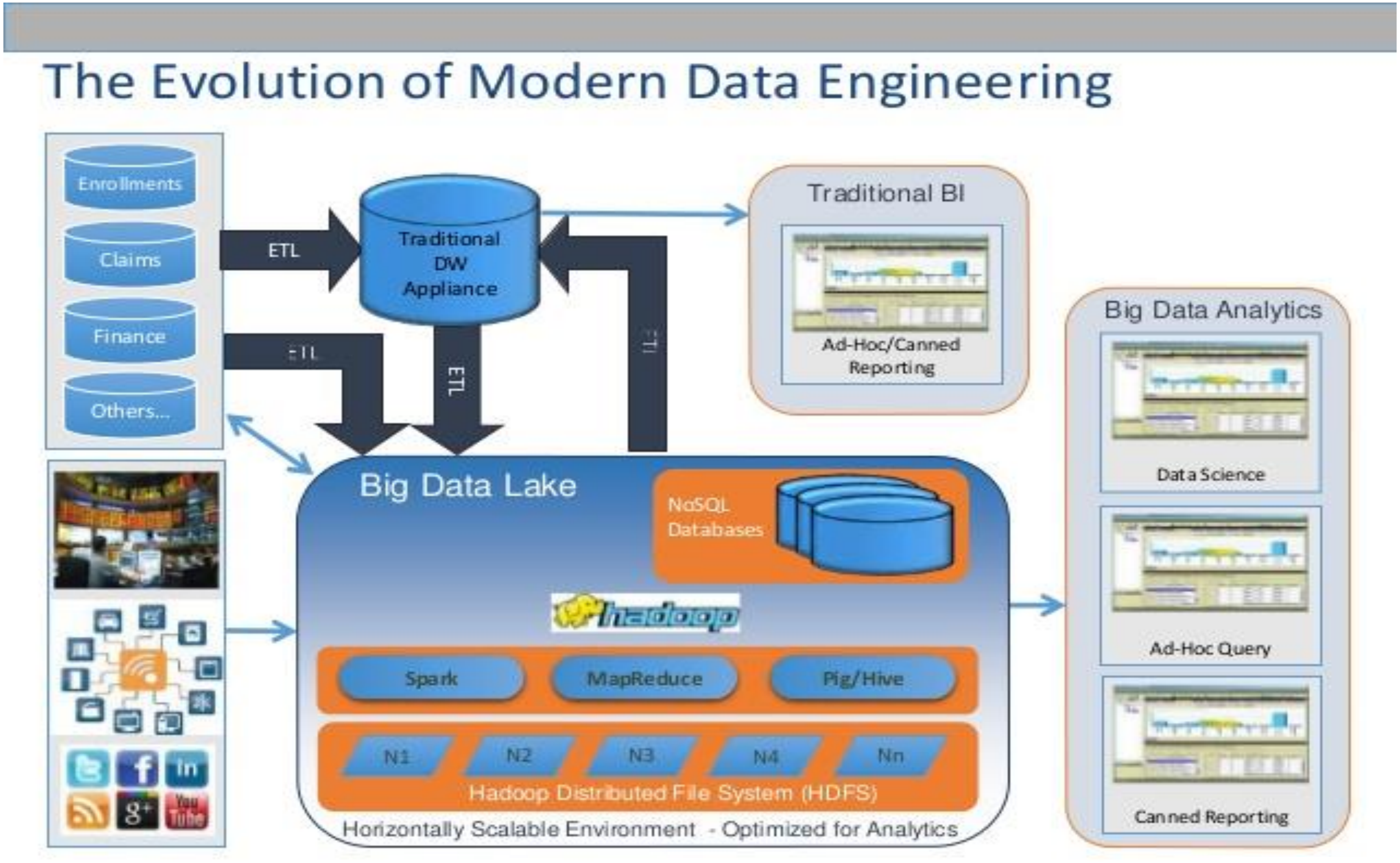

Input Processor
Clinical Text
Document Models
Input Structure Identification
Section Data Type
Structured
Structured Data Processing
Textual
Ontological Coder
Language Models
Ontologies and Terminologies
Syntactic Analysis
Single Term Coding
Term-Sense Disambiguation
Negation and Uncertainty Identification
Compositional Expression Generation
Output Processor
Annotation Viewer
AND/OR
Graph Database
+
Berkeley DB
AND/OR
JSON Results File



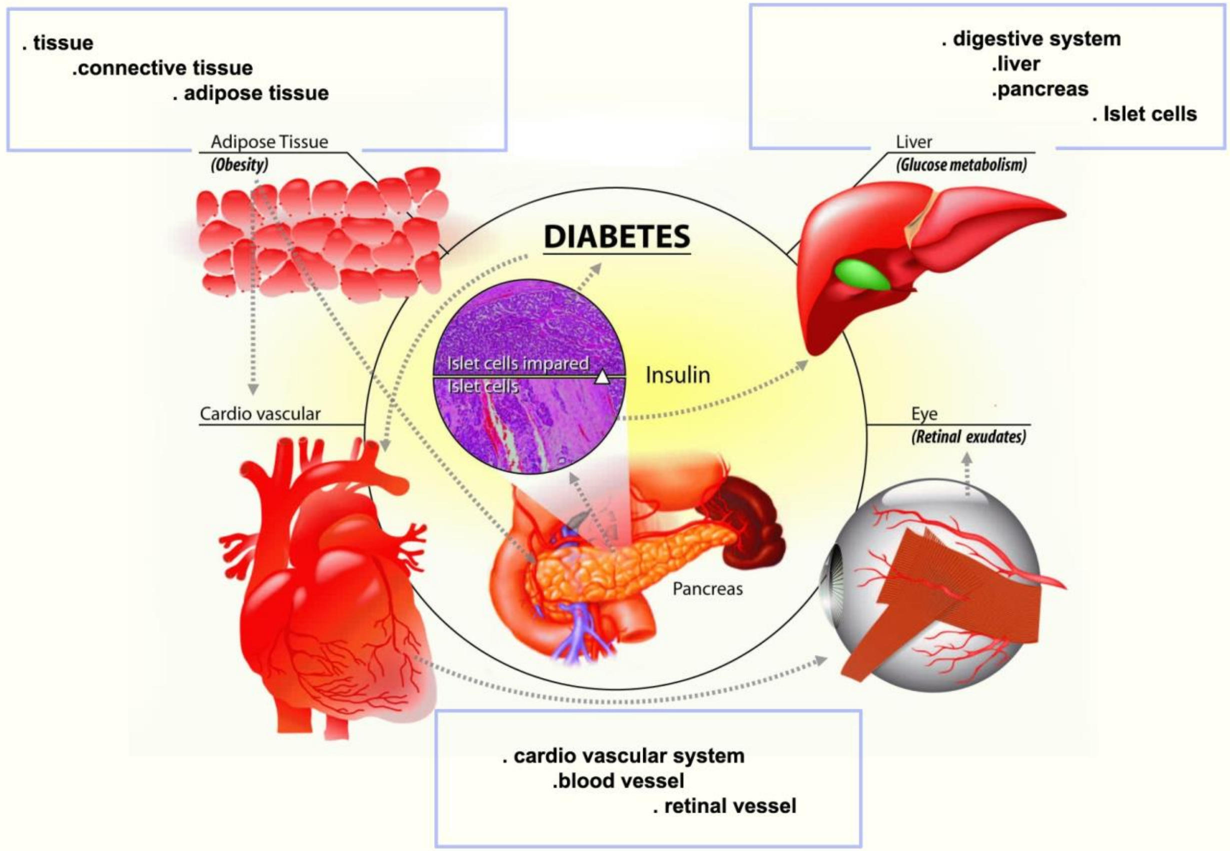
Results

We have 212,343 patients in our observational database. We have 7,000,000 clinical notes and reports and they have generated 750,000,000 SNOMED CT codes. Structured data are held in SQLServer^{TM} in OMOP / OHDSI format. The ontology codes such as in SNOMED CT are held in a Berkley DB, NOSQL database. The compositional expressions are held in Neo4J (a graph database) and also in Graph DB (a triple store). Our retrieval times for real clinical questions average between 2 and 3 seconds.

Observational Data are formatted for OMOP (OHDSI) and i2b2



Medical Ontology : Relationships between diseases, disorders, & systems, organs and tissues

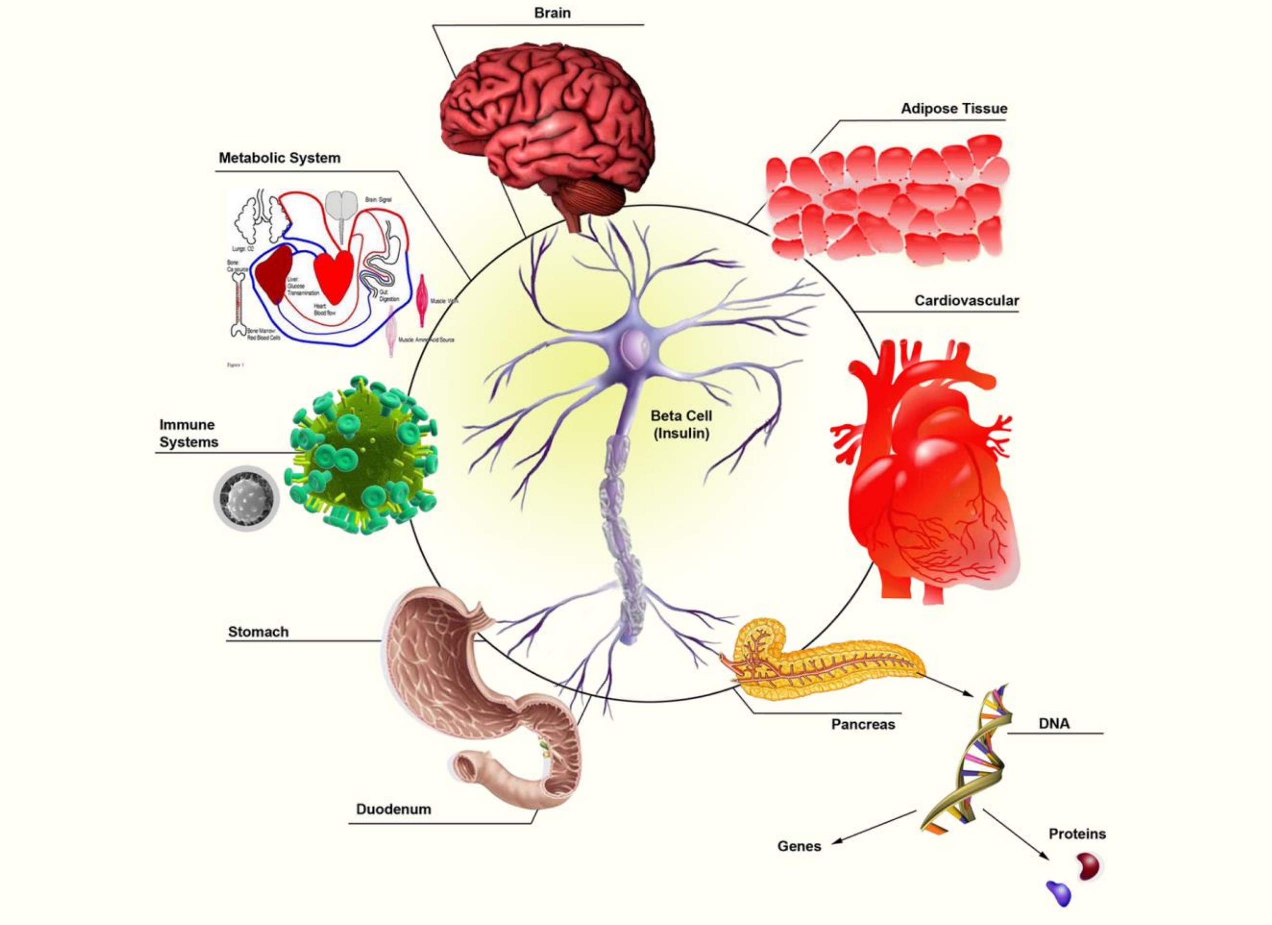


. tissue
.connective tissue
. adipose tissue

. digestive system
.liver
.pancreas
. Islet cells

. cardio vascular system
.blood vessel
. retinal vessel

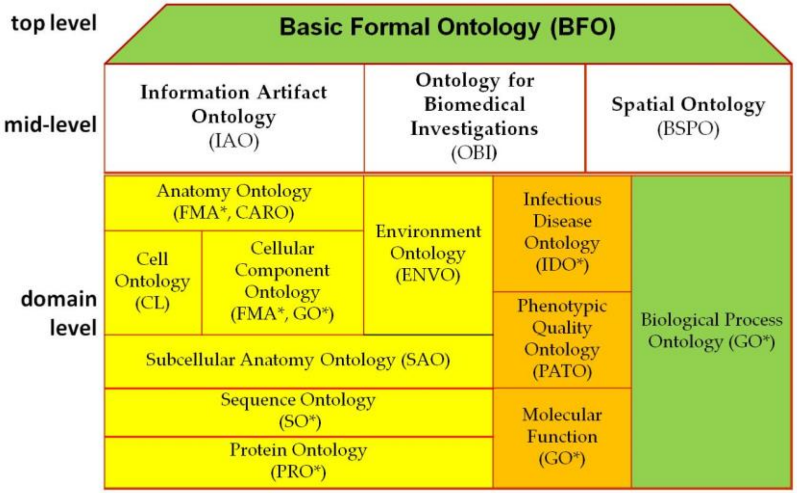
Biomedical Ontology : Neuronal interaction between diseases, systems, organs, substances, tissues, cells, proteins and genetics



Basic Formal Ontology (BFO)

Defines the high-level structures common to all domains

Connects → Health – Basic Science – Finance & Engineering



top level

Basic Formal Ontology (BFO)

mid-level

Information Artifact Ontology (IAO)

Ontology for Biomedical Investigations (OBI)

Spatial Ontology (BSPO)

domain level

Anatomy Ontology (FMA*, CARO)

Cell Ontology (CL)

Cellular Component Ontology (FMA*, GO*)

Environment Ontology (ENVO)

Infectious Disease Ontology (IDO*)

Phenotypic Quality Ontology (PATO)

Biological Process Ontology (GO*)

Subcellular Anatomy Ontology (SAO)

Sequence Ontology (SO*)

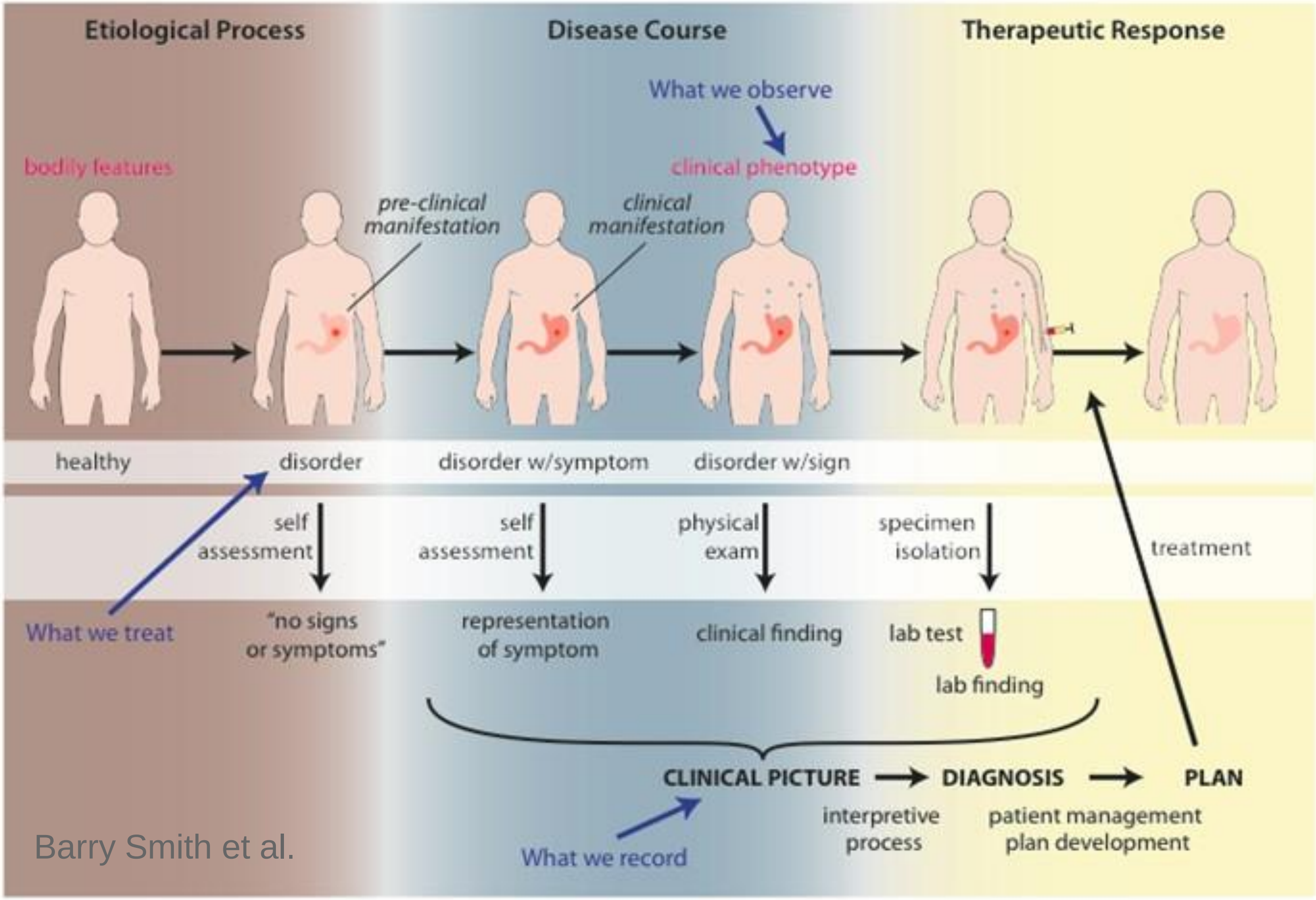
Molecular Function (GO*)

Protein Ontology (PRO*)

Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. Int J Med Inform. 2007 Dec;76 Suppl 3:S326-33.

- Cell Ontology (NHGRI, NIAID)
- eagle-i and VIVO (NCATS)
- Environment Ontology (GSC)
- Gene Ontology (NHGRI)
- IDO Infectious Disease Ontology (NIAID)
- Nanoparticle Ontology (PNNL)
- Ontology for Risks Against Patient Safety (EU)
- Ontology for Pain, Mental Health and Quality Of Life (NIDCR)
- Plant Ontology (NSF)
- Protein Ontology (NIGMS)
- Translational Medicine Ontology (W3C)
- US Army Biometrics Ontology (DOD)
- Vaccine Ontology (NHBLI)

Ontology of General Medical Sciences (OGMS)

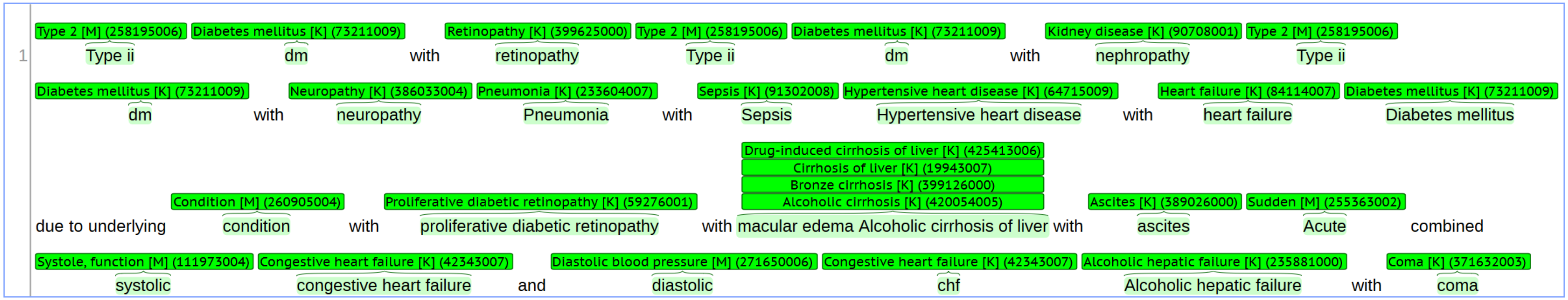


Barry Smith et al.

Level Three Ontology

- Fully Encoded Health Record
- Consistent with the Level One and Two Ontologies for Health
- Compositional Expressions are assigned Automagically
- Information is gathered through the usual documentation of patient care.
- Example…………….

SNOMED Codes:



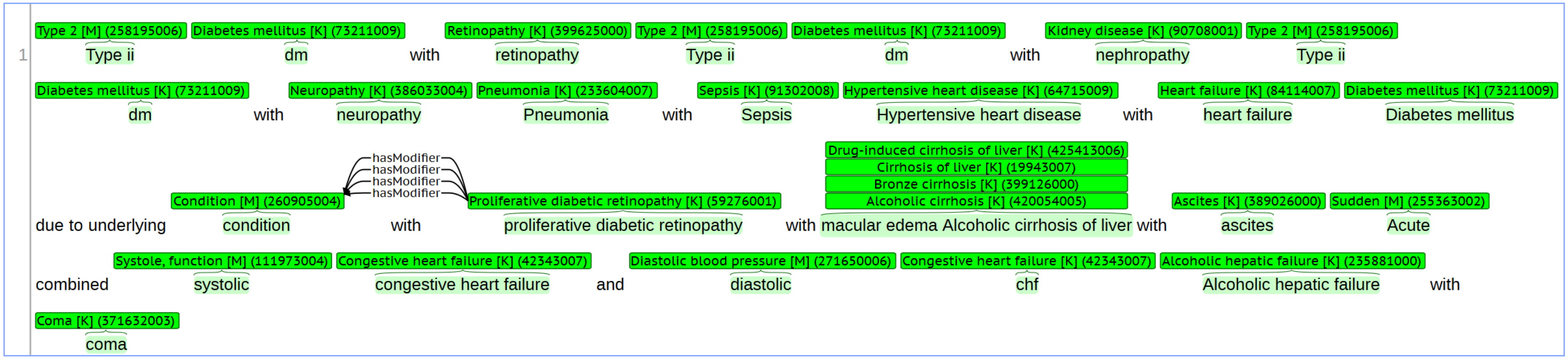
Type 2 [M] (258195006) Diabetes mellitus [K] (73211009) Retinopathy [K] (399625000) Type 2 [M] (258195006) Diabetes mellitus [K] (73211009) Kidney disease [K] (90708001) Type 2 [M] (258195006)
Type ii dm with retinopathy Type ii dm with nephropathy Type ii

Diabetes mellitus [K] (73211009) Neuropathy [K] (386033004) Pneumonia [K] (233604007) Sepsis [K] (91302008) Hypertensive heart disease [K] (64715009) Heart failure [K] (84114007) Diabetes mellitus [K] (73211009)
dm with neuropathy Pneumonia with Sepsis Hypertensive heart disease with heart failure Diabetes mellitus

Condition [M] (260905004) Proliferative diabetic retinopathy [K] (59276001) Drug-induced cirrhosis of liver [K] (425413006) Cirrhosis of liver [K] (19943007) Bronze cirrhosis [K] (399126000) Alcoholic cirrhosis [K] (420054005) Ascites [K] (389026000) Sudden [M] (255363002)
due to underlying condition with proliferative diabetic retinopathy with macular edema Alcoholic cirrhosis of liver with ascites Acute combined

Systole, function [M] (111973004) Congestive heart failure [K] (42343007) Diastolic blood pressure [M] (271650006) Congestive heart failure [K] (42343007) Alcoholic hepatic failure [K] (235881000) Coma [K] (371632003)
systolic congestive heart failure and diastolic chf Alcoholic hepatic failure with coma

Compositional Expressions:



Type 2 [M] (258195006) Diabetes mellitus [K] (73211009) Retinopathy [K] (399625000) Type 2 [M] (258195006) Diabetes mellitus [K] (73211009) Kidney disease [K] (90708001) Type 2 [M] (258195006)
Type ii dm with retinopathy Type ii dm with nephropathy Type ii

Diabetes mellitus [K] (73211009) Neuropathy [K] (386033004) Pneumonia [K] (233604007) Sepsis [K] (91302008) Hypertensive heart disease [K] (64715009) Heart failure [K] (84114007) Diabetes mellitus [K] (73211009)
dm with neuropathy Pneumonia with Sepsis Hypertensive heart disease with heart failure Diabetes mellitus

Condition [M] (260905004) Proliferative diabetic retinopathy [K] (59276001) Drug-induced cirrhosis of liver [K] (425413006) Cirrhosis of liver [K] (19943007) Bronze cirrhosis [K] (399126000) Alcoholic cirrhosis [K] (420054005) Ascites [K] (389026000) Sudden [M] (255363002)
due to underlying condition with proliferative diabetic retinopathy with macular edema Alcoholic cirrhosis of liver with ascites Acute

Systole, function [M] (111973004) Congestive heart failure [K] (42343007) Diastolic blood pressure [M] (271650006) Congestive heart failure [K] (42343007) Alcoholic hepatic failure [K] (235881000)
combined systolic congestive heart failure and diastolic chf Alcoholic hepatic failure with

Coma [K] (371632003)
coma

Case

HISTORY OF PRESENT ILLNESS:

#1 Chest pain

Patient is a 57-year old gentleman with a 80-pack-year smoking history. He has a family history of early coronary disease on his father's side, as his father had a heart attack at age 43. Patient does not exercise very much. He drinks 2 ounces of alcohol a day. He has type ii diabetes mellitus, hypertension, nor does he know his cholesterol level. Patient was in his usual state of health until 2 months ago when he began having exertional dyspnea and chest pain at peak exercise. Patient could walk 4 blocks and up 2 flights of stairs before he would have crushing substernal chest pain, which radiated to his left arm. On a scale of 0 to 10, it was as bad as 8 out of 10. Patient had some diaphoresis and dyspnea associated with the chest pain. He would sit down and this would be relieved after about 15 minutes. Patient has taken it upon himself to limit his activities based on this symptomatology. Patient has an interest in quitting smoking. He denies palpitations, syncope, pre-syncope, PND, or orthopnea. Patient has had no peripheral edema or shortness of breath at rest. He has had no episodes where the pain lasted greater than one-half hour.

#2 Right knee pain

Patient has had an 8-year history of right knee pain. Patient works as a construction worker and had a fork lift injury 8 years ago. Since that time, he has had more difficulty getting around on his right knee. It pops occasionally, but it never locks. It has not given out on him, but he has constant pain for which he takes ibuprofen on a regular basis. Patient used to be an avid golfer, but he has not been able to participate since the injury. This has also effected his work, as he has had difficulty climbing which is sometimes required in his profession.

#3 Nicotine dependence

Patient smokes a pack a day and has a 80-pack-year smoking history. He was smoking less than this until last year. Patient states his stress at work is the factor that has caused an increase in smoking, and he will be willing to see the Nicotine Dependence Center. In the past, he has tried to quit on his own without help of nicotine patches or any other nicotine replacement or Wellbutrin.

#4 Obesity

Patient is somewhat overweight and has had difficulty losing weight despite being a smoker. Patient has tried dieting and exercising programs, but since his inability to exercise with the right knee injury, he has had more difficulty with exercise and has not been able to lose weight. Patient states he watches his diet quite closely and has been limiting his caloric intake. To that end, he has actually lost 8 pounds over the last 6 months.

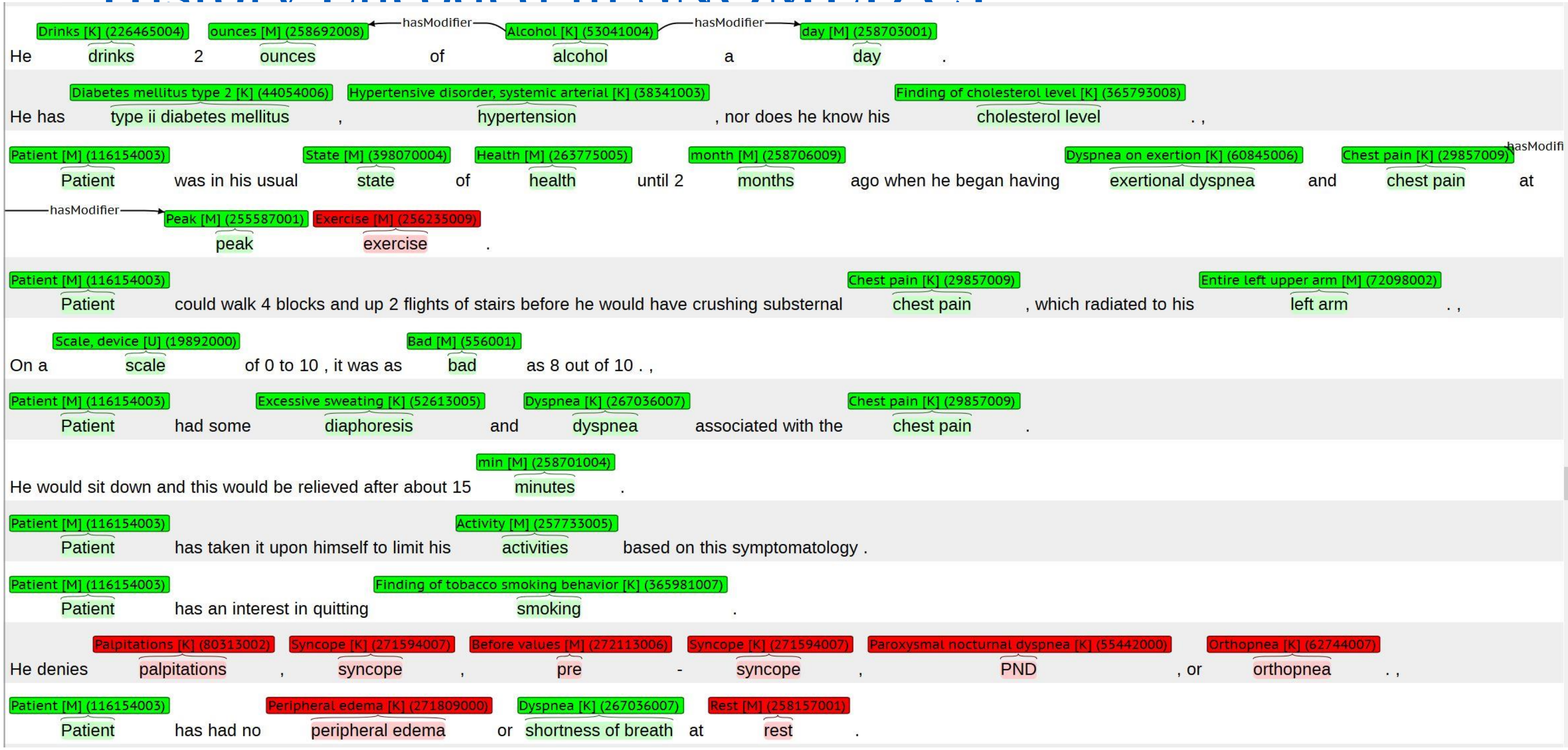
#5 Diabetes Mellitus Type ii

Patient denies polyuria and polydipsia however he is well controlled with Levemir Insulin 28 U SQ bid and Metformin 1000 mg bid. He has peripheral diabetic neuropathy, nephropathy and retinopathy.

Physical Examination (Relevant Sections)

- Extremities – Without clubbing, cyanosis, or edema. + Neuropathy with 3+/5+ loss of sensation in both feet to the ankle.
- Neuro – Cranial nerves 2 through 12 were intact. Visual fields were within normal limits. Pupils were equal and reactive to light and accomodation. Sensation was intact and bilaterally symmetric in his arms but a loss of sensation was found in his feet using a microfilliment examination. Motor was 5+/5+ bilaterally symmetric. Deep tendon reflexes were 2+/2+ and were symmetric bilaterally. Romberg was normal. Cerebellar signs were absent. Babinski was down going bilaterally.

History Encoded in SNOMED CT



Drinks [K] (226465004) ounces [M] (258692008) ←hasModifier— Alcohol [K] (53041004) —hasModifier→ day [M] (258703001)

He drinks 2 ounces of alcohol a day .

Diabetes mellitus type 2 [K] (44054006) Hypertensive disorder, systemic arterial [K] (38341003) Finding of cholesterol level [K] (365793008)

He has type ii diabetes mellitus , hypertension , nor does he know his cholesterol level . ,

Patient [M] (116154003) State [M] (398070004) Health [M] (263775005) month [M] (258706009) Dyspnea on exertion [K] (60845006) Chest pain [K] (29857009) hasModifi

Patient was in his usual state of health until 2 months ago when he began having exertional dyspnea and chest pain at

—hasModifier→ Peak [M] (255587001) Exercise [M] (256235009)

peak exercise .

Patient [M] (116154003) Chest pain [K] (29857009) Entire left upper arm [M] (72098002)

Patient could walk 4 blocks and up 2 flights of stairs before he would have crushing substernal chest pain , which radiated to his left arm . ,

Scale, device [U] (19892000) Bad [M] (556001)

On a scale of 0 to 10 , it was as bad as 8 out of 10 . ,

Patient [M] (116154003) Excessive sweating [K] (52613005) Dyspnea [K] (267036007) Chest pain [K] (29857009)

Patient had some diaphoresis and dyspnea associated with the chest pain .

min [M] (258701004)

He would sit down and this would be relieved after about 15 minutes .

Patient [M] (116154003) Activity [M] (257733005)

Patient has taken it upon himself to limit his activities based on this symptomatology .

Patient [M] (116154003) Finding of tobacco smoking behavior [K] (365981007)

Patient has an interest in quitting smoking .

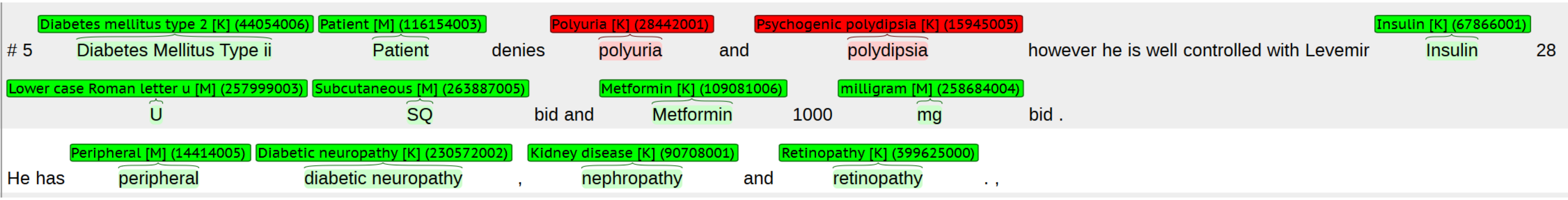
Palpitations [K] (80313002) Syncope [K] (271594007) Before values [M] (272113006) Syncope [K] (271594007) Paroxysmal nocturnal dyspnea [K] (55442000) Orthopnea [K] (62744007)

He denies palpitations , syncope , pre - syncope , PND , or orthopnea . ,

Patient [M] (116154003) Peripheral edema [K] (271809000) Dyspnea [K] (267036007) Rest [M] (258157001)

Patient has had no peripheral edema or shortness of breath at rest .

History



Diabetes mellitus type 2 [K] (44054006) Patient [M] (116154003) Polyuria [K] (28442001) Psychogenic polydipsia [K] (15945005) Insulin [K] (67866001)

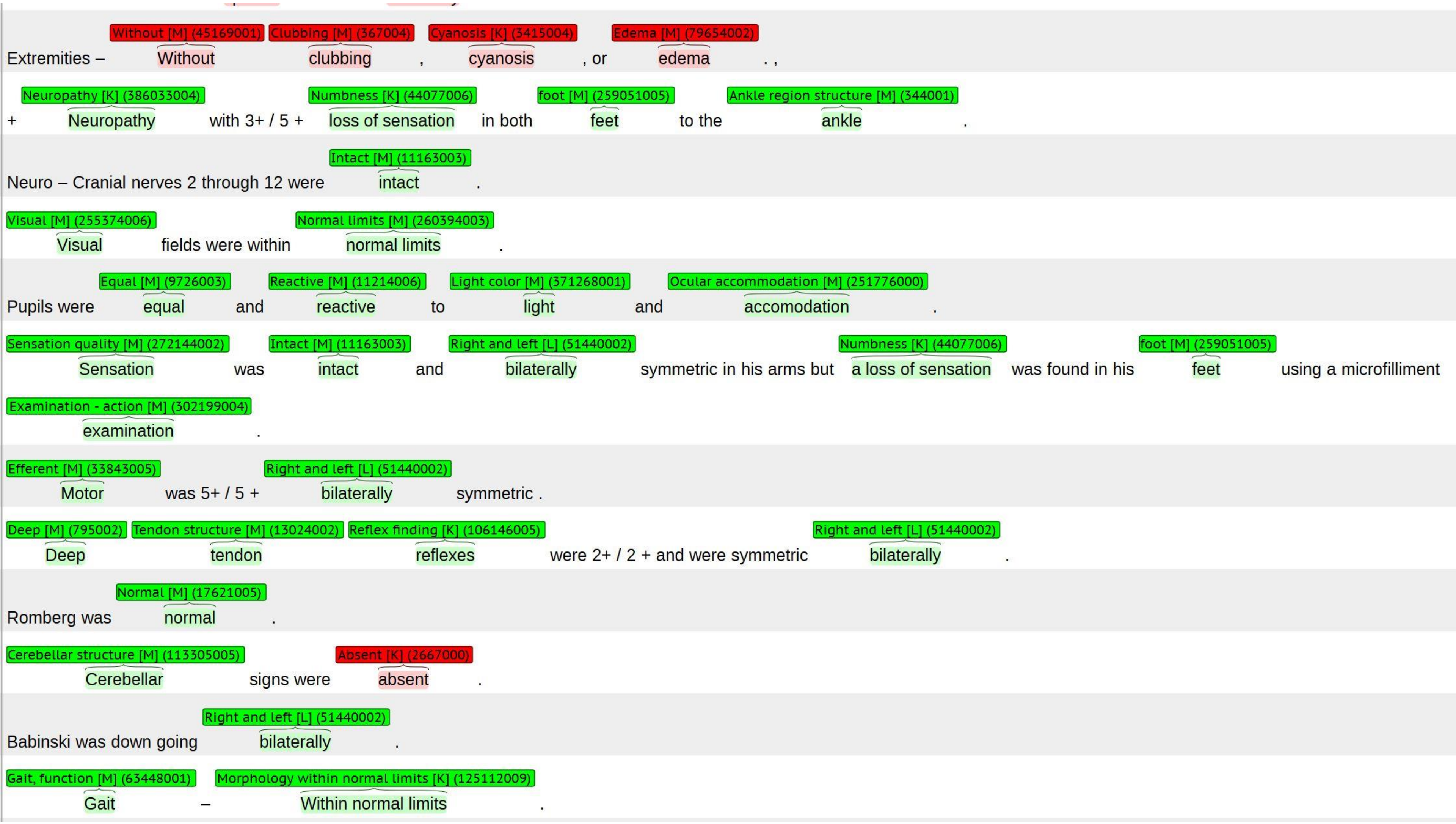
5 Diabetes Mellitus Type ii Patient denies polyuria and polydipsia however he is well controlled with Levemir Insulin 28

Lower case Roman letter u [M] (257999003) Subcutaneous [M] (263887005) Metformin [K] (109081006) milligram [M] (258684004)

U SQ bid and Metformin 1000 mg bid .

Peripheral [M] (14414005) Diabetic neuropathy [K] (230572002) Kidney disease [K] (90708001) Retinopathy [K] (399625000)

He has peripheral diabetic neuropathy , nephropathy and retinopathy . ,



Without [M] (45169001) Clubbing [M] (367004) Cyanosis [K] (3415004) Edema [M] (79654002)
Extremities – Without clubbing , cyanosis , or edema . ,

Neuropathy [K] (386033004) Numbness [K] (44077006) foot [M] (259051005) Ankle region structure [M] (344001)
+ Neuropathy with 3+ / 5 + loss of sensation in both feet to the ankle .

Intact [M] (11163003)
Neuro – Cranial nerves 2 through 12 were intact .

Visual [M] (255374006) Normal limits [M] (260394003)
Visual fields were within normal limits .

Equal [M] (9726003) Reactive [M] (11214006) Light color [M] (371268001) Ocular accommodation [M] (251776000)
Pupils were equal and reactive to light and accomodation .

Sensation quality [M] (272144002) Intact [M] (11163003) Right and left [L] (51440002) Numbness [K] (44077006) foot [M] (259051005)
Sensation was intact and bilaterally symmetric in his arms but a loss of sensation was found in his feet using a microfilliment

Examination - action [M] (302199004)
examination .

Efferent [M] (33843005) Right and left [L] (51440002)
Motor was 5+ / 5 + bilaterally symmetric .

Deep [M] (795002) Tendon structure [M] (13024002) Reflex finding [K] (106146005) Right and left [L] (51440002)
Deep tendon reflexes were 2+ / 2 + and were symmetric bilaterally .

Normal [M] (17621005)
Romberg was normal .

Cerebellar structure [M] (113305005) Absent [K] (2667000)
Cerebellar signs were absent .

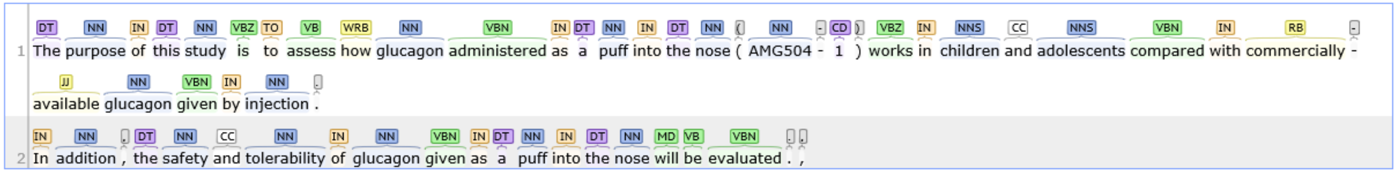
Right and left [L] (51440002)
Babinski was down going bilaterally .

Gait, function [M] (63448001) Morphology within normal limits [K] (125112009)
Gait – Within normal limits .

Assessment of Intranasal Glucagon in Children and Adolescents With Type 1 Diabetes

The purpose of this study is to assess how glucagon administered as a puff into the nose (AMG504-1) works in children and adolescents compared with commercially-available glucagon given by injection. In addition, the safety and tolerability of glucagon given as a puff into the nose will be evaluated.

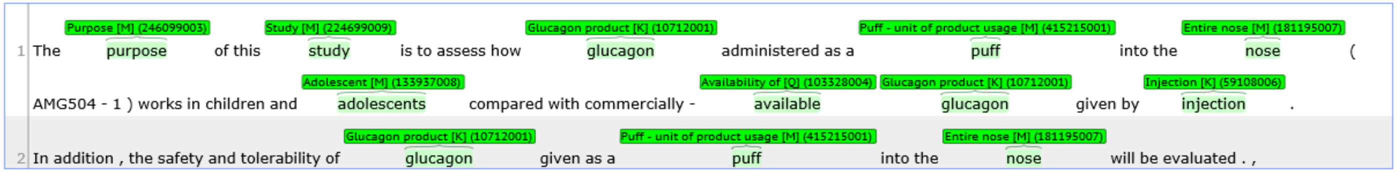
Part-of-Speech:


DT NN IN DT NN VBZ TO VB WRB NN VBN IN DT NN IN DT NN (NN - CD) VBZ IN NNS CC NNS VBN IN RB -
JJ NN VBN IN NN .
IN NN , DT NN CC NN IN NN VBN IN DT NN IN DT NN MD VB VBN . ,

1 The purpose of this study is to assess how glucagon administered as a puff into the nose (AMG504 - 1) works in children and adolescents compared with commercially - available glucagon given by injection .

2 In addition , the safety and tolerability of glucagon given as a puff into the nose will be evaluated . ,

SNOMED Codes:


Purpose [M] (246099003)
Study [M] (224699009)
Glucagon product [K] (10712001)
Puff - unit of product usage [M] (415215001)
Entire nose [M] (181195007)
Adolescent [M] (133937008)
Availability of [Q] (103328004)
Glucagon product [K] (10712001)
Injection [K] (59108006)
Glucagon product [K] (10712001)
Puff - unit of product usage [M] (415215001)
Entire nose [M] (181195007)

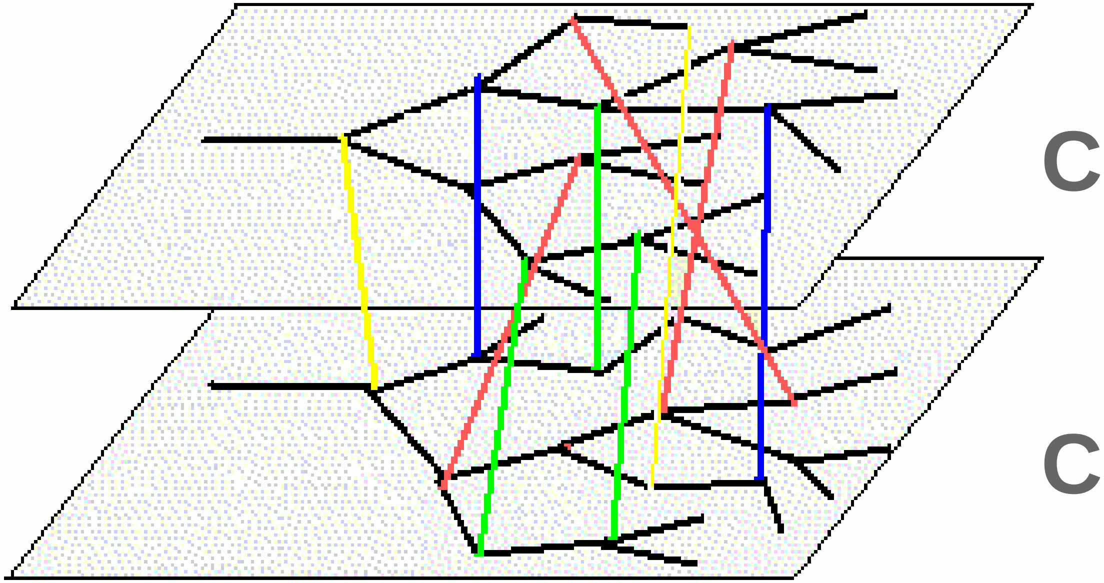
1 The purpose of this study is to assess how glucagon administered as a puff into the nose (AMG504 - 1) works in children and adolescents compared with commercially - available glucagon given by injection .

2 In addition , the safety and tolerability of glucagon given as a puff into the nose will be evaluated . ,

Rational Knowledge Representation

- Cellulitis of the left foot with Osteomyelitis of the Third metatarsal without Lymphangitis

- [AND]
  - [WITH]
    - Cellulitis (disorder) [128045006]
      - [has Finding Site]
        - Entire foot (body structure) [302545001]
          - [has Laterality]
            - Left (qualifier value) [7771000]
    - Osteomyelitis (disorder) [60168000]
      - [has Finding Site]
        - Entire third metatarsal (body structure) [182134006]
  - [WITHOUT]
    - Lymphangitis (disorder) [1415005]

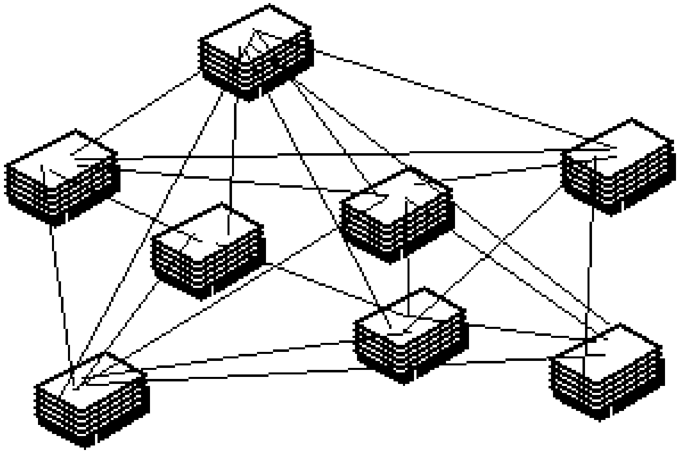


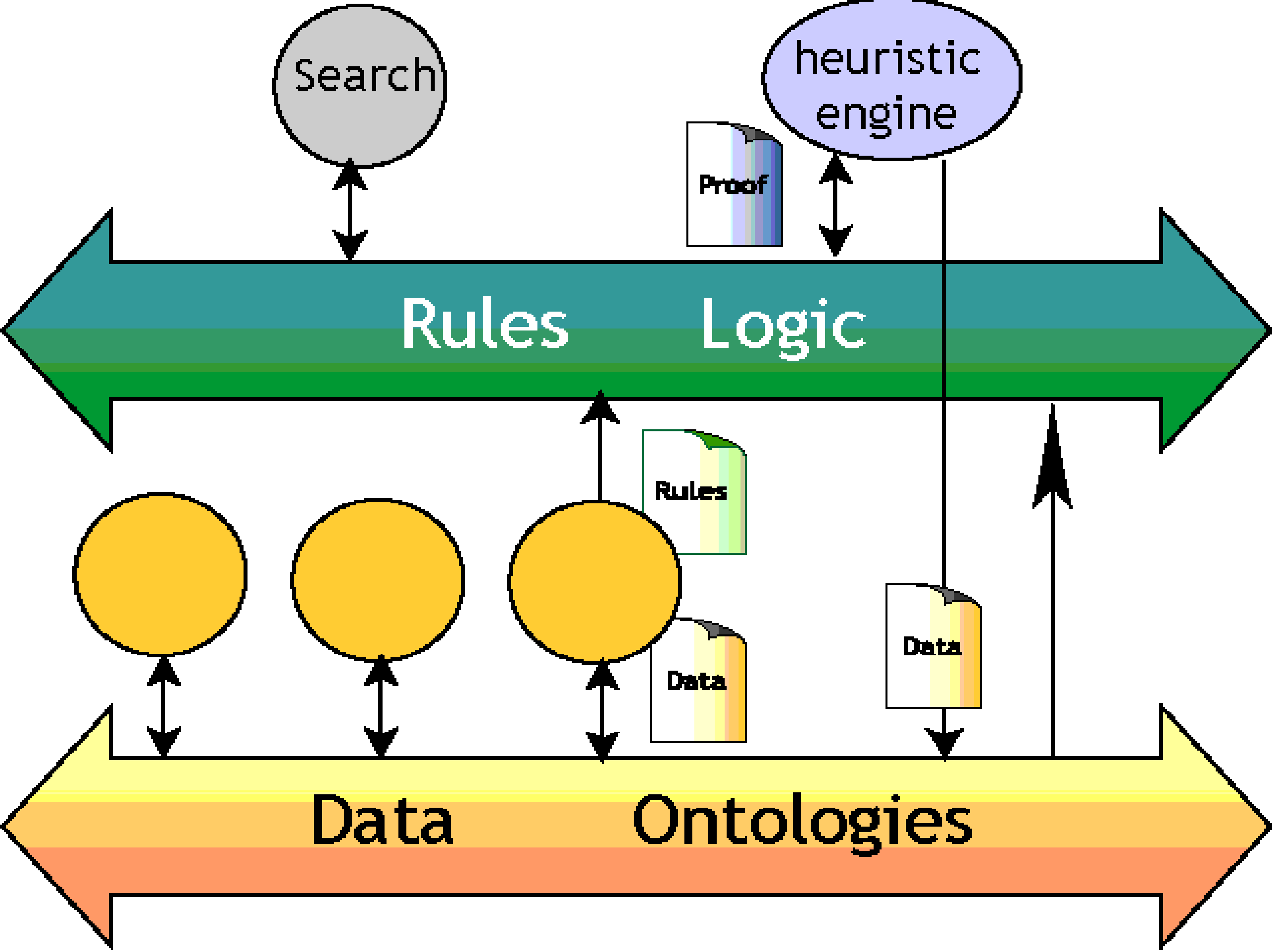
Semantic Network

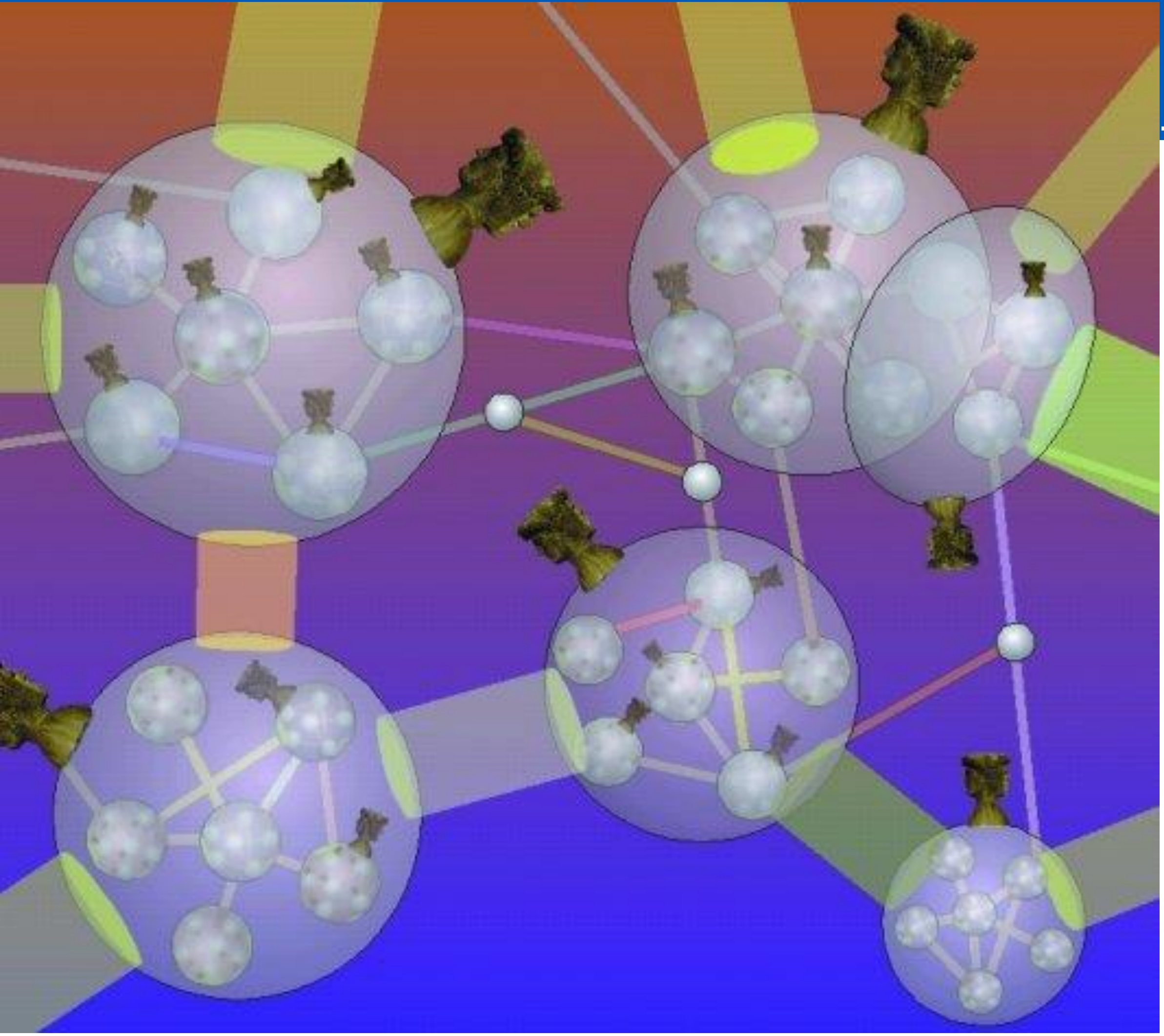
Case One

Case Two

Multi-Center Data Sharing and Interchange

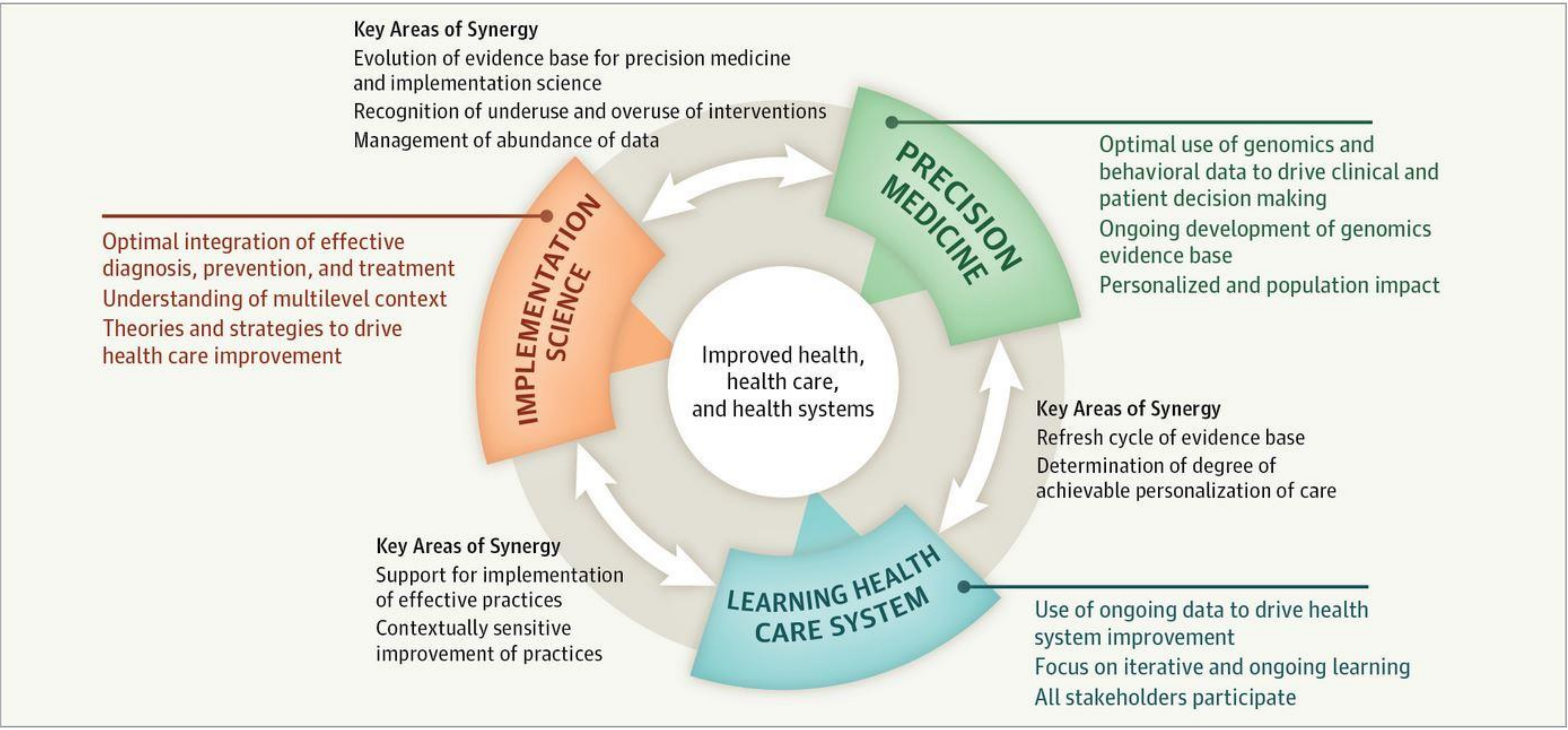



Search
heuristic engine
Proof
Rules
Logic
Rules
Data
Data
Data
Ontologies



Intelligent Agents

The Evolution of Healthcare



Key Areas of Synergy

Evolution of evidence base for precision medicine and implementation science

Recognition of underuse and overuse of interventions

Management of abundance of data

Optimal use of genomics and behavioral data to drive clinical and patient decision making

Ongoing development of genomics evidence base

Personalized and population impact

Optimal integration of effective diagnosis, prevention, and treatment

Understanding of multilevel context

Theories and strategies to drive health care improvement

Key Areas of Synergy

Refresh cycle of evidence base

Determination of degree of achievable personalization of care

Key Areas of Synergy

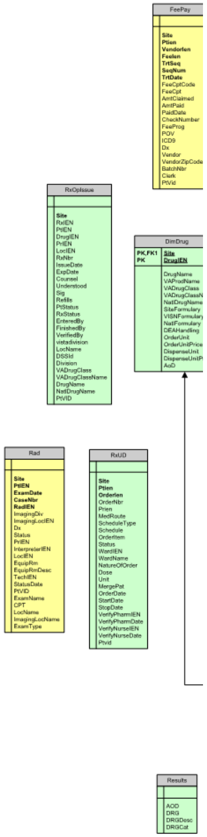
Support for implementation of effective practices

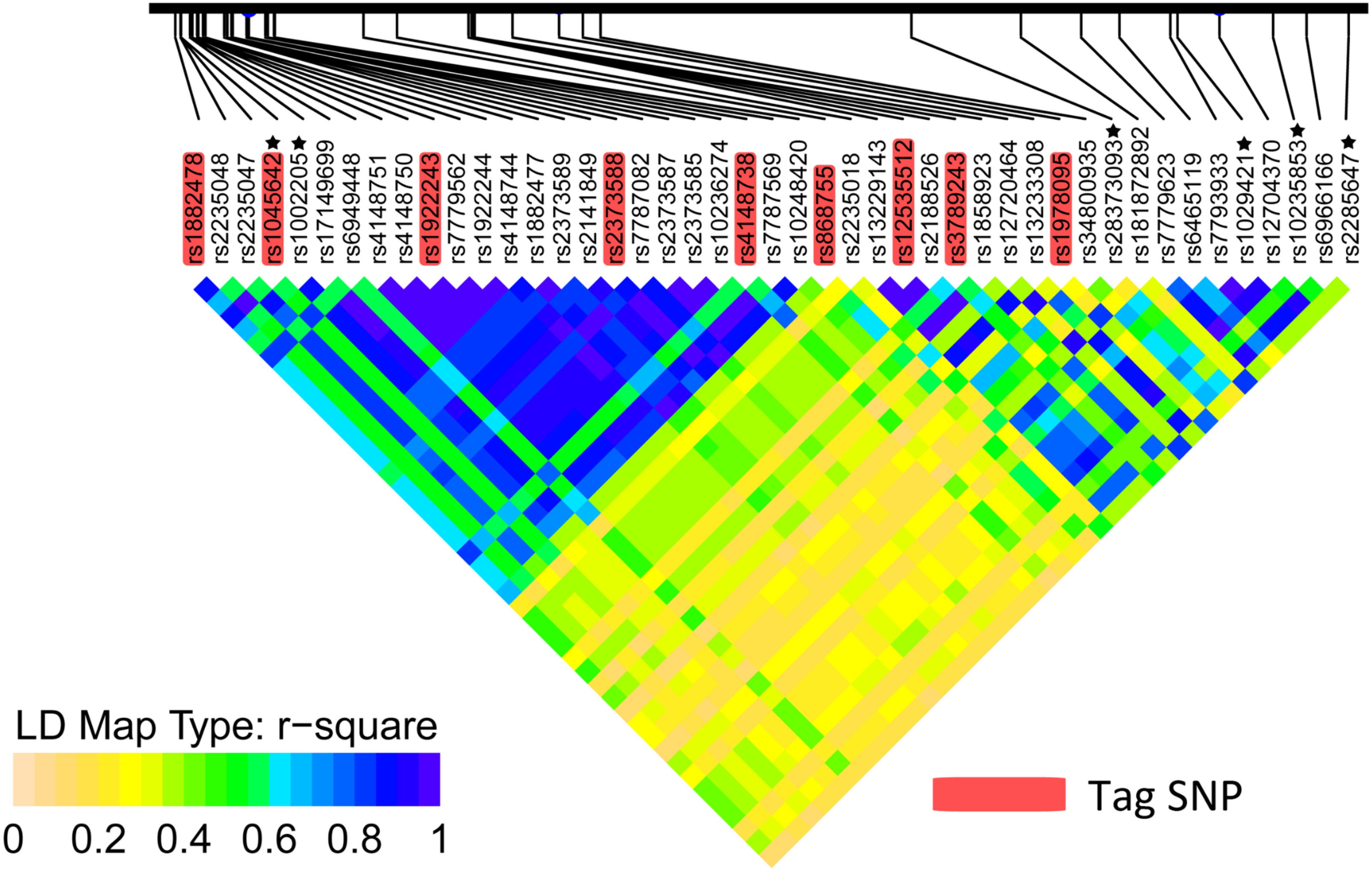
Contextually sensitive improvement of practices

Use of ongoing data to drive health system improvement

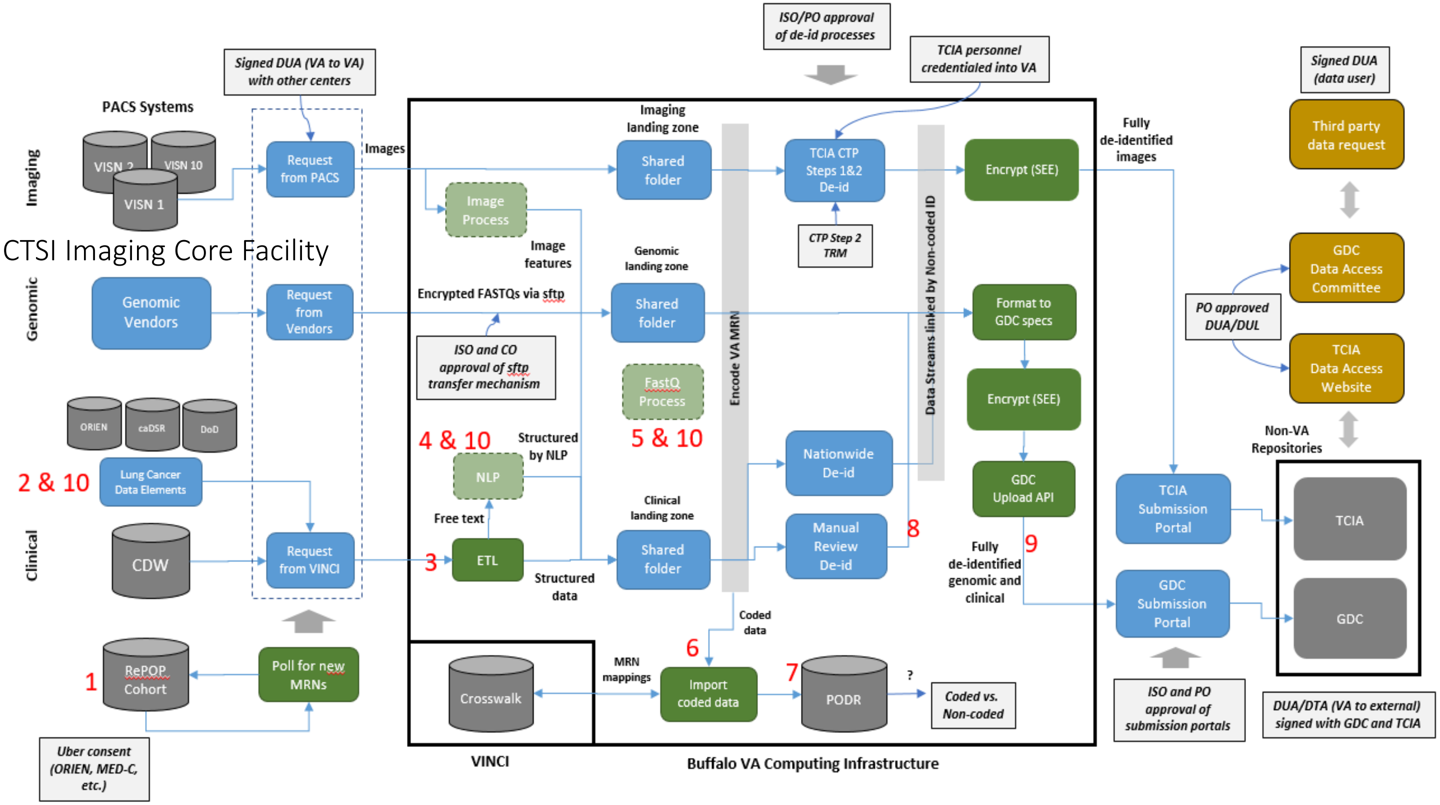
Focus on iterative and ongoing learning

All stakeholders participate


FeePay
RxOpIssue
DimDrug
Rad
RxUD
Results


rs1882478
rs2235048
rs2235047
rs1045642
rs1002205
rs17149699
rs6949448
rs4148751
rs4148750
rs1922243
rs7779562
rs1922244
rs4148744
rs1882477
rs2373589
rs2141849
rs2373588
rs7787082
rs2373587
rs2373585
rs10236274
rs4148738
rs7787569
rs10248420
rs868755
rs2235018
rs13229143
rs12535512
rs2188526
rs3789243
rs1858923
rs12720464
rs13233308
rs1978095
rs34800935
rs28373093
rs181872892
rs7779623
rs6465119
rs7793933
rs10294211
rs12704370
rs10235853
rs6966166
rs2285647
LD Map Type: r−square
0
0.2
0.4
0.6
0.8
1
Tag SNP

CTSI Biomedical Informatics Core Facility Architecture



Signed DUA (VA to VA) with other centers

PACS Systems

Imaging

VISN 2

VISN 10

VISN 1

Request from PACS

Images

CTSI Imaging Core Facility

Genomic

Genomic Vendors

Request from Vendors

ORIEN

caDSR

DoD

2 & 10

Lung Cancer Data Elements

Clinical

CDW

Request from VINCI

1

RePOP Cohort

Poll for new MRNs

Uber consent (ORIEN, MED-C, etc.)

ISO/PO approval of de-id processes

TCIA personnel credentialed into VA

Imaging landing zone

Shared folder

Image Process

Image features

TCIA CTP Steps 1&2 De-id

CTP Step 2 TRM

Encrypt (SEE)

Genomic landing zone

Encrypted FASTQs via sftp

Shared folder

ISO and CO approval of sftp transfer mechanism

FastQ Process

Format to GDC specs

Encrypt (SEE)

Encode VA MRN

Data Streams linked by Non-coded ID

4 & 10

Structured by NLP

5 & 10

NLP

Nationwide De-id

GDC Upload API

Free text

Clinical landing zone

3

ETL

Shared folder

Manual Review De-id

8

9

Structured data

Fully de-identified genomic and clinical

Coded data

Crosswalk

MRN mappings

6

Import coded data

7

PODR

?

Coded vs. Non-coded

VINCI

Buffalo VA Computing Infrastructure

Fully de-identified images

Signed DUA (data user)

Third party data request

GDC Data Access Committee

PO approved DUA/DUL

TCIA Data Access Website

Non-VA Repositories

TCIA Submission Portal

TCIA

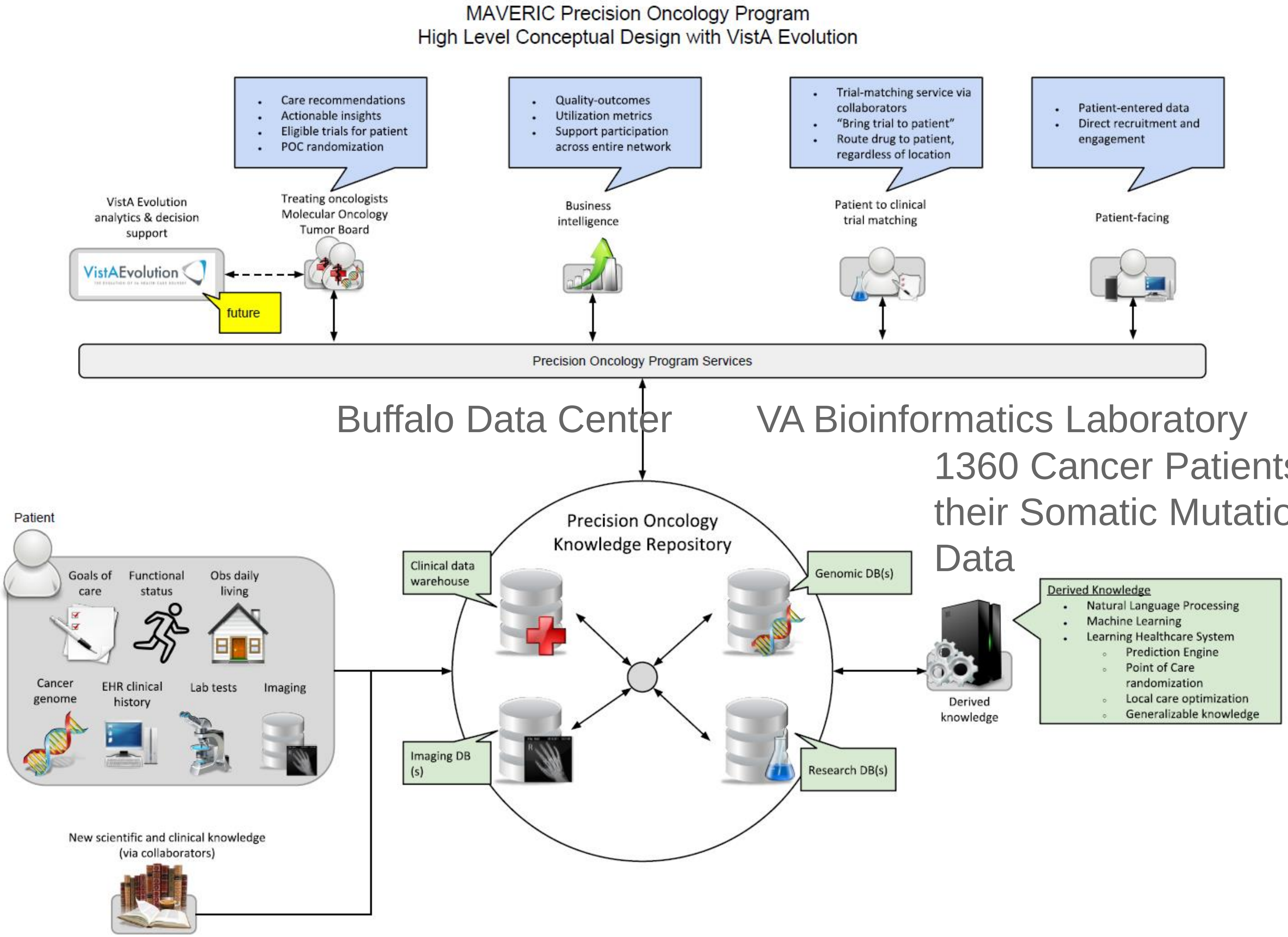
GDC Submission Portal

GDC

ISO and PO approval of submission portals

DUA/DTA (VA to external) signed with GDC and TCIA

Precision Oncology (POP) – Big Picture



MAVERIC Precision Oncology Program
High Level Conceptual Design with VistA Evolution

- Care recommendations
- Actionable insights
- Eligible trials for patient
- POC randomization

- Quality-outcomes
- Utilization metrics
- Support participation across entire network

- Trial-matching service via collaborators
- "Bring trial to patient"
- Route drug to patient, regardless of location

- Patient-entered data
- Direct recruitment and engagement

VistA Evolution analytics & decision support
VistAEvolution
future
Treating oncologists Molecular Oncology Tumor Board
Business intelligence
Patient to clinical trial matching
Patient-facing
Precision Oncology Program Services
Patient
Goals of care
Functional status
Obs daily living
Cancer genome
EHR clinical history
Lab tests
Imaging
New scientific and clinical knowledge (via collaborators)
Precision Oncology Knowledge Repository
Clinical data warehouse
Genomic DB(s)
Imaging DB (s)
Research DB(s)
Derived knowledge

Derived Knowledge
- Natural Language Processing
- Machine Learning
- Learning Healthcare System
  - Prediction Engine
  - Point of Care randomization
  - Local care optimization
  - Generalizable knowledge

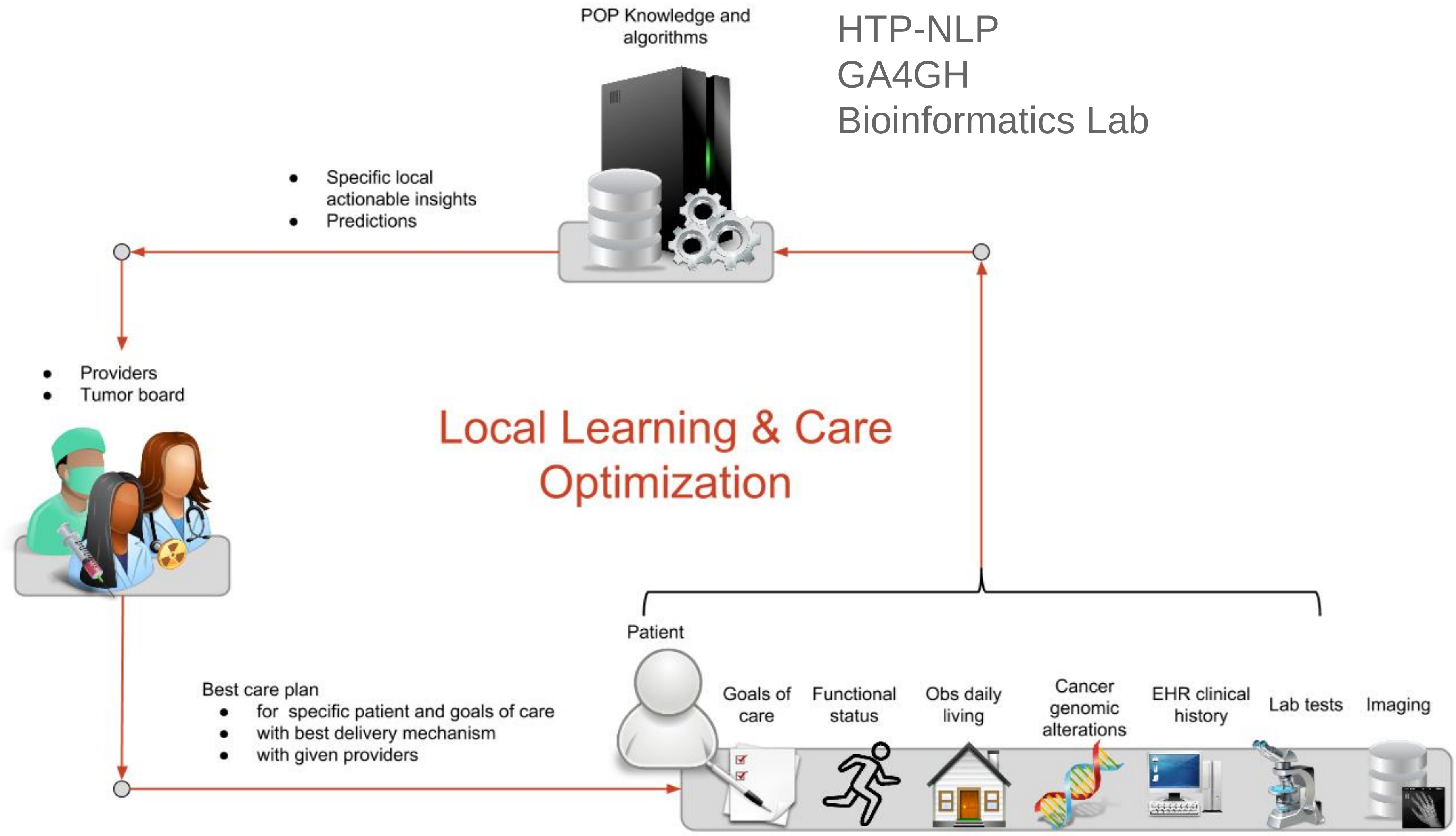
Buffalo Data Center

VA Bioinformatics Laboratory

1360 Cancer Patients with their Somatic Mutation Data

Learning Healthcare System Model

HTP-NLP
GA4GH
Bioinformatics Lab



POP Knowledge and algorithms

- Specific local actionable insights
- Predictions

- Providers
- Tumor board

Local Learning & Care Optimization

Best care plan

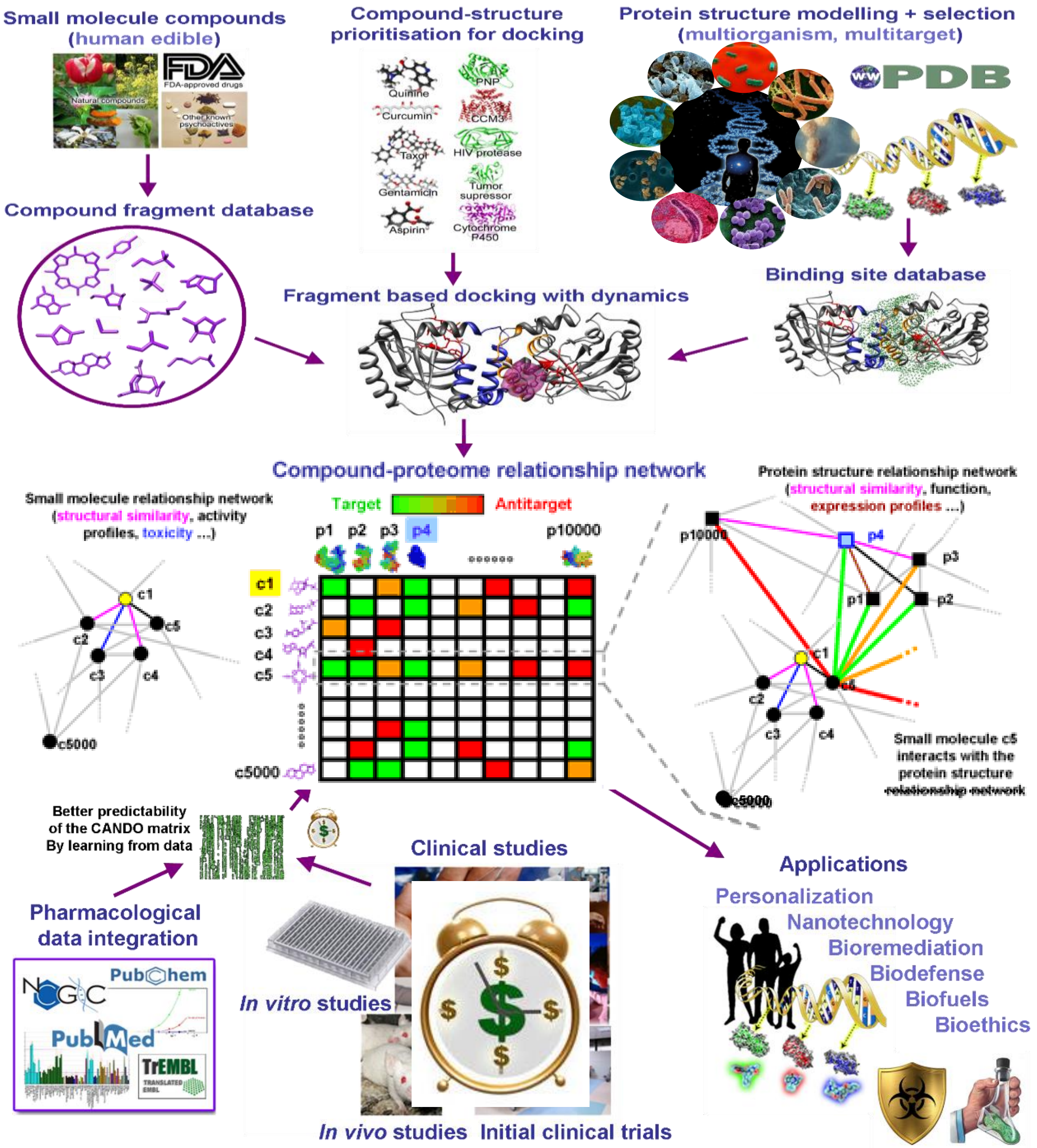
- for specific patient and goals of care
- with best delivery mechanism
- with given providers

SHOTGUN MULTITARGET DRUG DISCOVERY PIPELINE

Knowledge based fragment docking with dynamics

Systems based multitarget drug discovery

Prospective validation followed by clinical studies, other applications



Small molecule compounds (human edible)

Compound-structure prioritisation for docking

Protein structure modelling + selection (multiorganism, multitarget)

Compound fragment database

Fragment based docking with dynamics

Binding site database

Compound-proteome relationship network

Small molecule relationship network (structural similarity, activity profiles, toxicity ...)

Protein structure relationship network (structural similarity, function, expression profiles ...)

Small molecule c5 interacts with the protein structure relationship network

Better predictability of the CANDO matrix By learning from data

Pharmacological data integration

Clinical studies

In vitro studies

In vivo studies

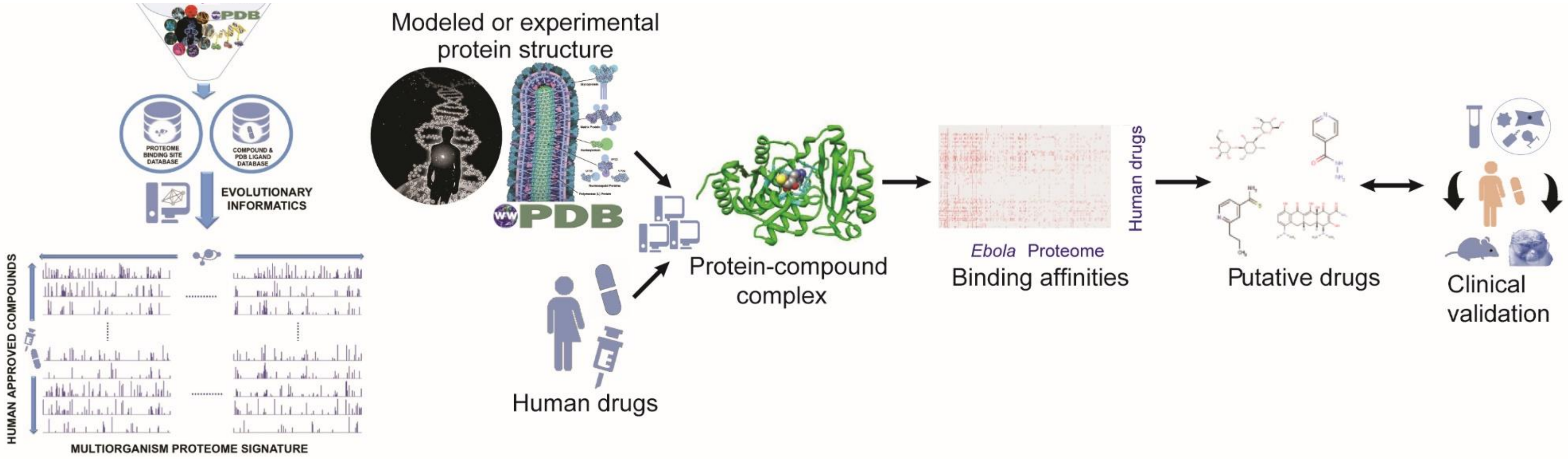
Initial clinical trials

Applications

Personalization
Nanotechnology
Bioremediation
Biodefense
Biofuels
Bioethics

CANDO PROSPECTIVE VALIDATIONS

Indication	Putative primary cause	Validations (total)	Hit rate (current) [★ = *in vivo*]	Source / Collaborator	Reference (or TBP)
Diabetes mellitus type 1	autoimmune, genetic	10	1/1 ★	Gaurav Chopra, UCSF	TBP
Dental caries	*S. mutans*	10	10/10	Jeremy Horst, UCSF	[5, 22], TBP
Dengue fever	Dengue virus	31	11/27	Scott Michael, FGCU	TBP
Herpes	HSV, CMV, KSHV (all)	29	6/29	Michael Lagunoff, UW; ImQuest Biosciences, Inc.	TBP
MDR Tuberculosis	*M. tuberculosis*	17	4/8	Michael Strong, NJHC	TBP
Systemic lupus erythematosus	autoimmune	≈20	1/1	Keith Elkon, UW	TBP
PB cirrhosis	HBRV	≈20	12 / 12	Andrew Mason, U. Alberta	TBP
Hepatitis B	Hepatitis B virus	31	3 / 31	ImQuest Biosciences, Inc.	[14], TBP
Flu	Influenza A virus	24	0 / 24	ImQuest Biosciences, Inc.	[14], TBP
AIDS	HIV 1 & 2	≈40	ongoing	James Mullins, UW	
Diabetes mellitus type 2	metabolic, genetic	≈80	ongoing	Jay Heinecke, UW	
Cholangiocarcinoma	neoplastic disorder	40	ongoing	Natini Jinawath, Ramathibodi Hospital, Thailand	
Ebola hemorrhagic fever	Ebola virus	≈40	ongoing	Michael Katze, UW	
Flu	Influenza viruses	≈40	ongoing	various	
Hepatitis C	Hepatitis C virus	≈20	ongoing	Lorne Tyrell, U. Alberta	
MDR Tuberculosis	*M. tuberculosis*	40	ongoing	Prasit Palittapongarnpim, Mahidol U, Thailand	
Soft tissue infections	*P. aeruginosa*	≈40	ongoing	Pradeep Singh, UW	
Yellow fever	Yellow fever virus	≈20	ongoing	Scott Michael, FGCU	



UPDATE: 58/163 (~36%) across 12 studies and 10 indications; first failure with infuenza.

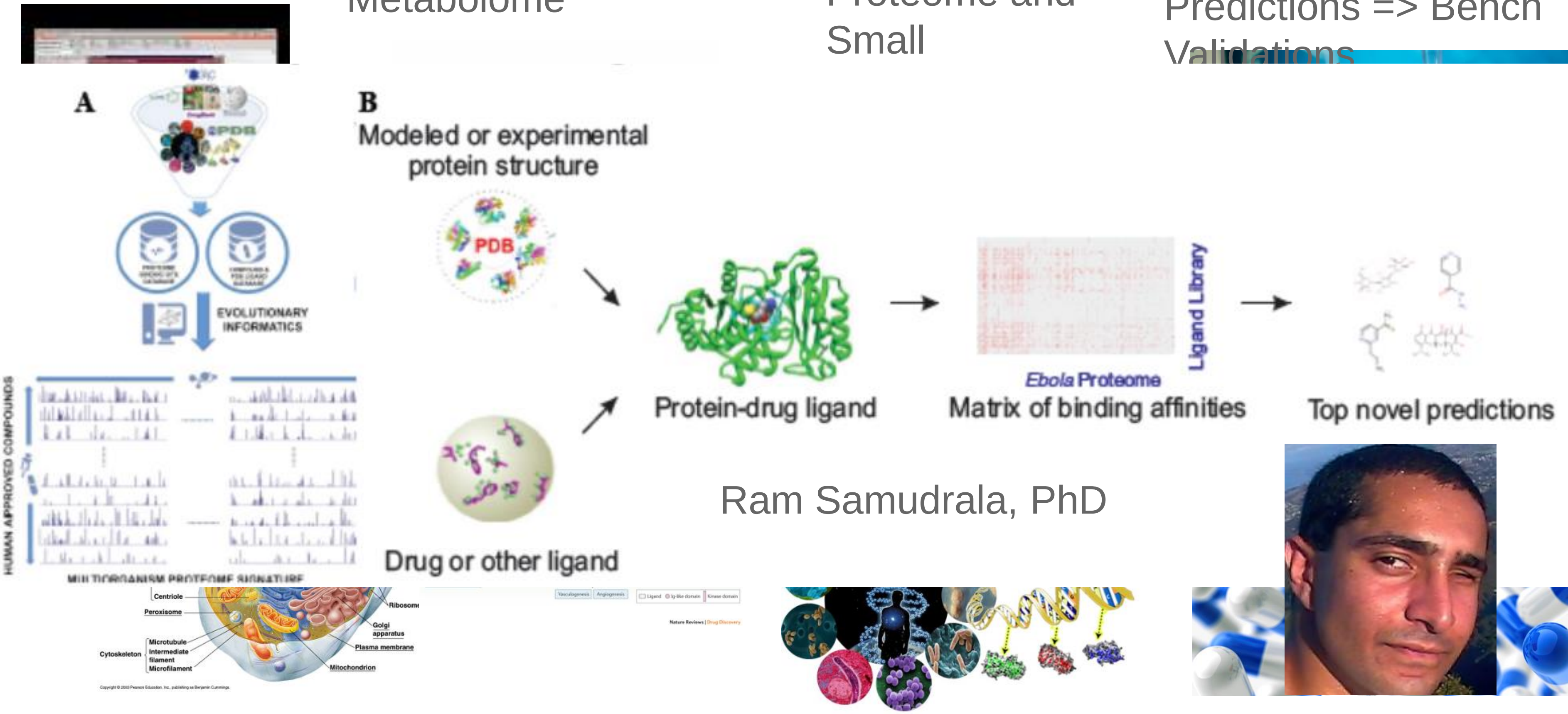
HTP-NLP & CANDO / CANDOCK

Clinical

Functional: Metabolome

Structural: Proteome and Small

Structure and Function = Accurate Predictions => Bench Validations

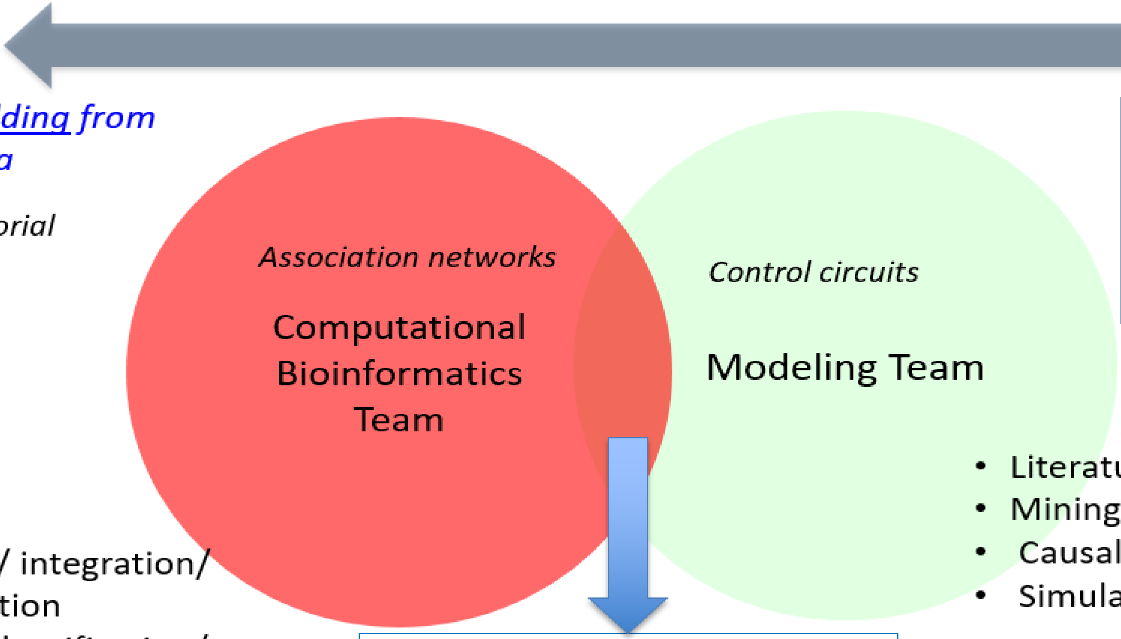


Ram Samudrala, PhD

Computational to Validation Components

Empirical / data-driven

Mechanistic/ first principles



Building from data

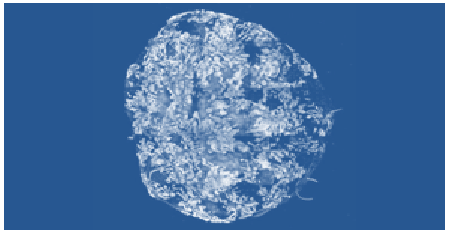
Combinatorial profiling

Association networks

Computational Bioinformatics Team

Control circuits

Modeling Team



Lung Organoid

- Literature Based Discovery
- Mining basic principles (NLP)
- Causal regulatory dynamics
- Simulation/ optimal trial design

- Data mgt/ integration/ normalization
- Feature identification/ pattern matching
- Network Generation
- Statistical inference / diagnostic classification

Evaluation of Candidate Drugs

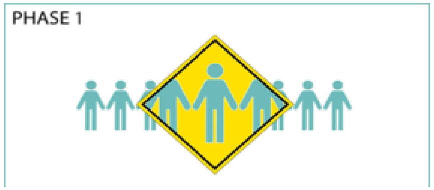
1. Cell lines

2. Organoids/Spheroids

3. PDX tumors

Future

Phase I Clinical Trials



Healthcare Value

- Value = Quality / Cost
- Quality is composed of:
  - Outcomes
  - Safety
  - Service
    - Reliability

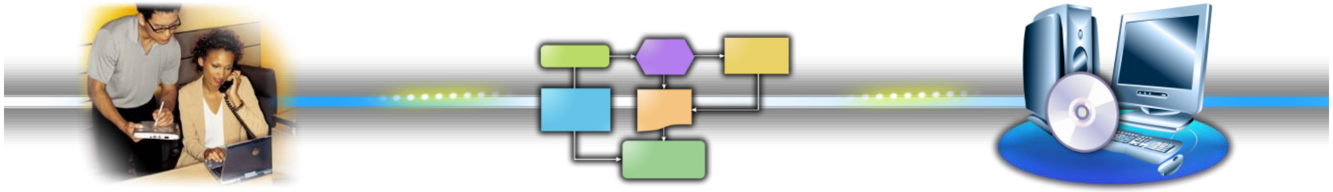
Measuring Strategic Performance

"You can't manage what you can't measure. You can't measure what you can't describe"

Robert Kaplan and David Norton

Authors of "The Balanced Scorecard"

Framework that **aligns the entire organization** to what is **important to the customer**, allowing the organization to **excel** at the **critical activities** and **reduce time spent** on the **things that don’t matter**



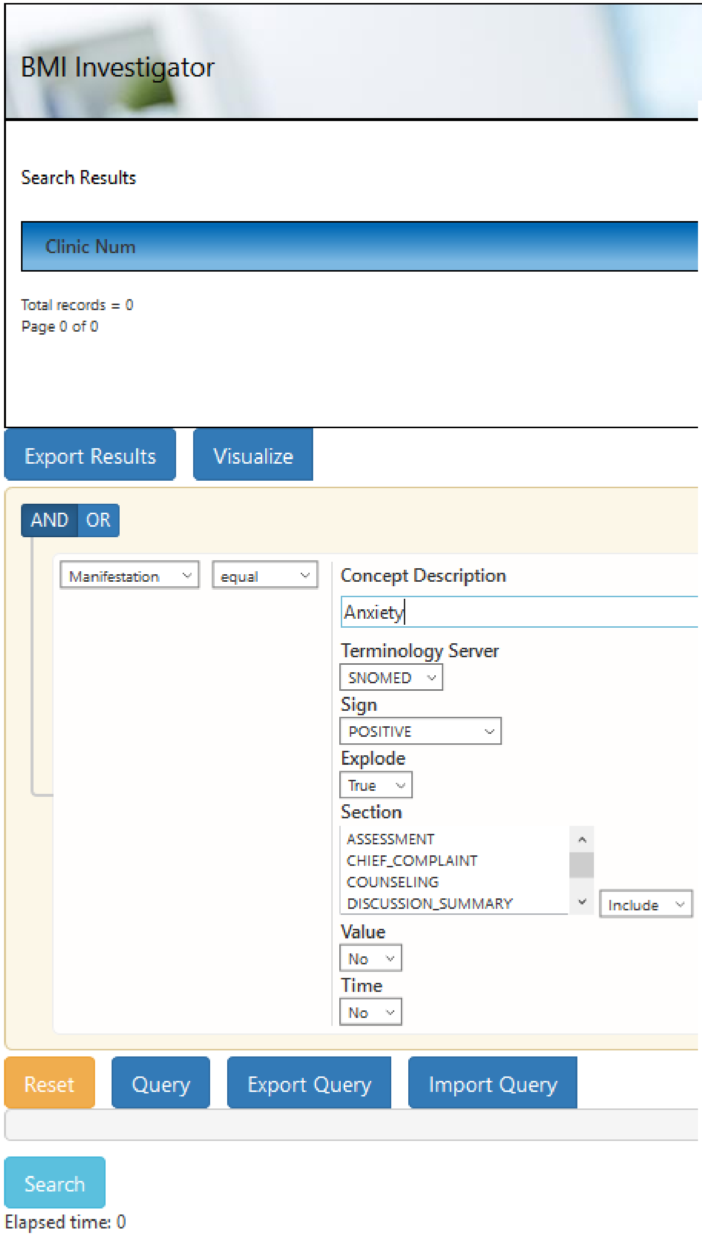
People

Process

Technology

University at Buffalo
Clinical and Translational Science Institute

University at Buffalo
The State University of New York



BMI Investigator

Search Results

Clinic Num

Total records = 0
Page 0 of 0

Manifestation: equal

Concept Description: Anxiety

Terminology Server: SNOMED

Sign: POSITIVE

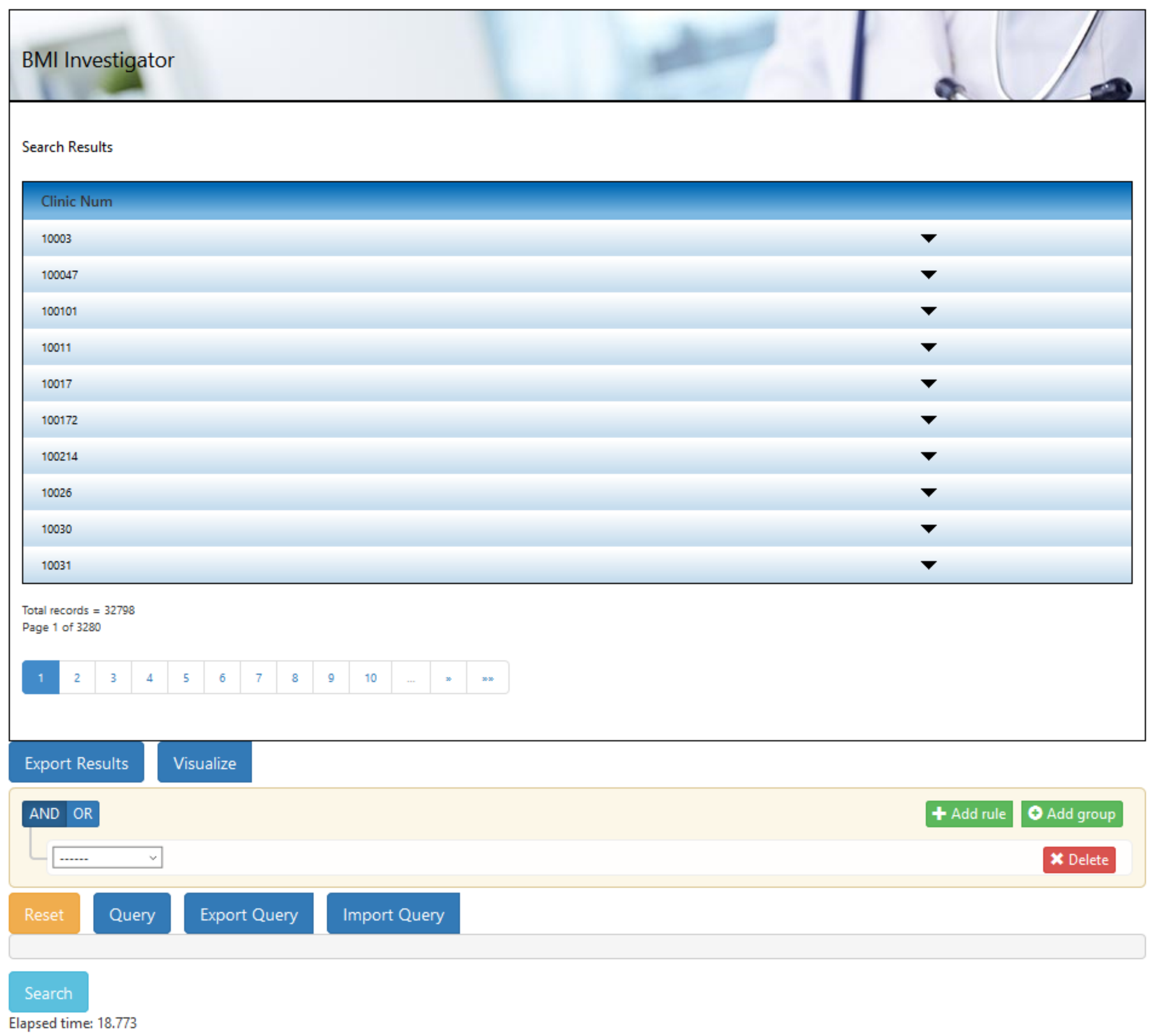
Explode: True

Section: ASSESSMENT, CHIEF_COMPLAINT, COUNSELING, DISCUSSION_SUMMARY — Include

Value: No

Time: No

Elapsed time: 0



BMI Investigator

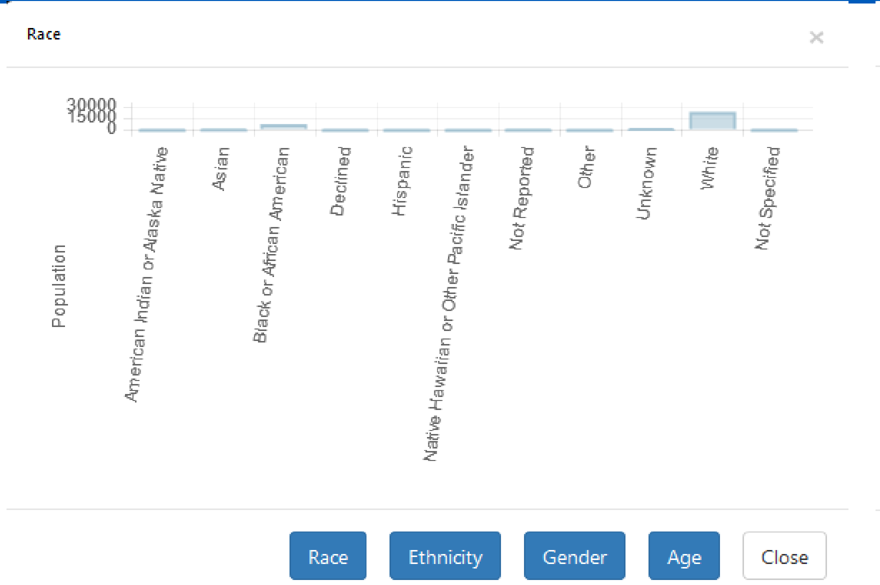
Search Results

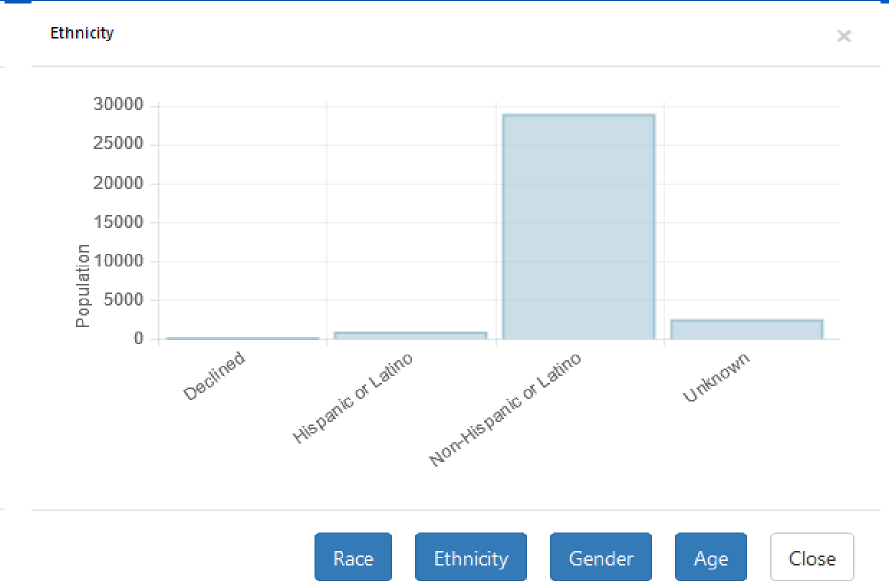
Clinic Num
10003
100047
100101
10011
10017
100172
100214
10026
10030
10031

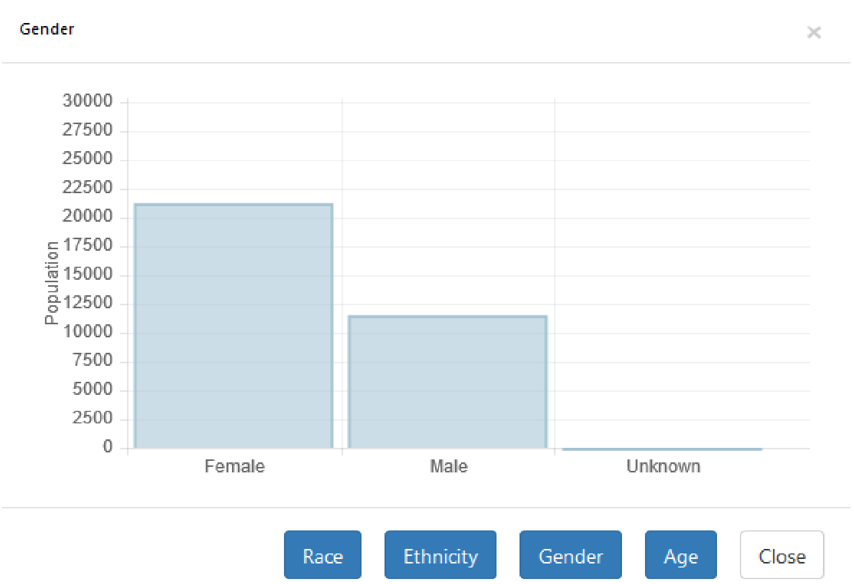
Total records = 32798
Page 1 of 3280

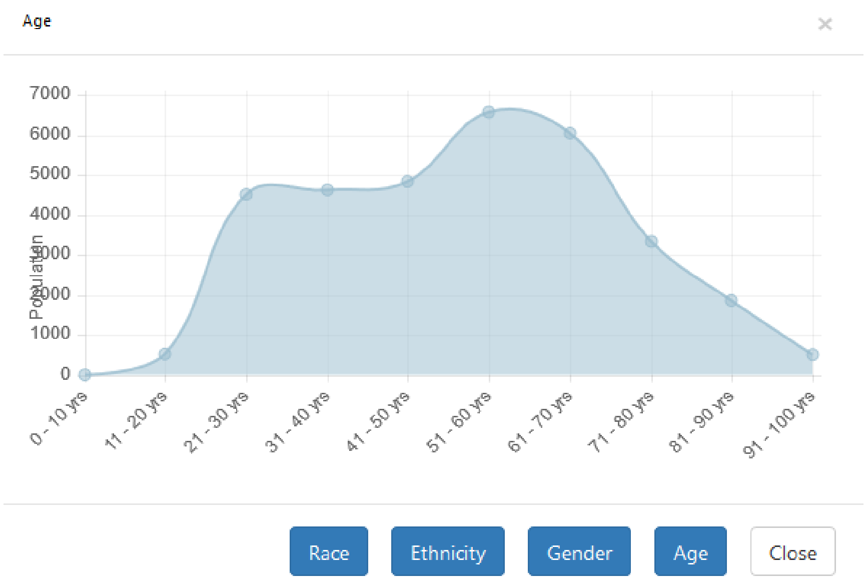
Elapsed time: 18.773

ANXIETY



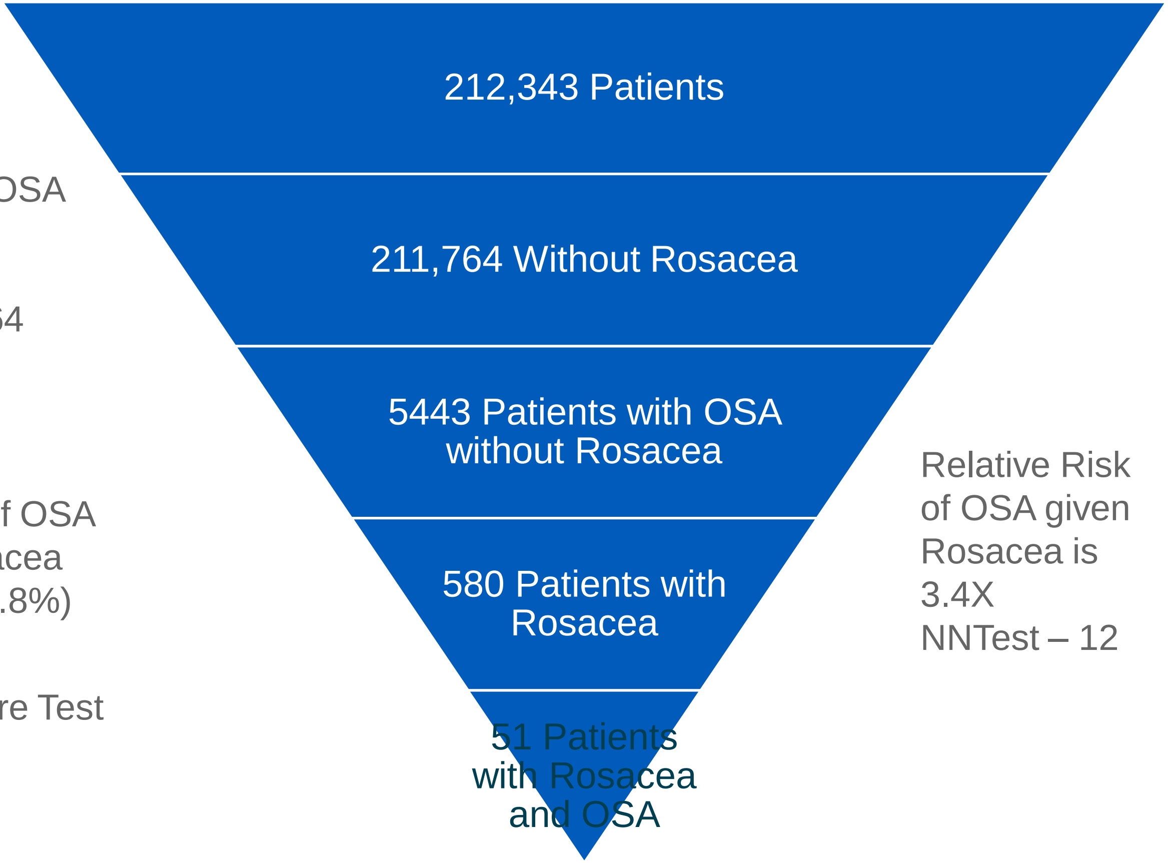






Social Determinants of Health

Study: Are patients with Rosacea at increased risk of having Obstructive Sleep Apnea?



212,343 Patients

211,764 Without Rosacea

5443 Patients with OSA without Rosacea

580 Patients with Rosacea

51 Patients with Rosacea and OSA

Chance of OSA without Rosacea 5443/211764 (2.6%)

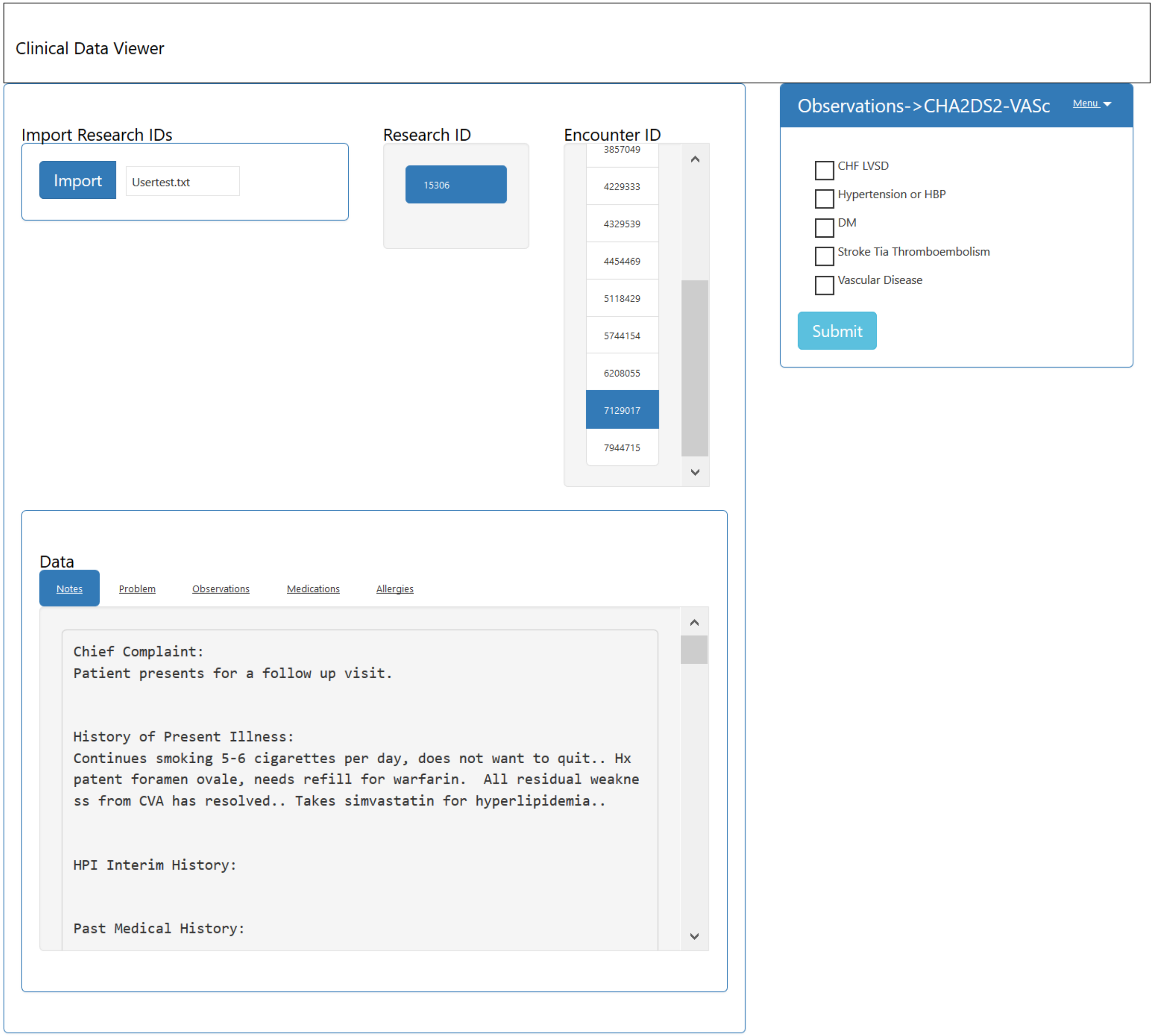
Chance of OSA with Rosacea 51/580 (8.8%)

Chi-Square Test P<0.0001

Relative Risk of OSA given Rosacea is 3.4X
NNTest – 12

Clinical Predication Rule Validation Engine

Electronic Health Record across all EHRs by using a common observational model (OMOP / OHDSI)



Clinical Data Viewer

Import Research IDs

Usertest.txt

Research ID

15306

Encounter ID

3857049
4229333
4329539
4454469
5118429
5744154
6208055
7129017
7944715

Observations->CHA2DS2-VASc

□ CHF LVSD
□ Hypertension or HBP
□ DM
□ Stroke Tia Thromboembolism
□ Vascular Disease

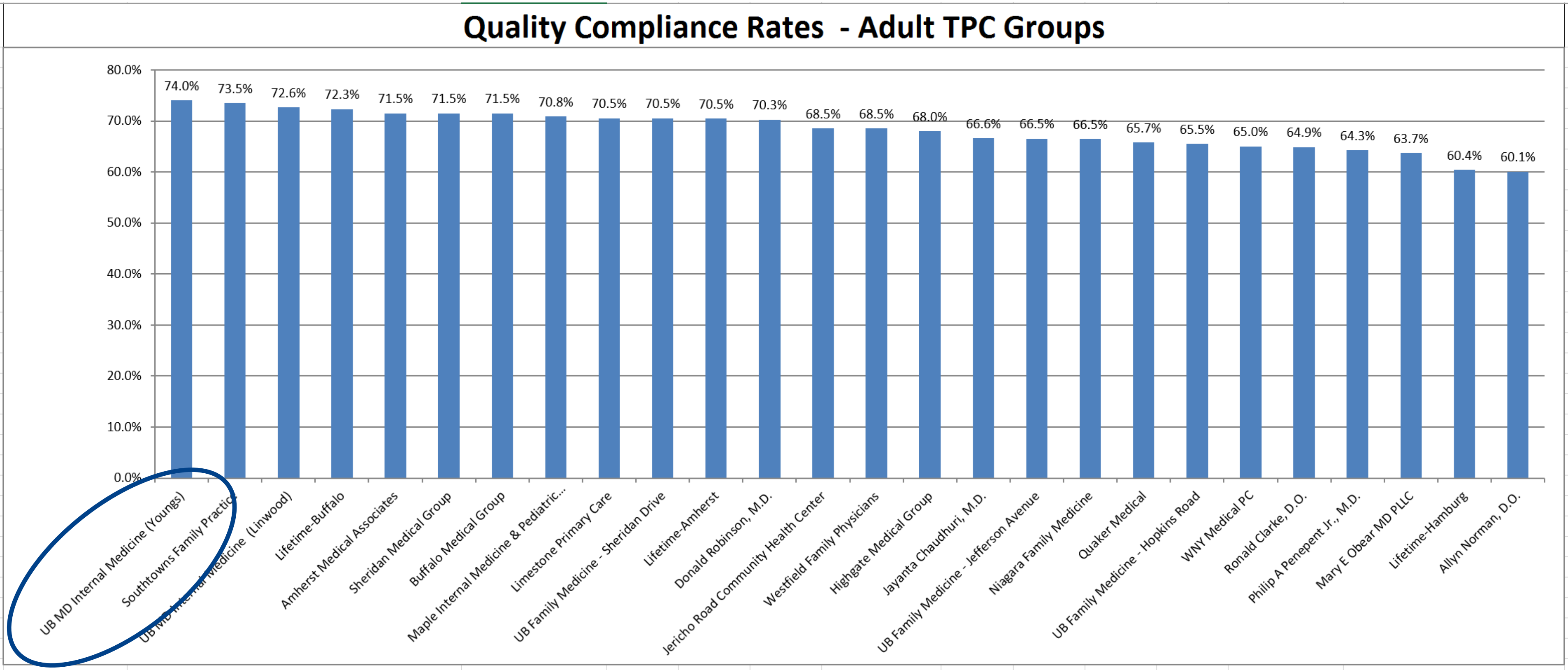
Data

```
Chief Complaint:
Patient presents for a follow up visit.


History of Present Illness:
Continues smoking 5-6 cigarettes per day, does not want to quit.. Hx
patent foramen ovale, needs refill for warfarin.  All residual weakne
ss from CVA has resolved.. Takes simvastatin for hyperlipidemia..


HPI Interim History:


Past Medical History:
```

Quality Accomplishments

- Improved Quality of Care
  - Metrics and Measurement of Practice Outcomes
  - Patient Centered Medical Home
  - Quality Improvement Project Registry
  - Improved outcomes in Payer Measures
- Improvement in Internal Referrals
  - Went from 54% to 82% Internal Referrals
- DOM Strategic Plan Implementation
  - Quality Tools
  - Quality Structures
  - Support of New Multispecialty Clinical and Research Centers

From third to the last to the best in IHA Quality metrics



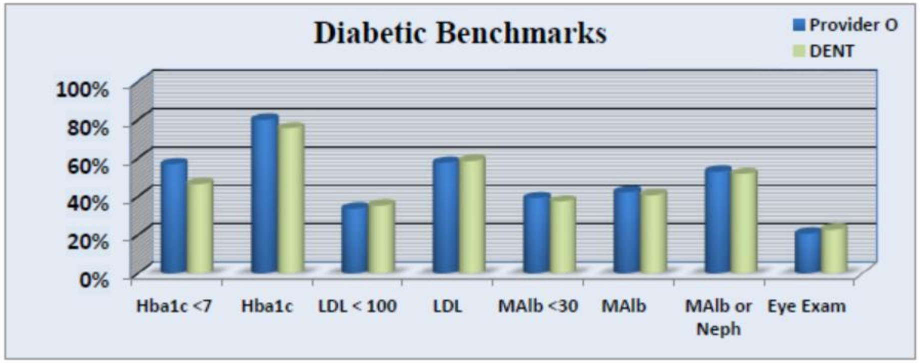
Period	Q2 2016			
METRIC	QUALITY	*Confidential*	**DATA SOURCE**	Quality Dashboard, VIPS
			TIME FRAME	July 1 2015 - June 30 2016
			GROUP COUNT	26

Internal Medicine Provider Report Cards for Target Patient Populations

Provider O

UB|MD INTERNAL MEDICINE

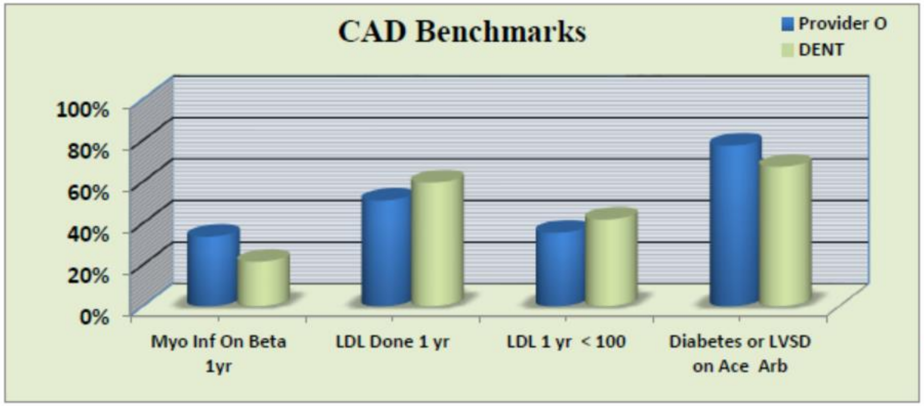
Patients with Diabetes	130



All benchmarks are within a one year period. Patient counts are on a provider level, unless otherwise noted.

	Provider	Practice
Hba1c <7	56.9%	46.6%
Hba1c 1 yr	80.0%	75.6%
LDL <100	33.8%	35.4%
LDL 1 yr	57.7%	58.7%
MAlb <30	39.2%	37.5%
MAlb 1yr	42.3%	40.9%
MAlb or Neph	53.1%	51.8%
Eye Exam	20.8%	22.8%

Patients with Coronary Artery Disease	65



Goals for benchmarks are 85% or higher for labs, vaccinations and exams. An 8% improvement from year to year is also considered meeting goals.

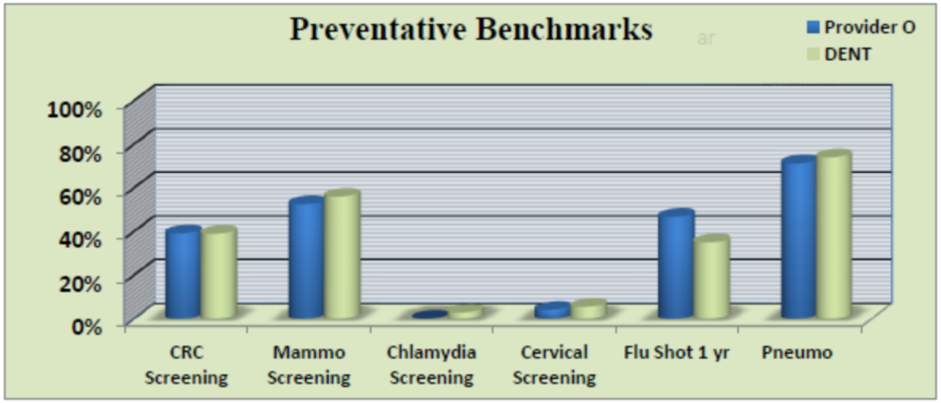
	Provider	Practice
Myocardial Infarction	3 Pts	47 Pts
Myo Inf on Beta	33.3%	21.3%
LDL Done 1yr	50.8%	59.5%
LDL <100	35.4%	41.8%
Diabetes or LVSD	22 Pts	219 Pts
Diab/LVSD on Ace/Arb	77.3%	67.1%

Internal Medicine Provider Report Cards for Target Patient Populations

Provider O

UB|MD INTERNAL MEDICINE

Patients Eligible for CRC Sreening	496	Patients Eligible for Mammo Sreening	423
Patients Eligible for Cervical Screening	584	Patients Eligible for Chlamydia Sreening	54
Patients Eligible for Flu Shot	957	Patients Eligible for Pneumo Shot	264

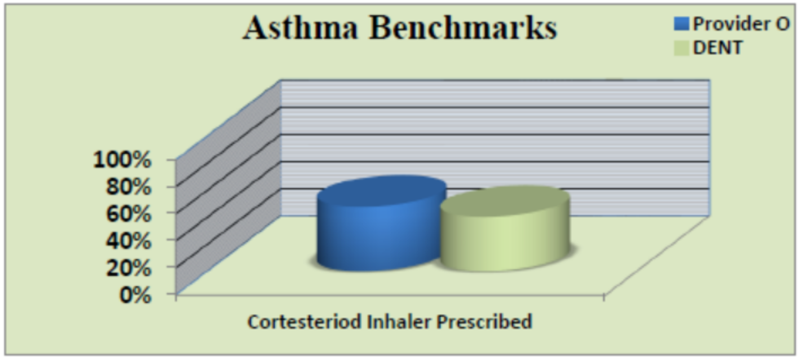


Colorectal Screening is colonoscopy in the last 10 yrs or FOBT in the last 2 yrs for patients between 50 and 80. Mammogran Screening is reporting on women ages 42 to 69. Chlamydia Screening is reporting on patients between 18 and 24. Cervical Screening is Pap Smear in the last 3 yrs. Flu shot is done with in the last yr yrs. Flu shot is done with in the last yr and Pneumo is a Pneumococcal vaccination lifetime

	Provider	Practice
CRC Screening	39.1%	39.0%
Mammo Screening	52.5%	56.0%
Chlamydia Screening	0.0%	2.9%
Cervical Cancer Screening	4.1%	5.4%
Flu shot 1yr	46.7%	35.0%
Pneumo	71.2%	74.0%

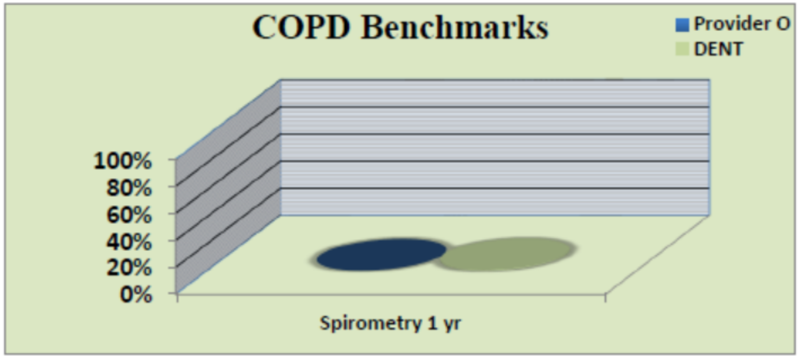
Internal Medicine Provider Report Cards for Target Patient Populations

Patients with Persistent Asthma	115



	Provider	Practice
Cortesteroid Inhaler Prescribed	47.8%	40.3%

Patients with COPD	38



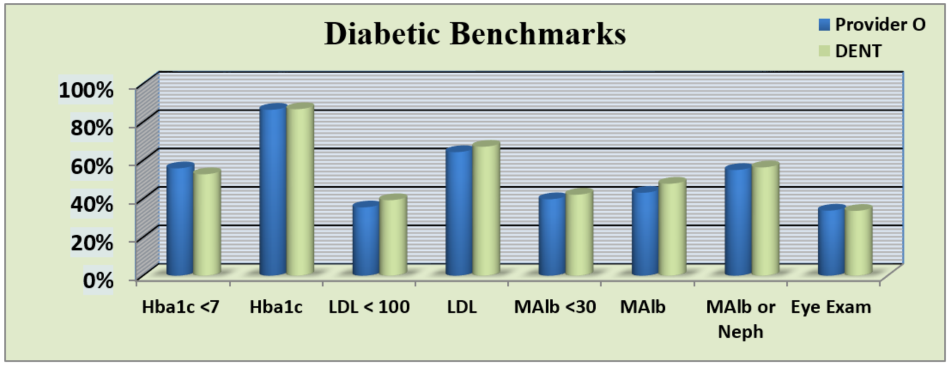
	Provider	Practice
Spirometry Test Done 1 yr	0.0%	0.4%

Internal Medicine Provider Report Cards for Target Patient Populations

Provider O

UB|MD INTERNAL MEDICINE

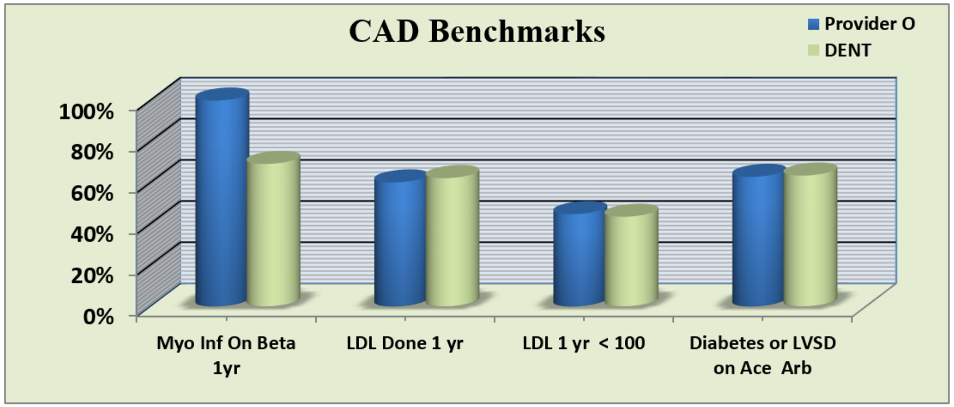
Patients with Diabetes	118



All benchmarks are within a one year period. Patient counts are on a provider level, unless otherwise noted.

	Provider	Practice
Hba1c <7	55.9%	52.8%
Hba1c 1 yr	86.4%	86.8%
LDL <100	35.6%	39.4%
LDL 1 yr	64.4%	67.2%
MAlb <30	39.8%	42.2%
MAlb 1yr	43.2%	47.9%
MAlb or Neph	55.1%	56.4%
Eye Exam	33.9%	33.8%

Patients with Coronary Artery Disease	78



Goals for benchmarks are 85% or higher for labs, vaccinations and exams. An 8% improvement from year to year is also considered

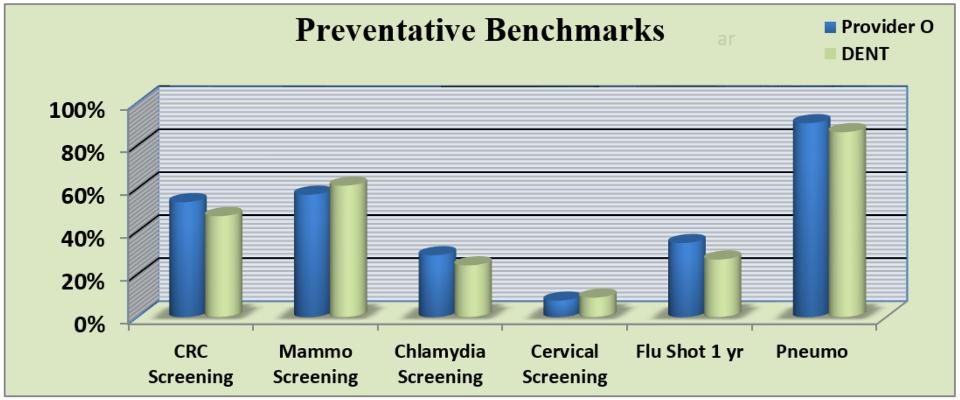
	Provider	Practice
Myocardial Infarction	1 Pts	13 Pts
Myo Inf on Beta	100.0%	69.2%
LDL Done 1yr	60.3%	62.3%
LDL <100	44.9%	43.6%
Diabetes or LVSD	27 Pts	206 Pts
Diab/LVSD on Ace/Arb	63.0%	63.6%

Internal Medicine Provider Report Cards for Target Patient Populations

Provider O

UB|MD INTERNAL MEDICINE

Patients Eligible for CRC Sreening	637	Patients Eligible for Mammo Sreening	486
Patients Eligible for Cervical Screening	699	Patients Eligible for Chlamydia Sreening	59
Patients Eligible for Flu Shot	1186	Patients Eligible for Pneumo Shot	361

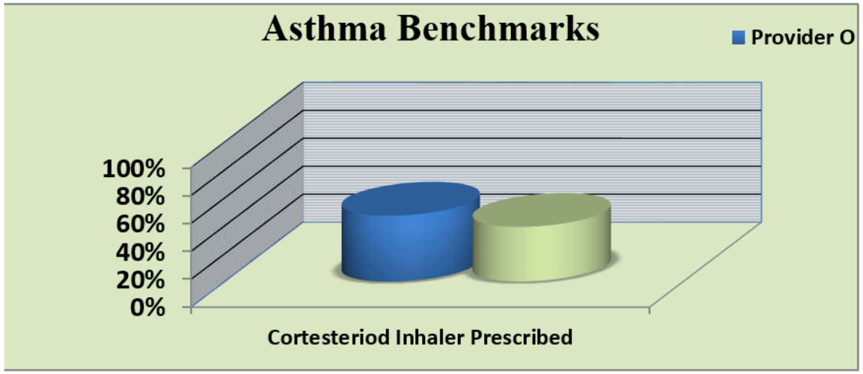


Colorectal Screening is colonoscopy in the last 10 yrs or FOBT in the last 2 yrs for patients between 50 and 80. Mammogran Screening is reporting on women ages 42 to 69. Chlamydia Screening is reporting on patients between 18 and 24. Cervical Screening is Pap Smear in the last 3 yrs. Flu shot is done with in the last yr and Pneumo is a Pneumococcal vaccination lifetime

	Provider	Practice
CRC Screening	53.5%	47.0%
Mammo Screening	57.0%	61.2%
Chlamydia Screening	28.8%	24.1%
Cervical Cancer Screening	7.7%	9.1%
Flu shot 1yr	34.5%	26.8%
Pneumo	90.3%	86.1%

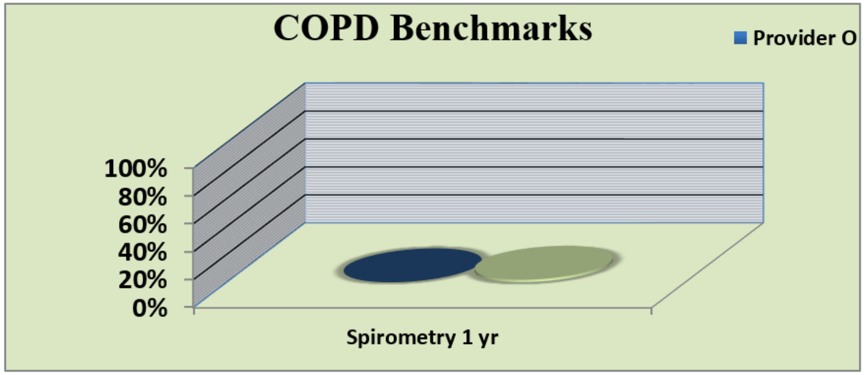
Internal Medicine Provider Report Cards for Target Patient Populations

Patients with Persistent Asthma	159

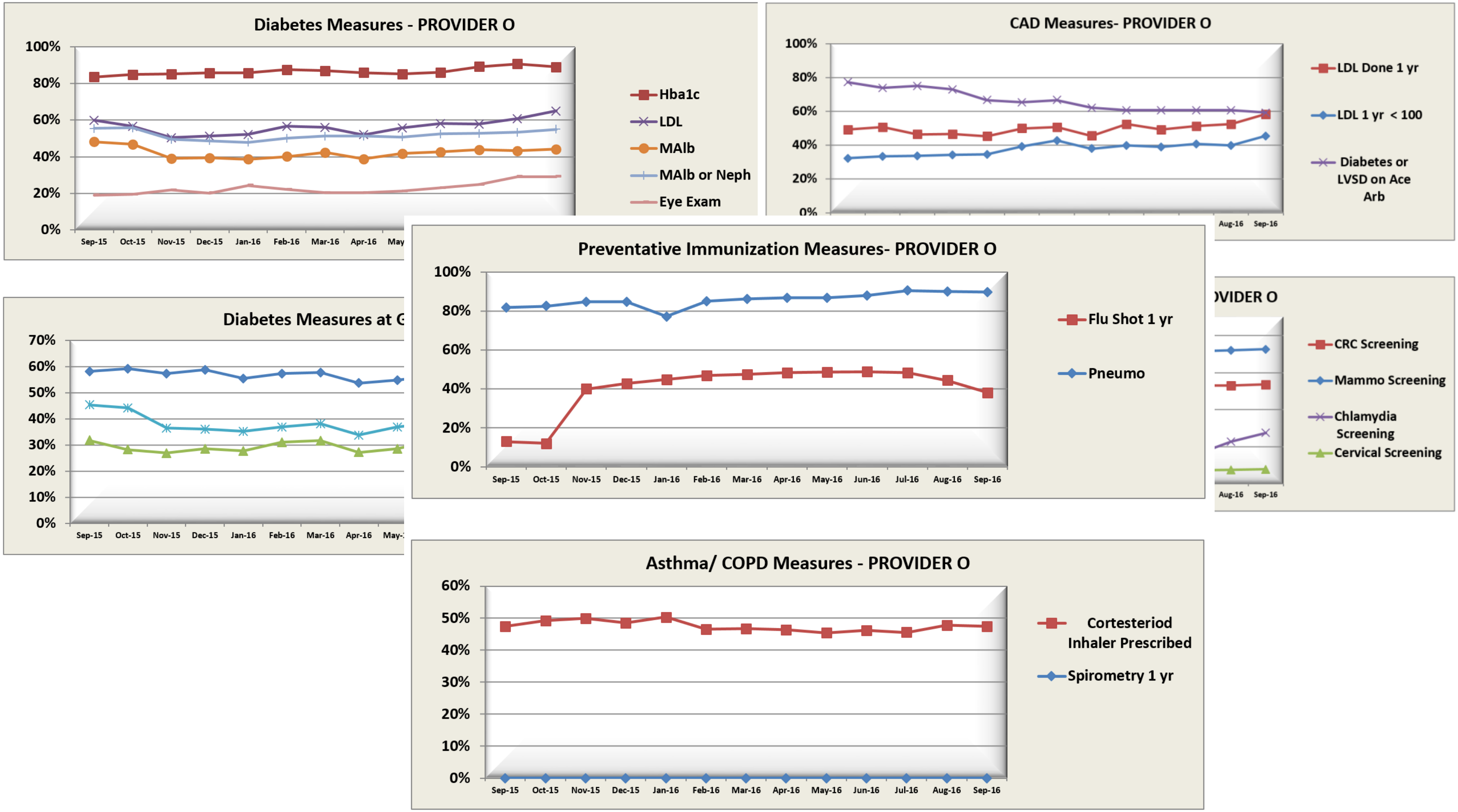


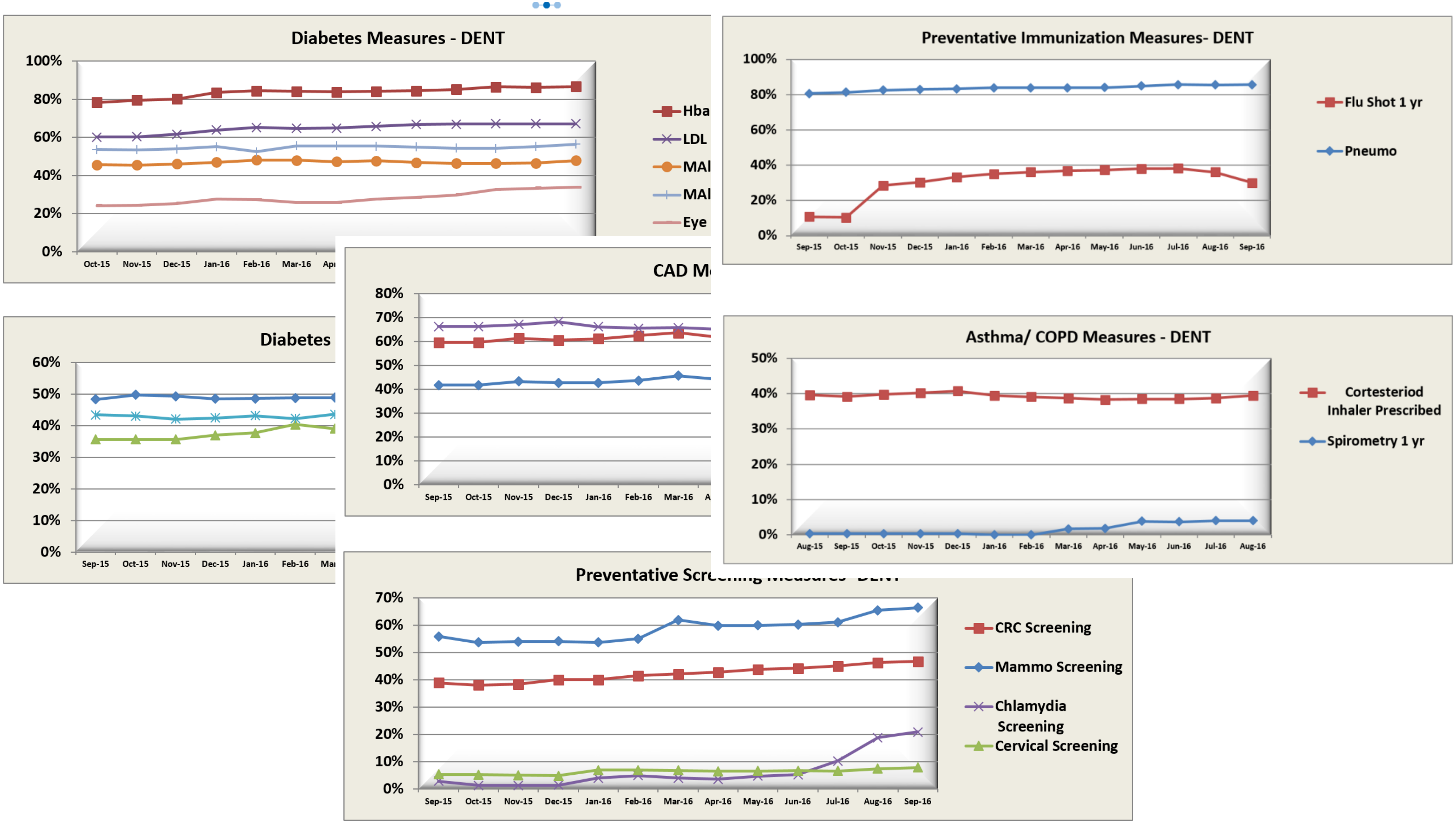
	Provider	Practice
Cortesteroid Inhaler Prescribed	47.2%	39.2%

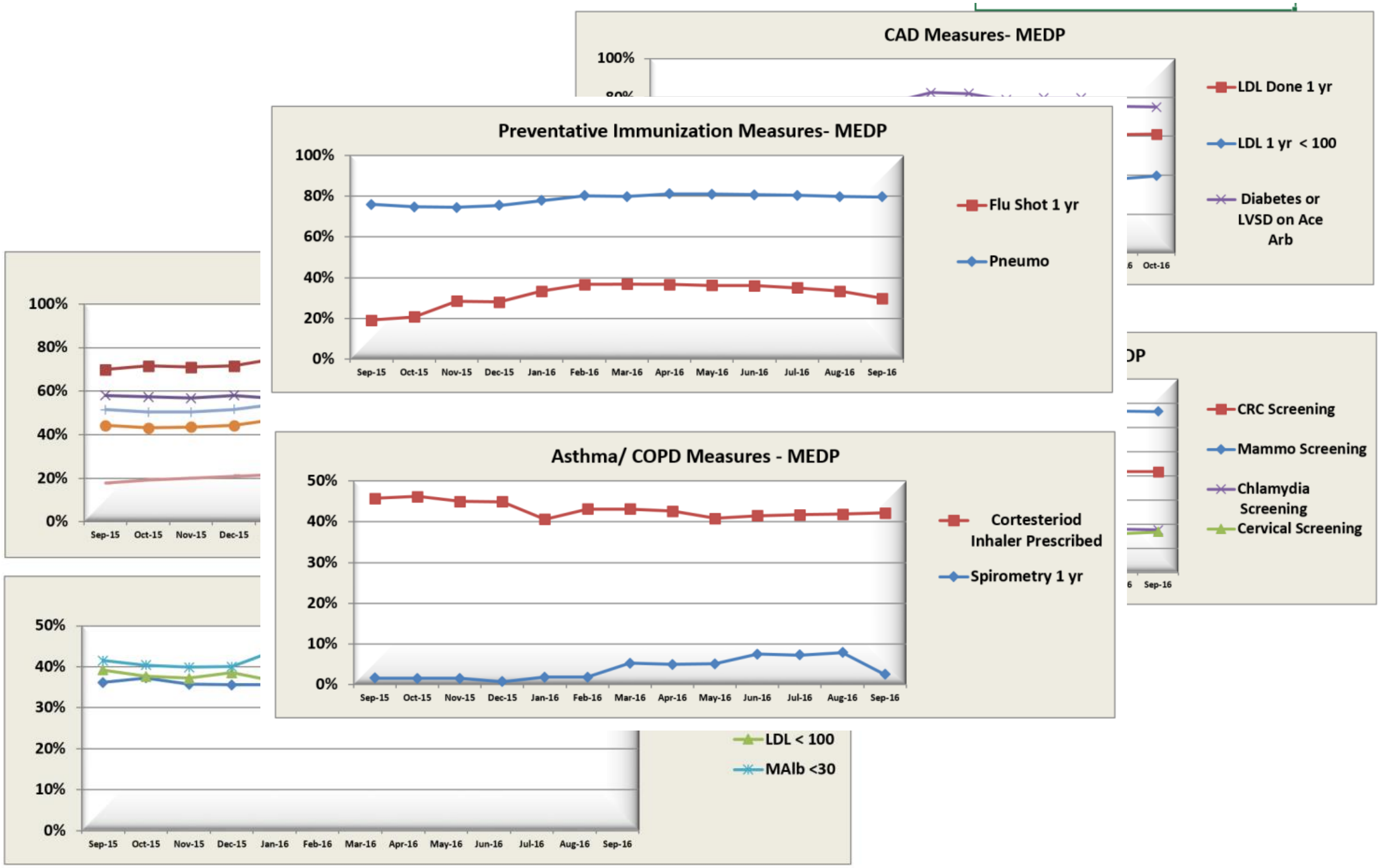
Patients with COPD	13



	Provider	Practice
Spirometry Test Done 1 yr	0.0%	1.8%


Diabetes Measures - PROVIDER O
100%
80%
60%
40%
20%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May
Hba1c
LDL
MAlb
MAlb or Neph
Eye Exam
CAD Measures- PROVIDER O
100%
80%
60%
40%
20%
0%
Aug-16
Sep-16
LDL Done 1 yr
LDL 1 yr < 100
Diabetes or LVSD on Ace Arb
Diabetes Measures at G
70%
60%
50%
40%
30%
20%
10%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-
Preventative Immunization Measures- PROVIDER O
100%
80%
60%
40%
20%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Flu Shot 1 yr
Pneumo
)VIDER O
Aug-16
Sep-16
CRC Screening
Mammo Screening
Chlamydia Screening
Cervical Screening
Asthma/ COPD Measures - PROVIDER O
60%
50%
40%
30%
20%
10%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Cortesteriod Inhaler Prescribed
Spirometry 1 yr

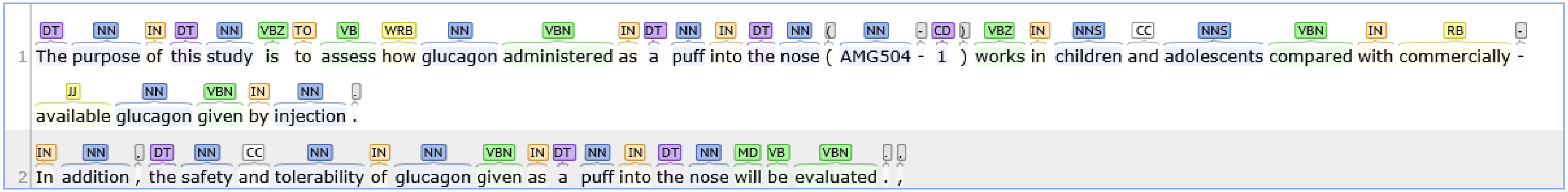

Diabetes Measures - DENT
100%
80%
60%
40%
20%
0%
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr
Hba
LDL
MAI
MAI
Eye
Preventative Immunization Measures- DENT
100%
80%
60%
40%
20%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Flu Shot 1 yr
Pneumo
CAD M
80%
70%
60%
50%
40%
30%
20%
10%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
A
Diabetes
60%
50%
40%
30%
20%
10%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar
Asthma/ COPD Measures - DENT
50%
40%
30%
20%
10%
0%
Aug-15
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Cortesteriod Inhaler Prescribed
Spirometry 1 yr
Preventative Scre
70%
60%
50%
40%
30%
20%
10%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
CRC Screening
Mammo Screening
Chlamydia Screening
Cervical Screening


CAD Measures- MEDP
100%
LDL Done 1 yr
LDL 1 yr < 100
Diabetes or LVSD on Ace Arb
Oct-16
Preventative Immunization Measures- MEDP
100%
80%
60%
40%
20%
0%
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Flu Shot 1 yr
Pneumo
CRC Screening
Mammo Screening
Chlamydia Screening
Cervical Screening
Sep-16
Asthma/ COPD Measures - MEDP
50%
40%
30%
20%
10%
0%
Cortesteriod Inhaler Prescribed
Spirometry 1 yr
LDL < 100
MAlb <30

Assessment of Intranasal Glucagon in Children and Adolescents With Type 1 Diabetes

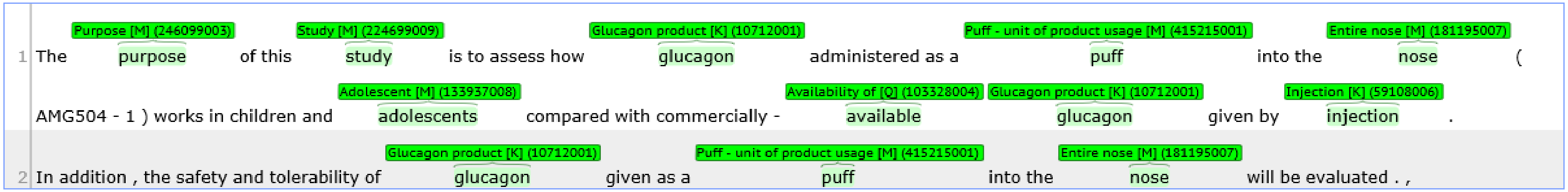
The purpose of this study is to assess how glucagon administered as a puff into the nose (AMG504-1) works in children and adolescents compared with commercially-available glucagon given by injection. In addition, the safety and tolerability of glucagon given as a puff into the nose will be evaluated.

Part-of-Speech:



```
1  DT  NN      IN DT   NN    VBZ TO VB     WRB NN       VBN          IN DT NN   IN   DT  NN   (  NN     - CD )  VBZ   IN NNS      CC  NNS         VBN      IN   RB          -
   The purpose of this study is  to assess how glucagon administered as a  puff into the nose ( AMG504 - 1  )  works in children and adolescents compared with commercially -

   JJ        NN       VBN   IN NN        .
   available glucagon given by injection .

2  IN NN       , DT  NN     CC  NN           IN NN       VBN   IN DT NN   IN   DT  NN   MD   VB VBN       . ,
   In addition , the safety and tolerability of glucagon given as a  puff into the nose will be evaluated . ,
```

SNOMED Codes:



```
1       Purpose [M] (246099003)         Study [M] (224699009)                    Glucagon product [K] (10712001)                       Puff - unit of product usage [M] (415215001)               Entire nose [M] (181195007)
   The  purpose                of this  study                 is to assess how   glucagon                        administered as a     puff                                         into the      nose                         (

                                                 Adolescent [M] (133937008)                                 Availability of [Q] (103328004)  Glucagon product [K] (10712001)              Injection [K] (59108006)
   AMG504 - 1 ) works in children and            adolescents                compared with commercially -    available                        glucagon                        given by     injection                 .

                                                 Glucagon product [K] (10712001)             Puff - unit of product usage [M] (415215001)            Entire nose [M] (181195007)
2  In addition , the safety and tolerability of  glucagon                        given as a  puff                                         into the       nose                        will be evaluated . ,
```

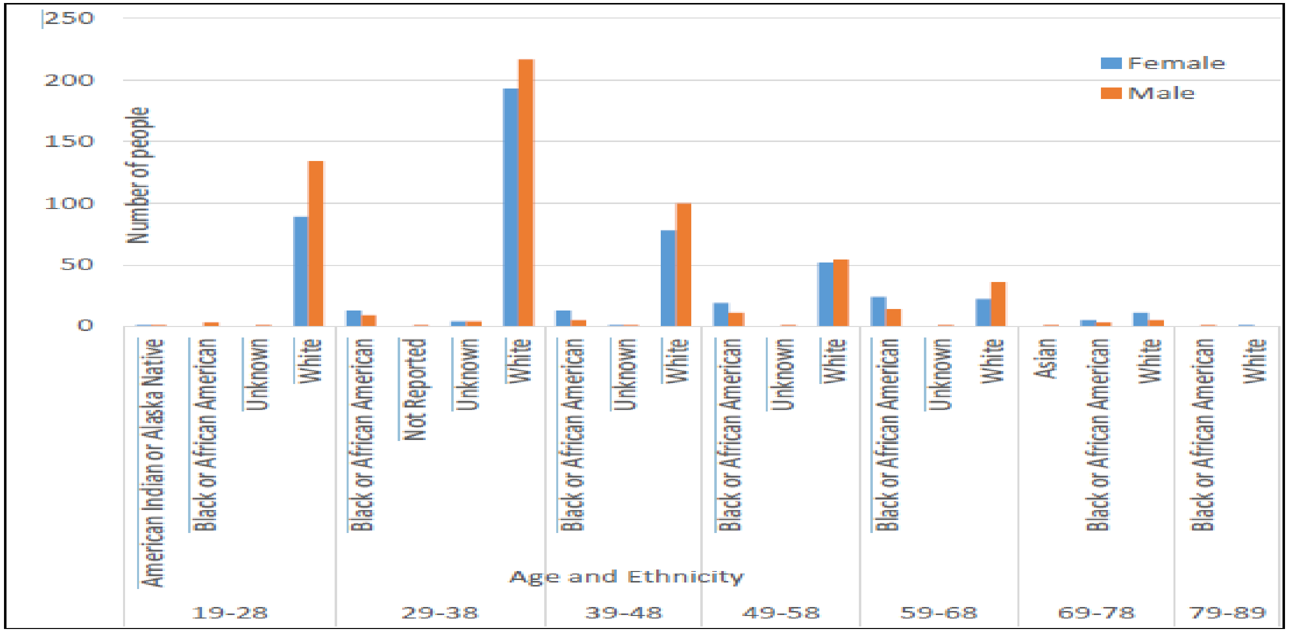
Prescription Opioid Dependence in Western New York: Using Data Analytics to find an answer to the Opioid Epidemic

Shyamashree Sinha, Gale R Burstein, Kenneth E Leonard, Timothy F Murphy, Peter L Elkin

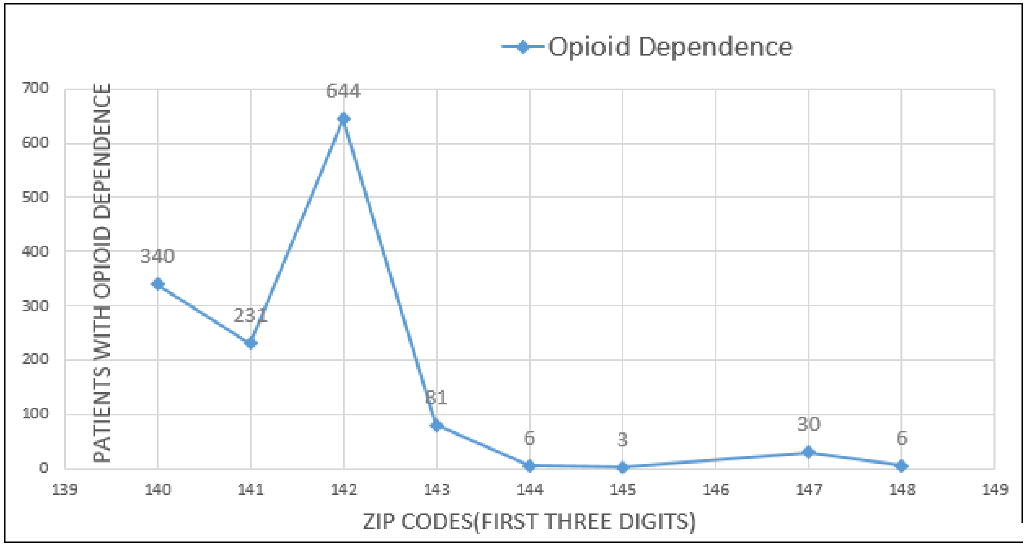
Department of Biomedical Informatics/ Department of Anesthesiology

Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, The State University of New York, Buffalo, New York

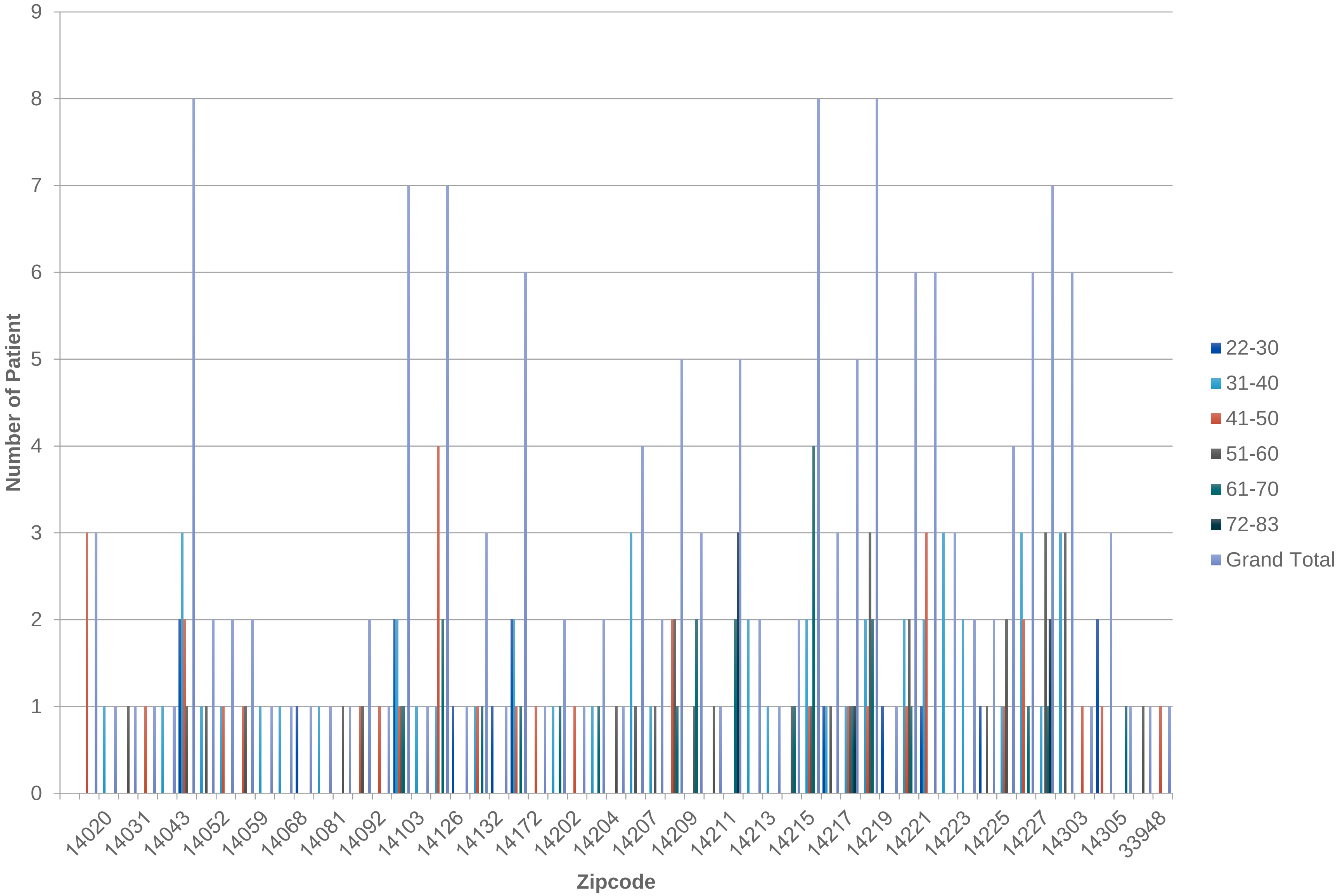
Distribution of Opioid Dependence among the Non-Hispanic community in the clinic population of Western New York

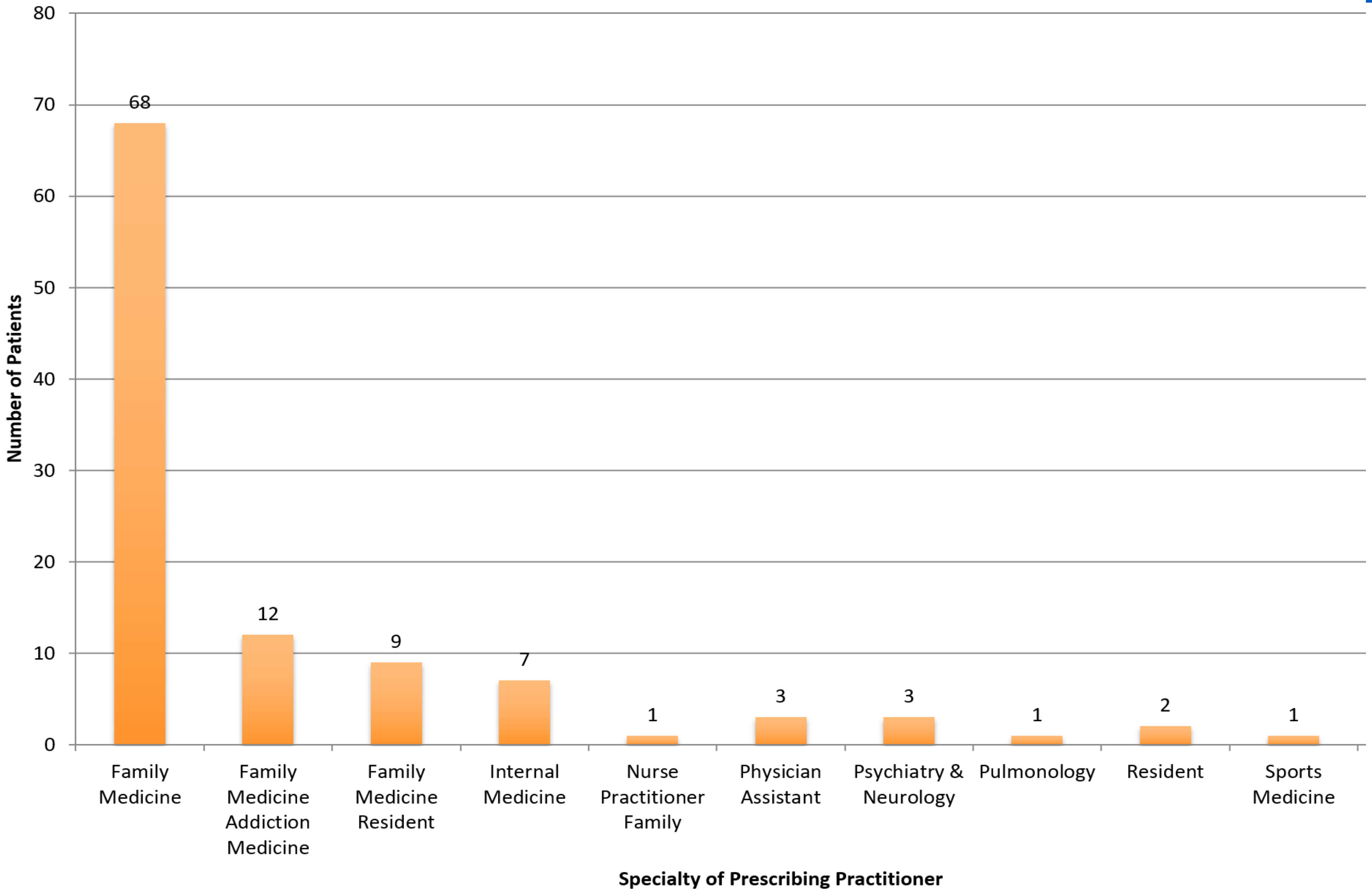


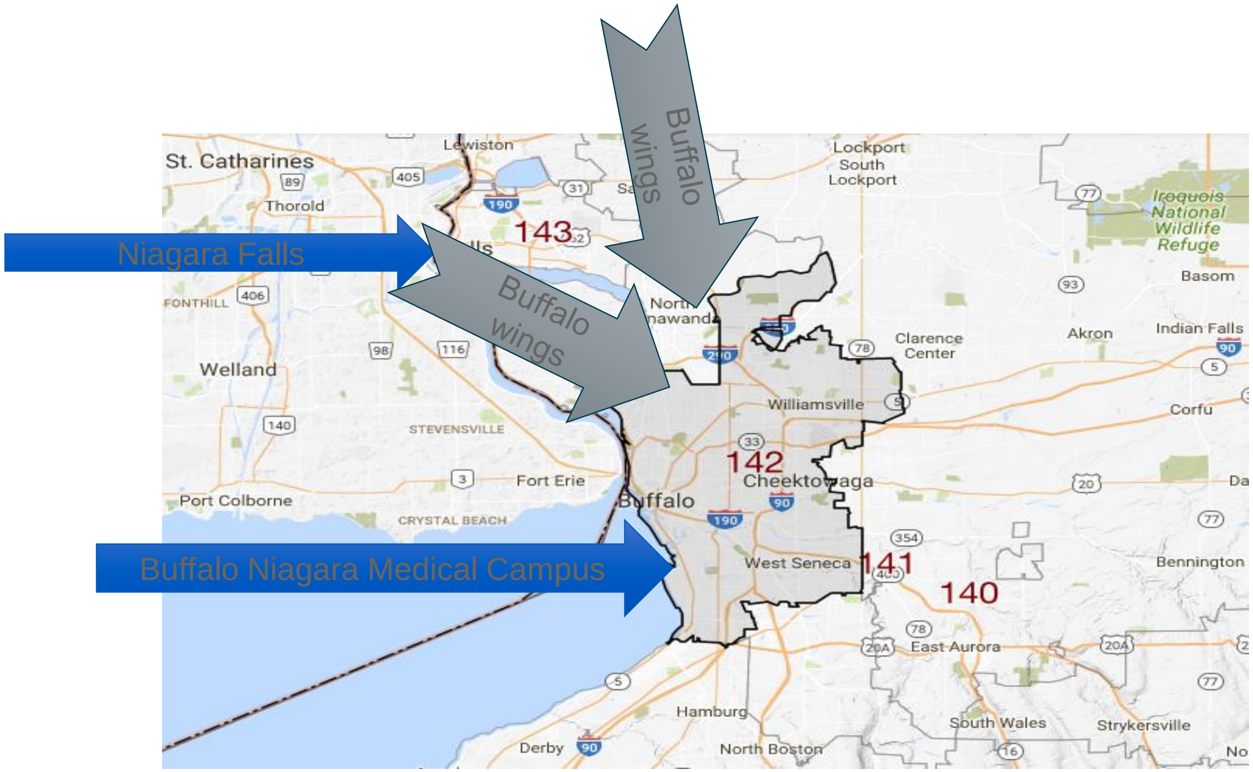
Distribution of Opioid Dependence based on geographical location



The distribution of the patients based on the first three numbers of the zip code showed area 142 had the highest number of opioid dependent population


9
8
7
6
5
4
3
2
1
0
Number of Patient
14020
14031
14043
14052
14059
14068
14081
14092
14103
14126
14132
14172
14202
14204
14207
14209
14211
14213
14215
14217
14219
14221
14223
14225
14227
14303
14305
33948
Zipcode
22-30
31-40
41-50
51-60
61-70
72-83
Grand Total


80
70
60
50
40
30
20
10
0
Number of Patients
68
12
9
7
1
3
3
1
2
1
Family Medicine
Family Medicine Addiction Medicine
Family Medicine Resident
Internal Medicine
Nurse Practitioner Family
Physician Assistant
Psychiatry & Neurology
Pulmonology
Resident
Sports Medicine
Specialty of Prescribing Practitioner



Map showing boundaries of area with zip code 142:
https://www.maptechnica.com/zip3-prefix-map/142

University at Buffalo
Clinical and Translational Science Institute

AI AND NATURAL LANGUAGE PROCESSING (NLP) TO ENHANCE STRUCTURED DATA'S ABILITY TO IDENTIFY NONVALVULAR ATRIAL FIBRILLATION PATIENTS AND THEIR STROKE AND BLEEDING RISK

Peter L. Elkin, MD, MACP, FACMI, FNYAM

For the NVAF Surveillance Study team

Advancing research discoveries to improve health for all

Goal of the study

- The goal of this study is to compare clinician-rated stroke and bleed risk assessments in Nonvalvular Atrial Fibrillation (NVAF) patients with assessments utilizing NLP derived codified EHR data for CHA_2DS_2-VASc and HAS-BLED scores.

Research Questions

- ***Research Question: 1***
- What is the accuracy of using structured data (ICD and CPT and Medication codes) alone vs. unstructured (ie, Clinical notes and reports, labs and Medications) plus structured data to identify patients who have Atrial Fibrillation?
- ***Objectives:***
- Compare structured data to structured and unstructured data using NLP to identify NVAF Patients - validated by clinician assessment

Research Question 4

Does the method (using structured data only vs. structured plus unstructured data) of determining risk scores affect the treatment of NVAF patients for stroke prevention with OAC?

Objectives:

1. Using structured and unstructured data assessments of CHA_2DS_2-VASc, HAS-BLED scores and contraindications for OAC, classify the patient cohorts as follows and compare the treatment rates with OAC.
    1. Would benefit and are on OAC;
    2. Would benefit but are not on OAC;
    3. Would not benefit and are on OAC;
    4. Would not benefit and are not on OAC

Semi-Supervised Machine Learning

- Small Amount of Labeled Data and Large Amounts of Unlabeled Data
- Cheaper and Faster than a Fully Supervised Approach
- More accurate than an unsupervised approach
- Can be used to create models from a mixed dataset. These models can be used for Biosurveillance.
- Example:
  - Intuitively, we can think of the learning problem as an exam and labeled data as the few example problems that the teacher solved in class. The educator also provides a set of unsolved problems. In transductive reasoning, these unsolved problems are a take-home exam questions and you want to do well on them in particular. In inductive reasoning, these are practice problems of the sort you will encounter on the in-class exam.
- NSQIP - Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011 Aug 24;306(8):848-55.
- NVAF Study – in press, Circulation, 2017.

Result

Table 1. Comparison of outcomes for Structured and Structured plus Unstructured data against the gold standard.

Outcome	Structured	Structured+NLP	P
Sensitivity	.773 (.68, .79)	1 (.979,1)	<0.001
Specificity	.47 (.258, .65)	.444 (.279, .619)	0.317
PPV	.91 (.87, .95)	.93 (.893, .956)	0.007
NPV	.215(.131, .322)	1 (.713, 1)	<0.001
kappa	.156 (.041, .271)	.585 (.414, .733)	<0.001

- Out of the 96,681 patients identified in the AllScripts EHR database, 2.8% (2722 cases) were identified with NVAF by the Structured+NLP method as opposed to 1.9% for Structured alone (1849 cases) with a difference of 873 cases

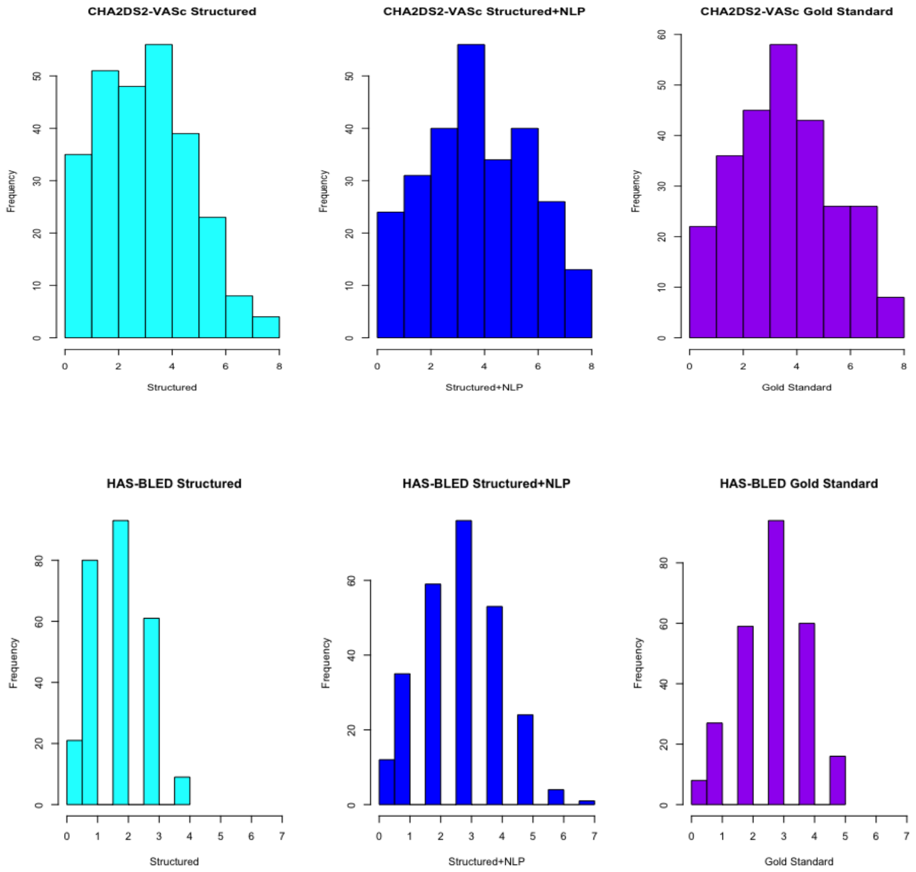
- Out of the 96,681 patients identified in the AllScripts EHR database, 2.8% (2722 cases) were identified with NVAF by the Structured+NLP method as opposed to 1.9% for Structured alone (1849 cases) with a difference of 873 cases

- Based on the PPV adjusting the true positive rates for both ICD9 and NLP alone this converts to a 36.3 % improvement identification of true cases in this NVAF cohort.

Histograms of CHA_2DS_2-VASC Scores and HAS-BLED scores

Results:

Table 2.1. Pearson Product Moment				
	Structured		Structured+NLP	
	estimate (95% CI)	p-value	estimate (95% CI)	p-value
CHA_2DS_2-VASC Score	0.819 (0.775,0.855)	<.001	0.898 (0.872,0.92)	<.001
HAS-BLED Score	0.688 (0.619,0.747)	<.001	0.717 (0.652,0.771)	<.001



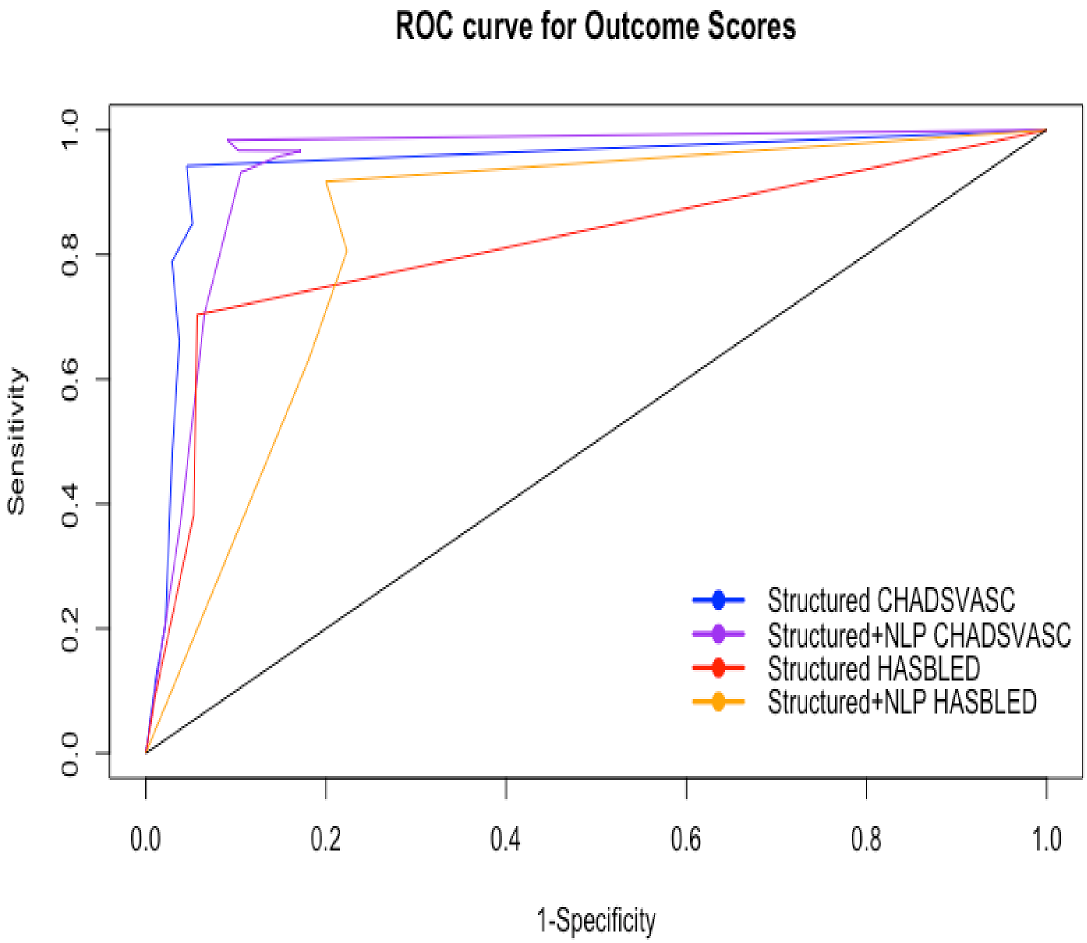
Sensitivity and Specificity of Outcomes Compared to Gold Standard			
HAS-BLED		CHA_2DS_2-VASC	
Method: McNemar		Method: Exact Binomial	
Sensitivity		Sensitivity	
Structured	0.382	Structured	**0.942**
Structured+NLP	0.806	Structured+NLP	**0.983**
Difference	0.424	Difference	**0.0413**
Test Statistic	72	Test Statistic	**-**
p-value	<.0001	p-value	**0.00195**
Method: McNemar		Method: Exact Binomial	
Specificity		Specificity	
Structured	0.947	Structured	**0.955**
Structured+NLP	0.777	Structured+NLP	**0.909**
Difference	-0.17	Difference	**-0.0455**
Test Statistic	16	Test Statistic	**-**
p-value	<.0001	p-value	**1**
Method: Generalized Score		Method: Generalized Score	
Positive Predictive Value		Positive Predictive Value	
Structured	0.929	Structured	**0.996**
Structured+NLP	0.867	Structured+NLP	**0.992**
Difference	.061	Difference	**0.004**
Test Statistic	4.487	Test Statistic	**0.915**
p-value	0.034	p-value	**0.339**
Negative Predictive Value		Negative Predictive Value	
Structured	0.459	Structured	**0.6**
Structured+NLP	0.689	Structured+NLP	**0.833**
Difference	0.23	Difference	**0.233**
Test Statistic	47.757	Test Statistic	**11.662**
p-value	<.00001	p-value	**<0.001**

Area under the Curves (AUC)
C-Index and Somer's D using Ordinal Logistic Regression (where probabilities are modelled as P(Y>=k|X))
(R *rms* and *Hmisc* packages)

C-index Structured *CHA_2DS_2-VASC:* 0.863 (CI:0.838, 0.887) (Somer's D (D_{xy}): 0.726, SD=0.025)

C-index Structured+NLP *CHA_2DS_2-VASC:* 0.914 (CI: 0.896, 0.933) (Somer's D (D_{xy}): 0.829, SD=0.0185)
Z=0.625/.0316=19.776

***CHA_2DS_2-VASC**:* Compared to Standard normal distribution*: **2-Sided p-value: <0.001**
1-Sided p-value: <0.001



Predictive Risk Model Generation of Requiring Rx with OAC and not being currently on treatment

		Would Benefit and On OAC	Would Benefit and Not on OAC	Would Not Benefit and Are on OAC	Would Not Benefit and Are Not on OAC
Gold Standard with Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and Contraindication	3	2	0	1
	CHA$_2$DS$_2$-VASc > 2AND HAS-BLED ≥ 3 and Contraindication	6	0	0	1
	CHA$_2$DS$_2$-VASc <2 and Contraindication	0	0	0	1
Gold Standard with No Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and No Contraindication	38	15	0	14
	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED ≥ 3 and No Contraindication	129	16	1	16
	CHA$_2$DS$_2$-VASc <2 and No Contraindication	10	3	0	8
Structured with Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and Contraindication	4	1	0	0
	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED ≥ 3 and Contraindication	3	1	0	0
	CHA$_2$DS$_2$-VASc <2 and Contraindication	0	0	0	0
Structured with No Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and No Contraindication	109	25	0	21
	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED ≥ 3 and No Contraindication	49	5	0	11
	CHA$_2$DS$_2$-VASc <2 and No Contraindication	21	4	1	8
Structured+NLP with Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and Contraindication	2	0	0	1
	CHA$_2$DS$_2$-VASc > AND HAS-BLED ≥ 3 and Contraindication	6	2	0	1
	CHA$_2$DS$_2$-VASc <2 and Contraindication	0	0	0	0
Structured+NLP with No Contraindication	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED <3 and No Contraindication	53	17	1	8
	CHA$_2$DS$_2$-VASc >2 AND HAS-BLED ≥ 3 and No Contraindication	113	13	0	23
	CHA$_2$DS$_2$-VASc <2 and No Contraindication	12	4	0	8

AI Biosurveillance:
Population of NVAF in the USA

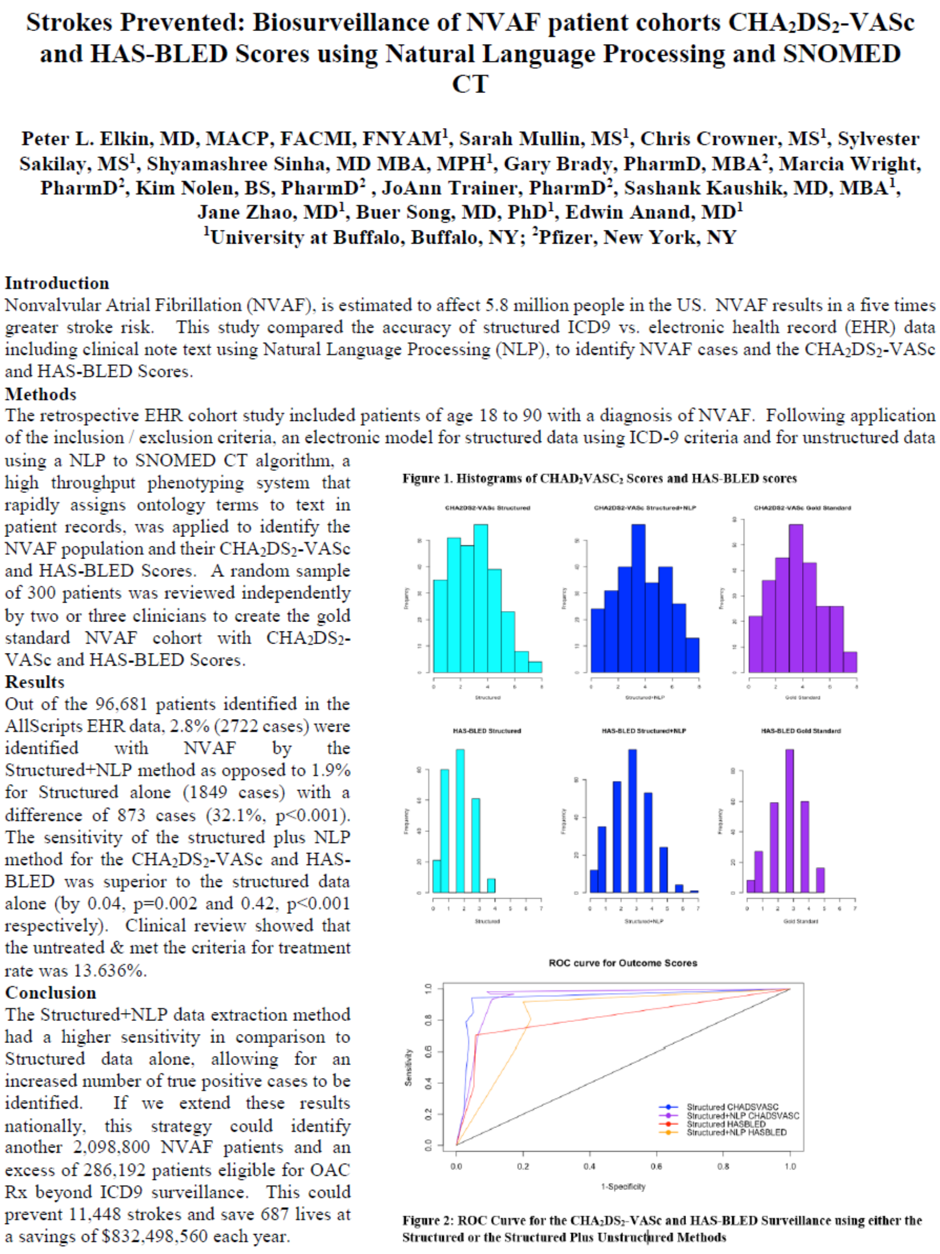
Population for Rates	Truven	Optum	Total	Event Rates in %
1. All the patients enrolled during Oct 2015 - Sep 2016	32,046,193	31,249,927	63,296,120	
2. (1) and age>=18 in 2016	25,400,465			
3. (2) and with any diagnosis of AF during Oct 2015 - Sep 2016 (first = index date)	422,092	865,072	1,287,164.00	
4. (3) and without VHD diagnosis during 1-year pre-index	355,811	611,990	967,801.00	1.52%
5. (4) and CHADS-VASc >= 2 and no contraindications to OAC	276,465	539,775	816,240.00	84.34%
6. (5) and Untreated	179,441	316,308	495,749.00	60.74%
Stroke Rate	11,530	10491	22,021.00	4.44%
Death Rate	727	593	1,320.00	5.99%

Cost the Year After Stroke	Costs the Year Prior to the Stroke	PMPM Difference	PMPM Inflation adjusted Difference	Annual PM Inflation adjusted Difference
$11,130.30	$2,665.40	$ 8,464.90	$ 8,253.42	$ 99,041.00

Artificial Intelligence Based Disease Surveillance: The Case of NVAF

Extrapolated Results	Structured	Structured Plus Unstructured	Difference Between the Two Methods
NVAF Population	4,955,284	6,754,052	1,798,768
NVAF Population with no contraindications and CHA2DS2-VASc >= 2	4,543,995	6,193,466	1,649,470
NVAF Population needing Treatment	3,009,840	4,102,411	1,092,572
Strokes Prevented	133,637	182,147	48,510
Deaths Prevented	8,005	10,911	2,906
Cost Savings*	$ 13,235,529,625.06	$ 18,040,026,878.96	$ 4,804,497,253.90
* Cost Basis is $99,041 / Untreated Ischemic Stroke's 1st year after event Cost (1.9% Inflation Adjusted)			

Circulation, 2017
Presented at the American Heart Association Meeting



Strokes Prevented: Biosurveillance of NVAF patient cohorts CHA_2DS_2-VASc and HAS-BLED Scores using Natural Language Processing and SNOMED CT

Peter L. Elkin, MD, MACP, FACMI, FNYAM[1], Sarah Mullin, MS[1], Chris Crowner, MS[1], Sylvester Sakilay, MS[1], Shyamashree Sinha, MD MBA, MPH[1], Gary Brady, PharmD, MBA[2], Marcia Wright, PharmD[2], Kim Nolen, BS, PharmD[2] , JoAnn Trainer, PharmD[2], Sashank Kaushik, MD, MBA[1], Jane Zhao, MD[1], Buer Song, MD, PhD[1], Edwin Anand, MD[1]
[1]University at Buffalo, Buffalo, NY; [2]Pfizer, New York, NY

Introduction

Nonvalvular Atrial Fibrillation (NVAF), is estimated to affect 5.8 million people in the US. NVAF results in a five times greater stroke risk. This study compared the accuracy of structured ICD9 vs. electronic health record (EHR) data including clinical note text using Natural Language Processing (NLP), to identify NVAF cases and the CHA_2DS_2-VASc and HAS-BLED Scores.

Methods

The retrospective EHR cohort study included patients of age 18 to 90 with a diagnosis of NVAF. Following application of the inclusion / exclusion criteria, an electronic model for structured data using ICD-9 criteria and for unstructured data using a NLP to SNOMED CT algorithm, a high throughput phenotyping system that rapidly assigns ontology terms to text in patient records, was applied to identify the NVAF population and their CHA_2DS_2-VASc and HAS-BLED Scores. A random sample of 300 patients was reviewed independently by two or three clinicians to create the gold standard NVAF cohort with CHA_2DS_2-VASc and HAS-BLED Scores.

Results

Out of the 96,681 patients identified in the AllScripts EHR data, 2.8% (2722 cases) were identified with NVAF by the Structured+NLP method as opposed to 1.9% for Structured alone (1849 cases) with a difference of 873 cases (32.1%, p<0.001). The sensitivity of the structured plus NLP method for the CHA_2DS_2-VASc and HAS-BLED was superior to the structured data alone (by 0.04, p=0.002 and 0.42, p<0.001 respectively). Clinical review showed that the untreated & met the criteria for treatment rate was 13.636%.

Conclusion

The Structured+NLP data extraction method had a higher sensitivity in comparison to Structured data alone, allowing for an increased number of true positive cases to be identified. If we extend these results nationally, this strategy could identify another 2,098,800 NVAF patients and an excess of 286,192 patients eligible for OAC Rx beyond ICD9 surveillance. This could prevent 11,448 strokes and save 687 lives at a savings of $832,498,560 each year.

Figure 1. Histograms of $CHAD_2VASC_2$ Scores and HAS-BLED scores

Figure 2: ROC Curve for the CHA_2DS_2-VASc and HAS-BLED Surveillance using either the Structured or the Structured Plus Unstructured Methods

Conclusions

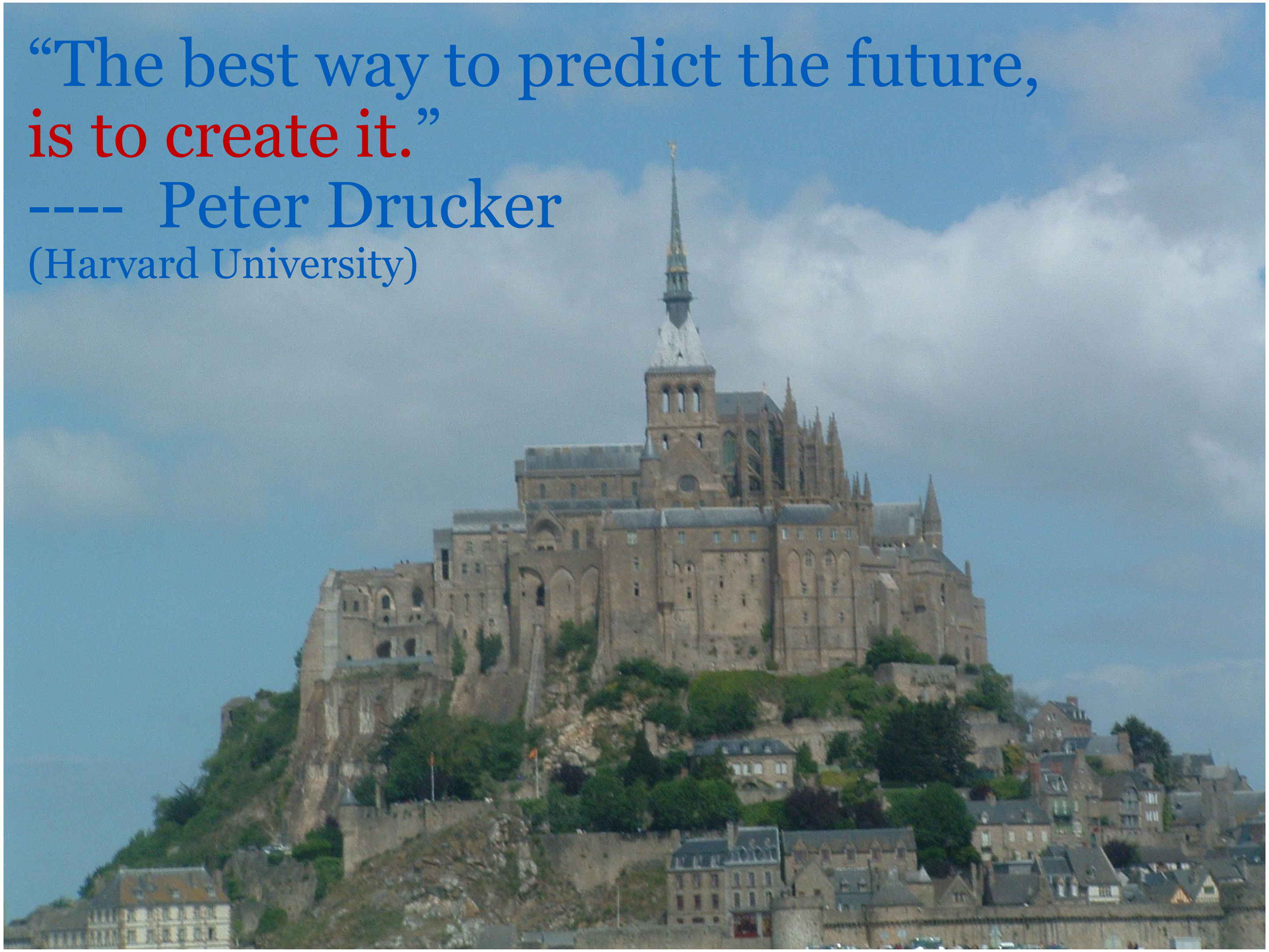
- Natural Language Processing is not only highly accurate, but also is now providing transaction speeds that make it practical for clinical applications.
- HTP-NLP is available for academic partnerships
- NLP is necessary to practically implement Semantic Interoperability
- Cross Validation of Data from a Variety of Datatypes is necessary to ensure accuracy
- Standardized Phenotypes can be shared and reused to ensure consistent population identification and data interoperability

Conclusions

- Clinical Decision Support assists clinicians in caring for their patients
- Biomedical Informatics partnering with Clinicians toward safer and more effective clinical care
- Biomedical Informatics as a Field deals with more than just computer in medicine
- Clinical Informatics is a new ABMS approved medical subspecialty that trains clinicians as future leaders of healthcare and healthcare organizations.

“…there is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things. Because the innovator has for enemies all those who have done well under the old conditions, and lukewarm defenders in those who may do well under the new. “

Nicolo Machiavelli c. 1505



"The best way to predict the future,
is to create it."
---- Peter Drucker
(Harvard University)

This program is supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412 to the University at Buffalo.