Regenstrief Institute
Better Care. Better Health.

Healthcare Data Analytics in the Age of Big Data: Real World Examples and Opportunities for the Future

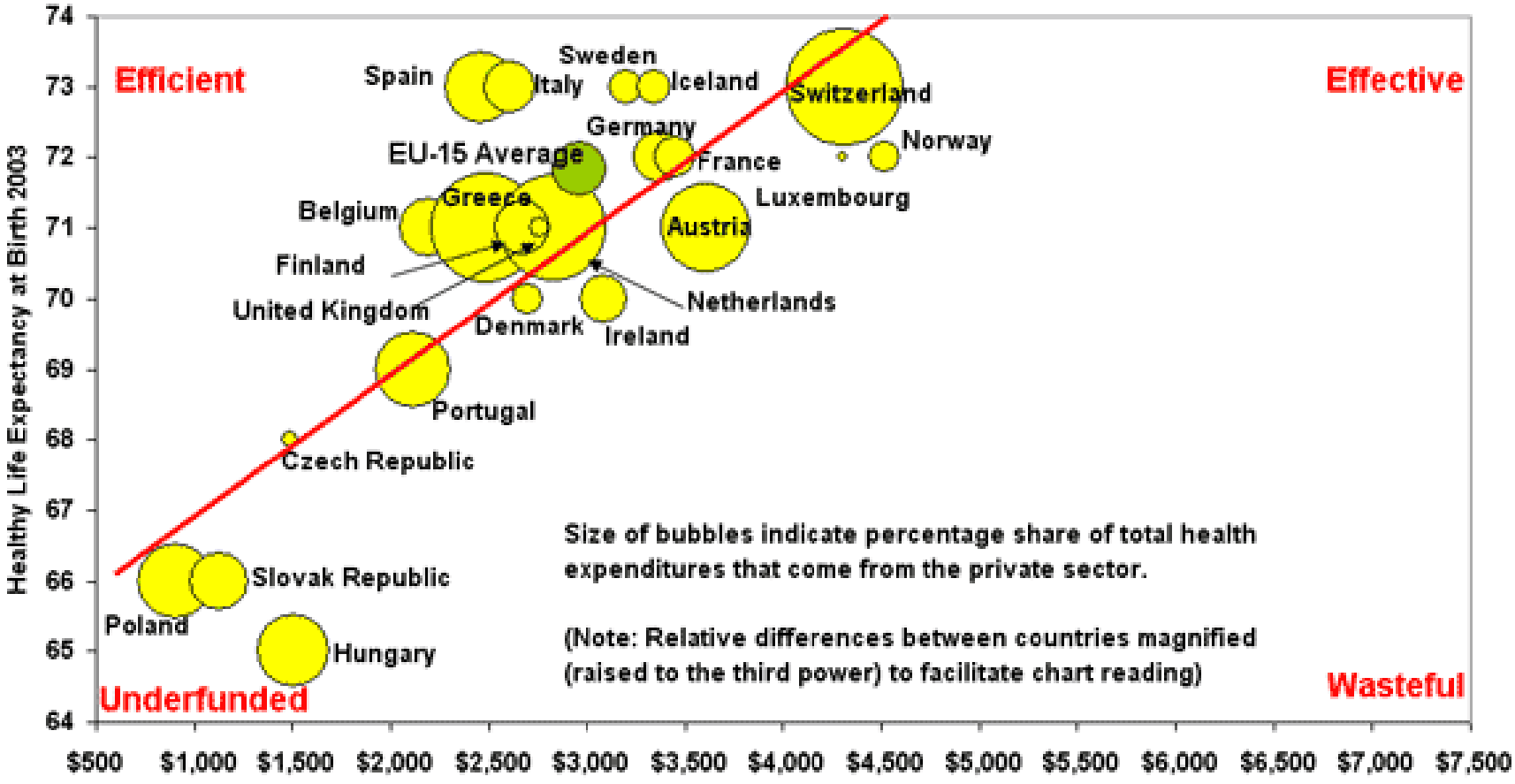
Shaun Grannis, Interim Director, Regenstrief Center for Biomedical Informatics

Associate Professor, Department of Family Medicine Indiana University School of Medicine

What We'll Cover

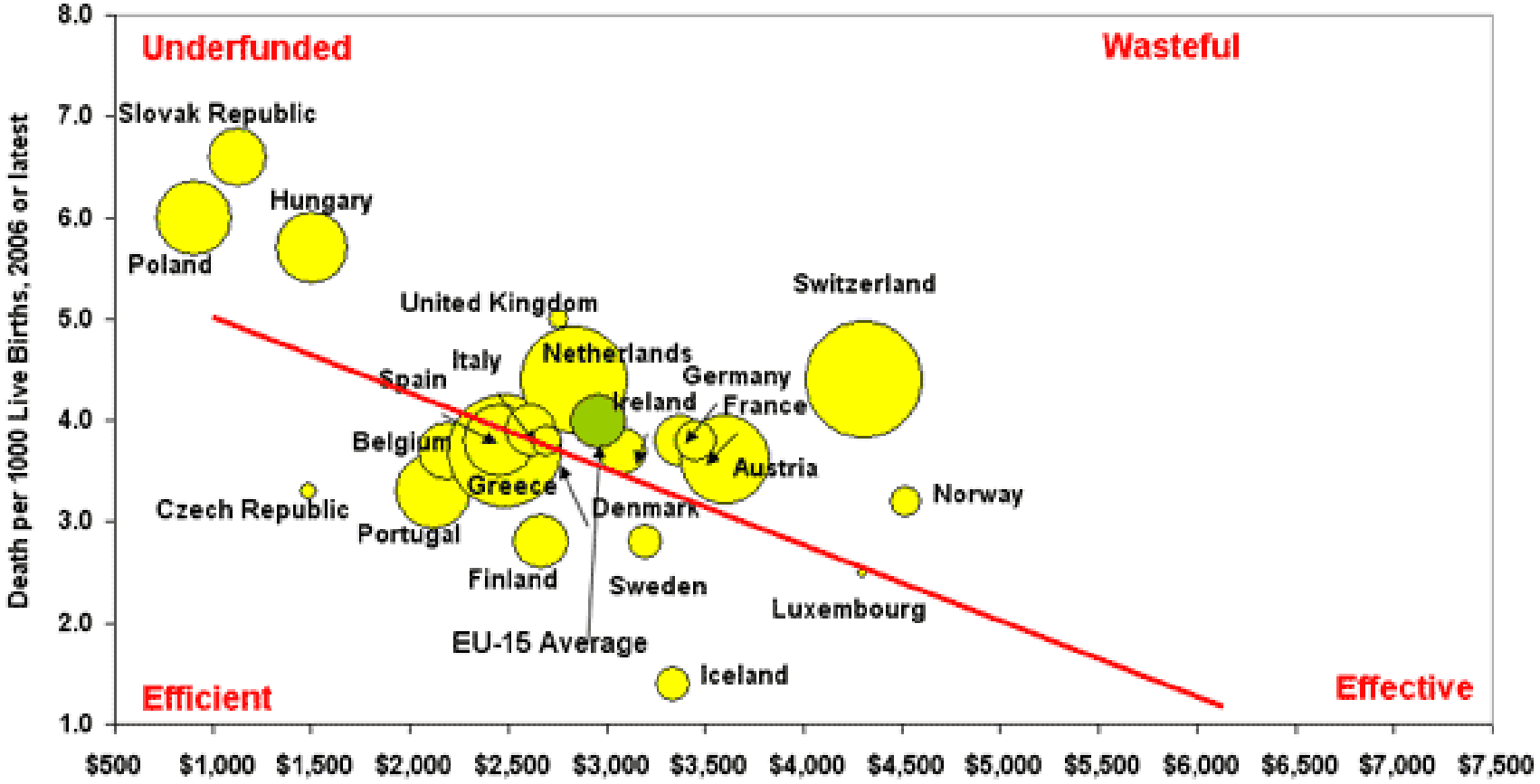
- Regenstrief
- The Problem
- Analytics
  - Phenotyping
  - Prediction
  - Examples putting it all together

Healthy Life Expectancy versus Expenditure per capita



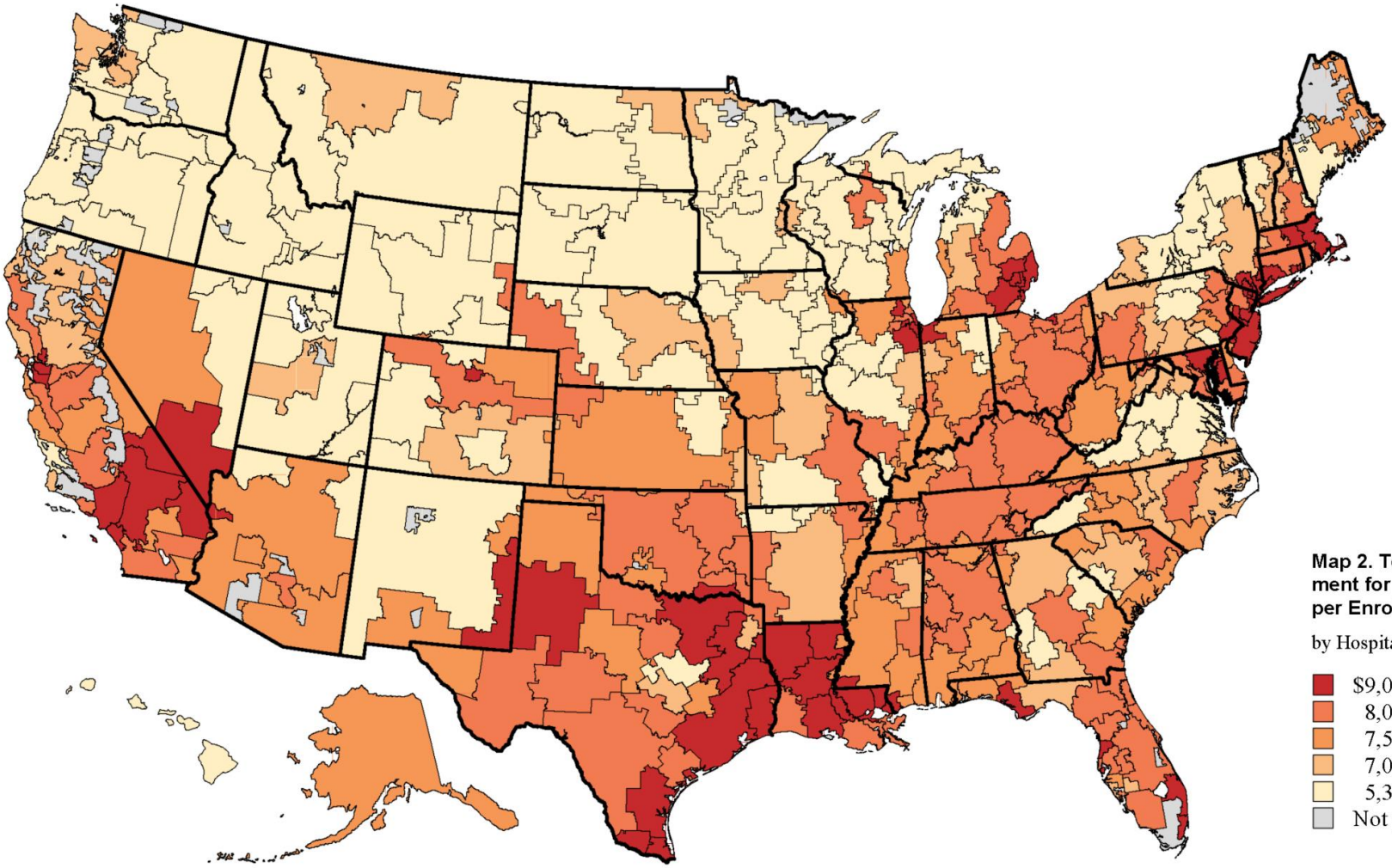
Source: OECD Health Database, June 2008 version; WHO World Health Data 2008; EU-15 average is the GDP weighted average

Infant Mortality versus Expenditure per capita



Source: OECD Health Database, June 2008 version; WHO World Health Data 2008; EU-15 average is the GDP weighted average

Variation in Medicare Reimbursement Rates

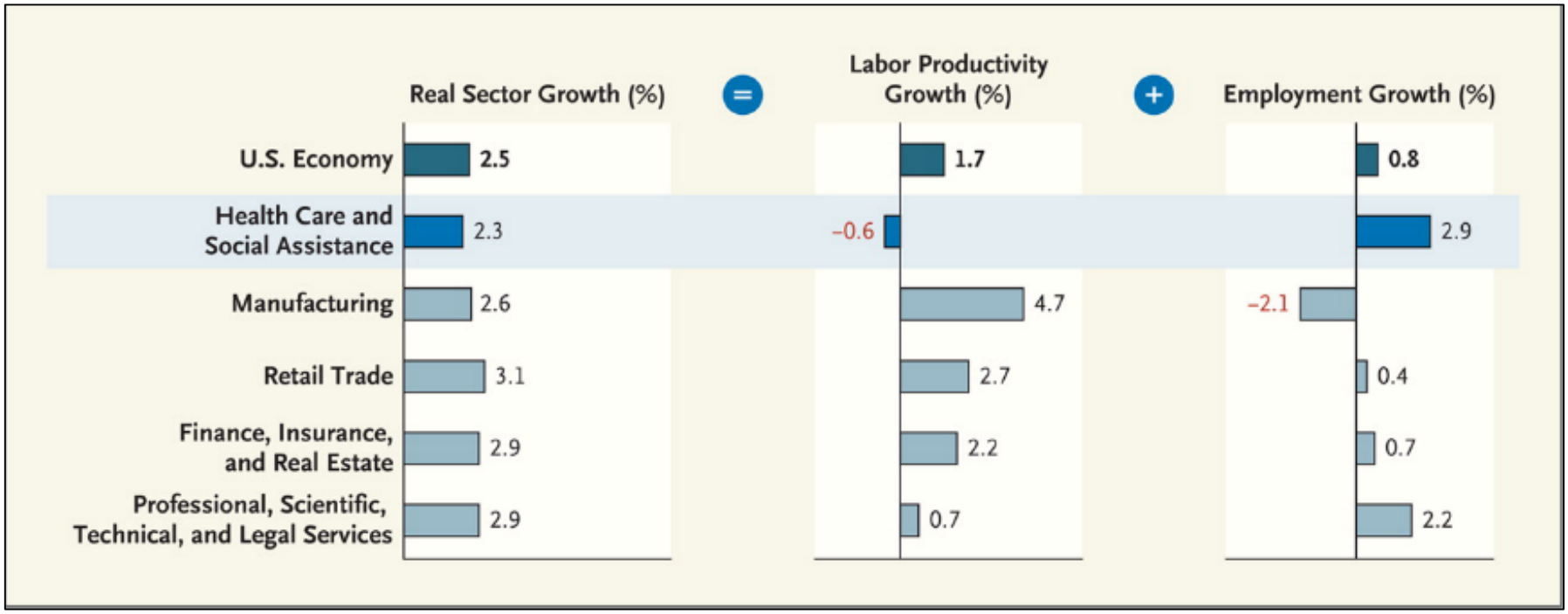


Map 2. Total Rates of Reimbursement for Noncapitated Medicare per Enrollee

by Hospital Referral Region (2006)

$9,000 to 16,352	(57)
8,000 to < 9,000	(79)
7,500 to < 8,000	(53)
7,000 to < 7,500	(42)
5,310 to < 7,000	(75)
Not populated	

Healthcare Labor Productivity



Real Sector Growth (Compound Annual Growth Rate), Broken into Labor Productivity Growth and Employment Growth in Various Sectors of the U.S. Economy, 1990–2010.

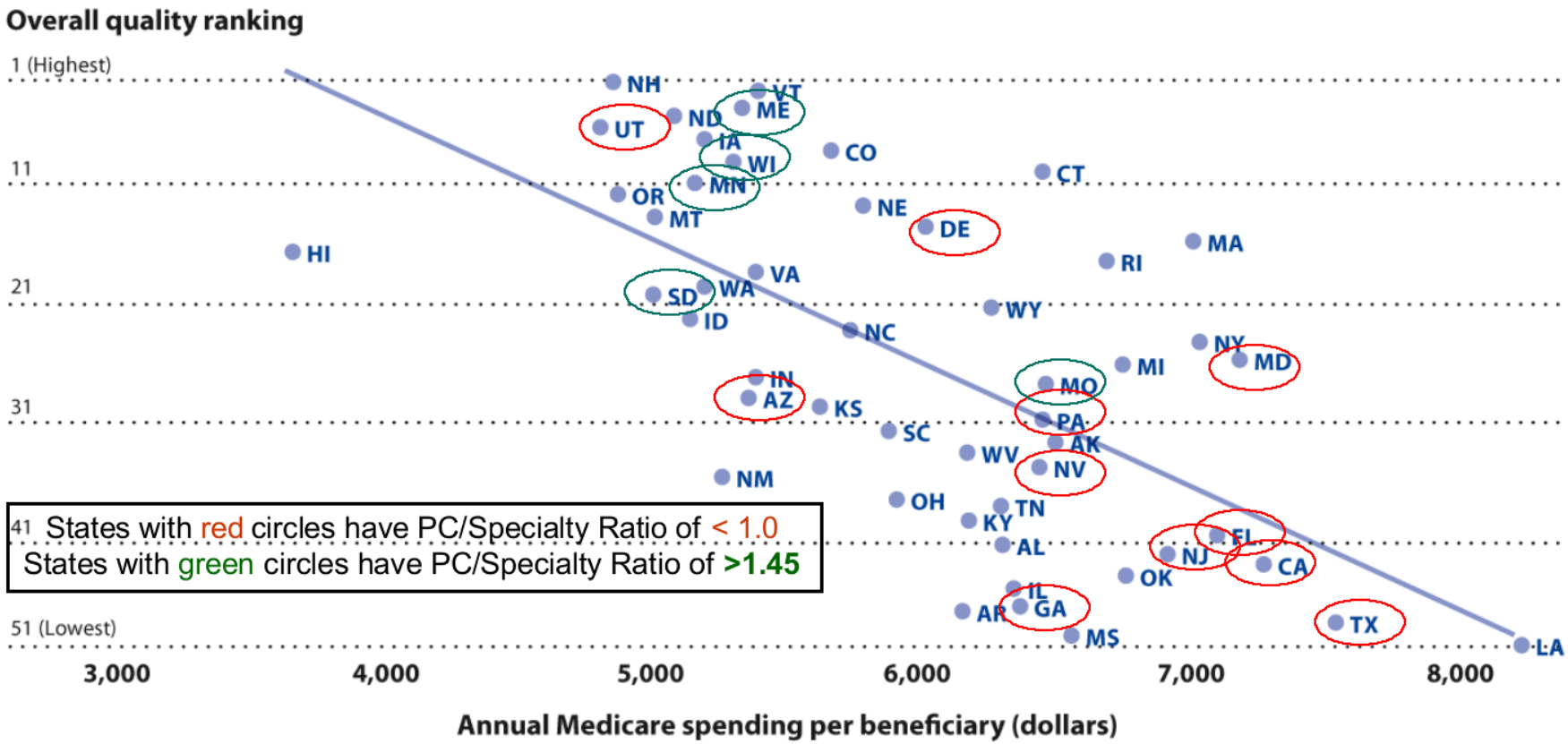
Real sector growth is defined as the value added by the industry to the gross domestic product. Data are from the Bureau of Labor Statistics and the Bureau of Economic Analysis.

Kocher R, Sahni NR. Rethinking Health Care Labor. N Engl J Med 2011; 365:1370-1372. October 13, 2011

Relationship Between Quality of Care and Medicare Spending

States with higher spending per Medicare beneficiary tended to rank lower on 22 quality of care indicators. This inverse relationship might reflect medical practice patterns that favor intensive, costly care rather than the effective care measured by these indicators.

Relationship between quality and Medicare spending, as expressed by overall quality ranking, 2000–2001

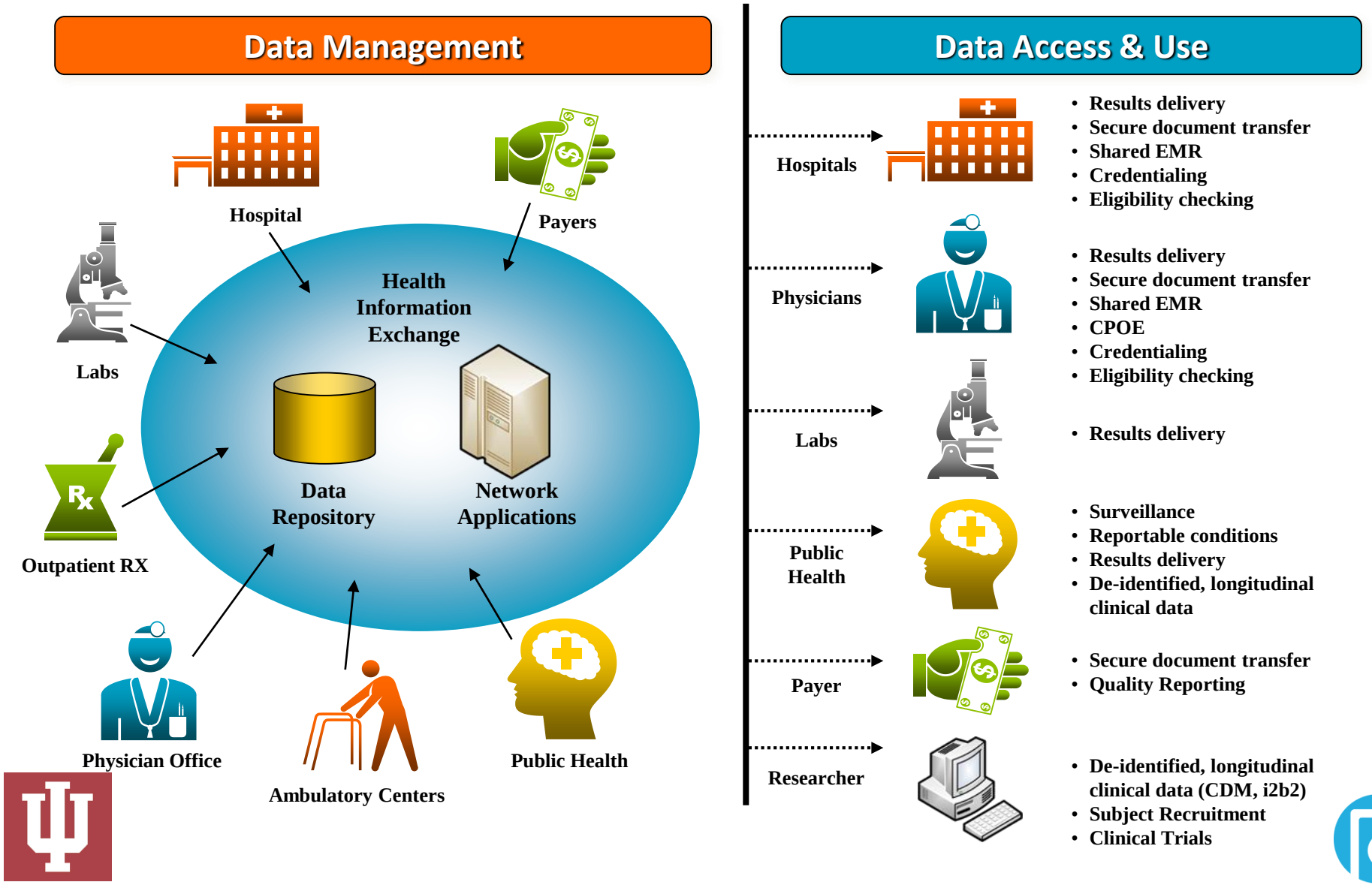


Source: Medicare administrative claims data and Medicare Quality Improvement Organization program data, as analyzed by Baicker and Chandra (2004). The solid line shows that for every $1,000 increase in Medicare spending per beneficiary, a state's quality ranking dropped by 10 positions. Adapted and republished with permission of *Health Affairs* from Baicker and Chandra, "Medicare spending, the physician workforce, and beneficiaries' quality of care" (Web Exclusive), 2004. Permission conveyed through the Copyright Clearance Center, Inc.

About Regenstrief Institute

- Regenstrief Institute was established in 1969 by philanthropist Samuel Regenstrief with the goal of improving the quality and efficiency of healthcare delivery, preventing medical errors, and enhancing patient safety.

EHR Integration: The Indiana Network for Patient Care (INPC)



Phenotype

- The composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, phenology, behavior, and products of behavior.
- A phenotype results from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two.

Source: wikipedia.org/wiki/Phenotype

- Phenotyping is the practice of developing algorithms designed to identify specific phenomic traits within an individual.
- These algorithms are created using multiple variables enabling researchers to accurately identify traits and perform analyses.

Source: emerge.mc.vanderbilt.edu

Phenotypic Map
Alcohol Abuse
Any Cancer
Asthma
Atrial Fibrillation
Bipolar Disorder
Breast Cancer
Chronic Kidney Disease
Chronic Obstructive Pulmonary Disease
Cirrhosis
Colorectal Cancer
Congestive Heart Failure
Coronary Artery Disease
Crohn's Disease
Dementia
Dental Caries
Depression
Diabetes Mellitus
Esophageal Cancer
Gastrointestinal Bleed
Hearing Loss
Hepatitis B
Hepatitis C
HIV/AIDS
Hyperlipids
Hypertension
Hyperthyroid
Hypothyroid
Irritable Bowel Syndrome
Leukemia
Liver Cancer
Low Back Pain
Lung Cancer
Melanoma
Myeloma
Myocardial Infarction
Neurofibromatosis
Obesity
Osteoarthritis
Osteoporosis
Ovarian Cancer
Pancreatic Cancer
Prostate Cancer
Psoriasis
Rheumatoid Arthritis
Schizophrenia
Seizure
Smoker
Stroke
Testicular Cancer
Thymoma
Ulcerative Colitis

How are Phenotypes Used?

- Identify cohorts to support a variety of research objectives
  - Recruitment (prospective)
  - Analyses (retrospective/prospective)
- Clinical decision support
- Case management/population health support

Secondary Use

“Primary” Use of Clinical Data

Healthcare providers record clinical data to support the processes necessary to conduct and sustain individual patient care:

- Aid information recall and decision making
- Convey information to other members of the healthcare team
- Provide medical-legal support for clinical decisions
- Support reimbursement processes

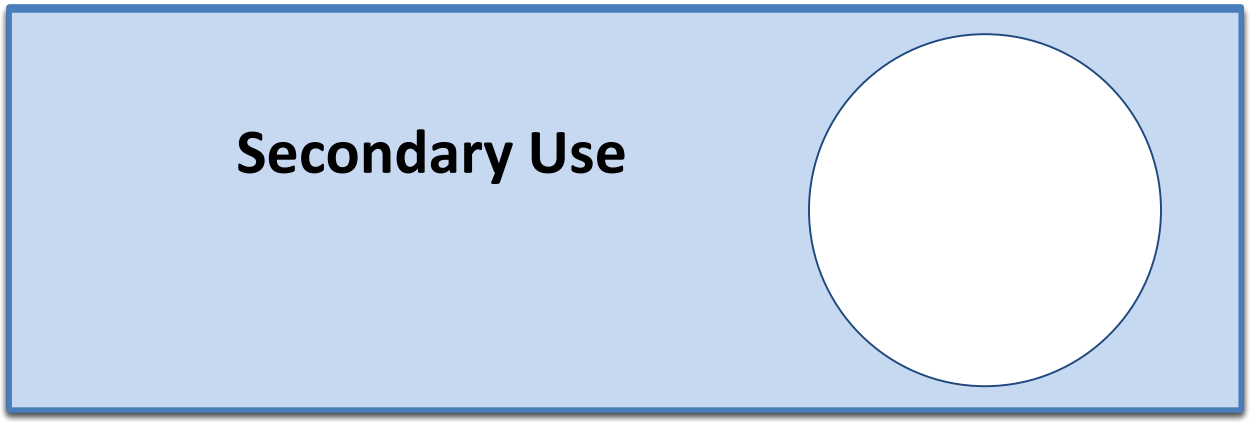
Secondary Use of Clinical Data

- Applies personal health information (PHI) for uses outside of direct health care delivery.
- It includes such activities as analysis, research, quality and safety measurement, public health, payment, provider certification or accreditation, marketing, and other business applications, including strictly commercial activities.

Source: Safran, et al. JAMIA (2007)

Secondary Use of Clinical Data

- Any use of health data not specifically intended or anticipated when data is initially obtained.
- Secondary use is defined in the complementary sense (i.e., NOT primary use.)



Bias Example: Diabetes and Obesity Cohort

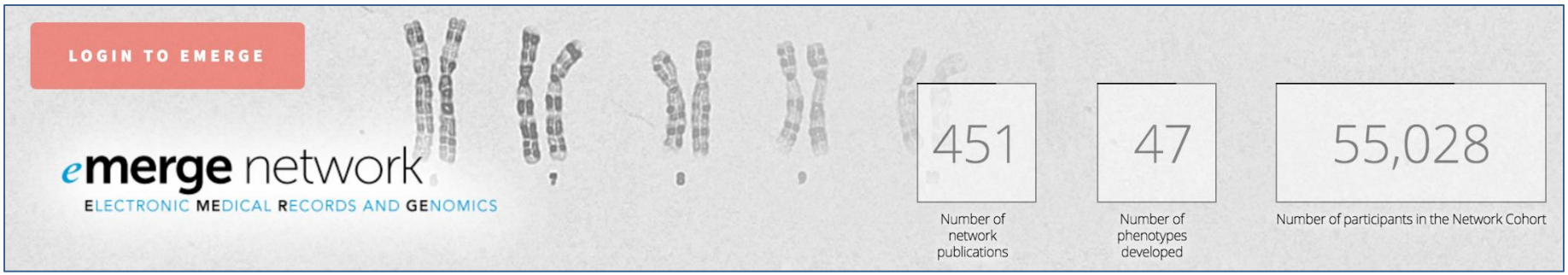
"Coders should pay attention to the BMI because it makes a difference in terms of reimbursement […]. A BMI of 40 or higher—diagnosis code V85.4—is considered a complicating condition, meaning higher reimbursement when reporting this code along with the appropriate principal diagnosis."

source: medicalcodingpro.wordpress.com/page/2/

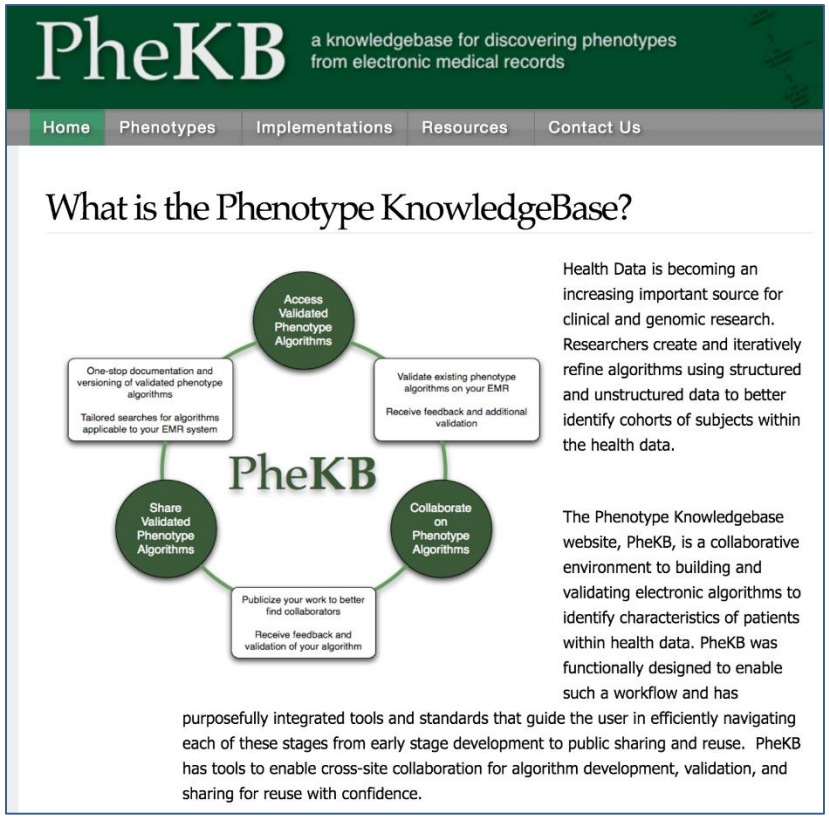
Data often reflect the financial incentives, not the true population distribution.

Counts reflect # of unique patients per cell. Queries are based on having these ICD9s ever in the INPC (and having no death date), except as noted below. Patients can appear in multiple rows and multiple columns, except as noted below.		Diabetes 250.*	Obesity 278.0*	Both Diabetes and Obesity 250.* AND 278.0*
HTN	401.*, 402.*, 403.*, or 404.*	333,483	164,544	73,877
CHF	428.*	61,198	24,774	17,557
Coronary Disease	410.*, 411.*, 412.*, 413.*, or 414.*	115,888	43,187	27,066
Stroke	433.x1 or 434.x1	18,475	5,726	3,953
Osteoarthritis	715.*	116,030	72,656	35,499
Only the most recent value from this entire list of V codes is shown for each patient.	V85.0*	177	13	5
	V85.1*	54	30	4
	V85.2	0	1	1
	V85.21	19	52	8
	V85.22	23	68	5
	V85.23	37	85	14
	V85.24	26	106	15
	V85.25	49	109	23
	V85.3	0	0	0
	V85.31	99	285	64
	V85.32	102	226	67
	V85.33	91	264	60
	V85.34	90	248	59
	V85.35	111	287	85
	V85.36	120	287	94
	V85.37	126	345	109
	V85.38	115	313	98
	V85.39	125	313	113
	V85.4	5,732	9,300	4,832
	V85.5	0	0	0
	V85.51	2	5	0
	V85.52	7	33	4
	V85.53	5	47	2
	V85.54	110	704	83
	V85.55	0	0	0

Phenotyping Initiatives



Source: emerge.mc.vanderbilt.edu



Source: phekb.org

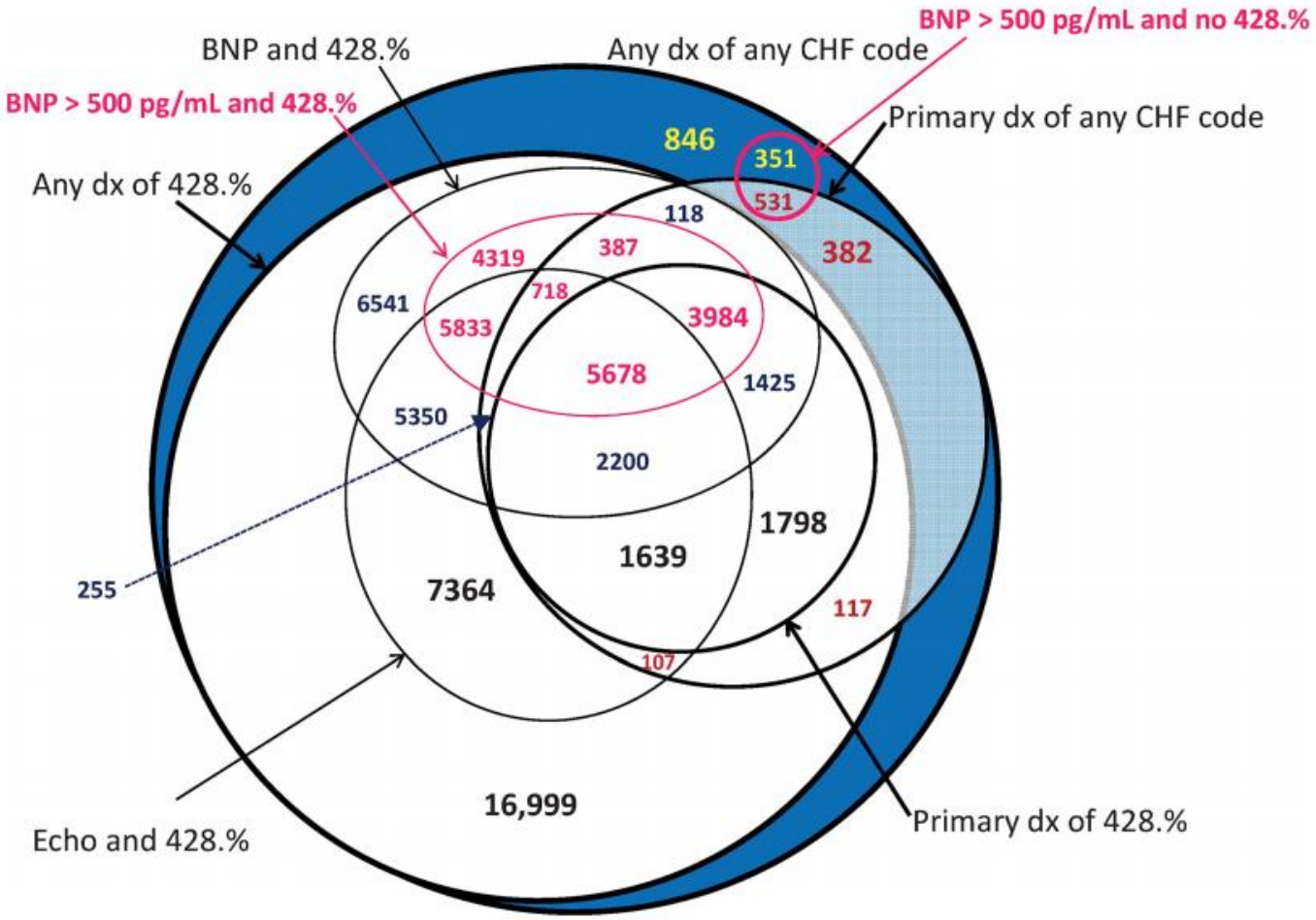
Table 3 Positive predictive value for phenotype case and control algorithms, and for phenotype eligibility algorithms across Electronic Medical Records and Genomics sites

Phenotype	Number validated	Positive predictive value				
		Group Health (%)	Marshfield Clinic (%)	Mayo Clinic (%)	Northwestern University (%)	Vanderbilt University (%)
Validated for case/control status and eligibility						
Cataract						
Case*	3234		97.7			96.0
Control*	3184		97.7			
Dementia						
Case*	3778	73.0	89.7			84.0
Control	505		96.7			
Type 2 diabetes						
Case	300		99.0		98.2	100
Control	143		98.0		100	100
Diabetic retinopathy						
Case	229		80.0			67.6
Control	80		98.0			100
Resistant hypertension						
Case	354	90.0	100		84.4	
Control	144	91.0	100		93.8	84.0
Peripheral arterial disease						
Case*	11504		87.5	90.7	95.0	
Control	100			100		
Chronic autoimmune hypothyroidism						
Case	389	92.0	91.3	82.0	98.1	98.0
Control	290	100	100	96.0	100	100
Validated for eligibility						
Low level of high density lipoprotein	440		81.6			
Lipids*	1054		78.8		92.3	
Red blood cell indices	391		96.4	98.0	98.0	
White blood cell indices	365		89.6		85.0	
QRS duration	245		100		96.9	97.0
Height	579		86.9		95.1	

Blank cells if did not participate in validation of that phenotype.
*Number large due to pre-existing study with validation.

Newton KM, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. JAMIA. (2013).

Heart Failure Phenotype Venn Diagram



Rosenman MR, et al. JAMIA. 2014

Phenotype Performance

Table 3 Results for the 10 congestive heart failure (CHF) phenotype queries

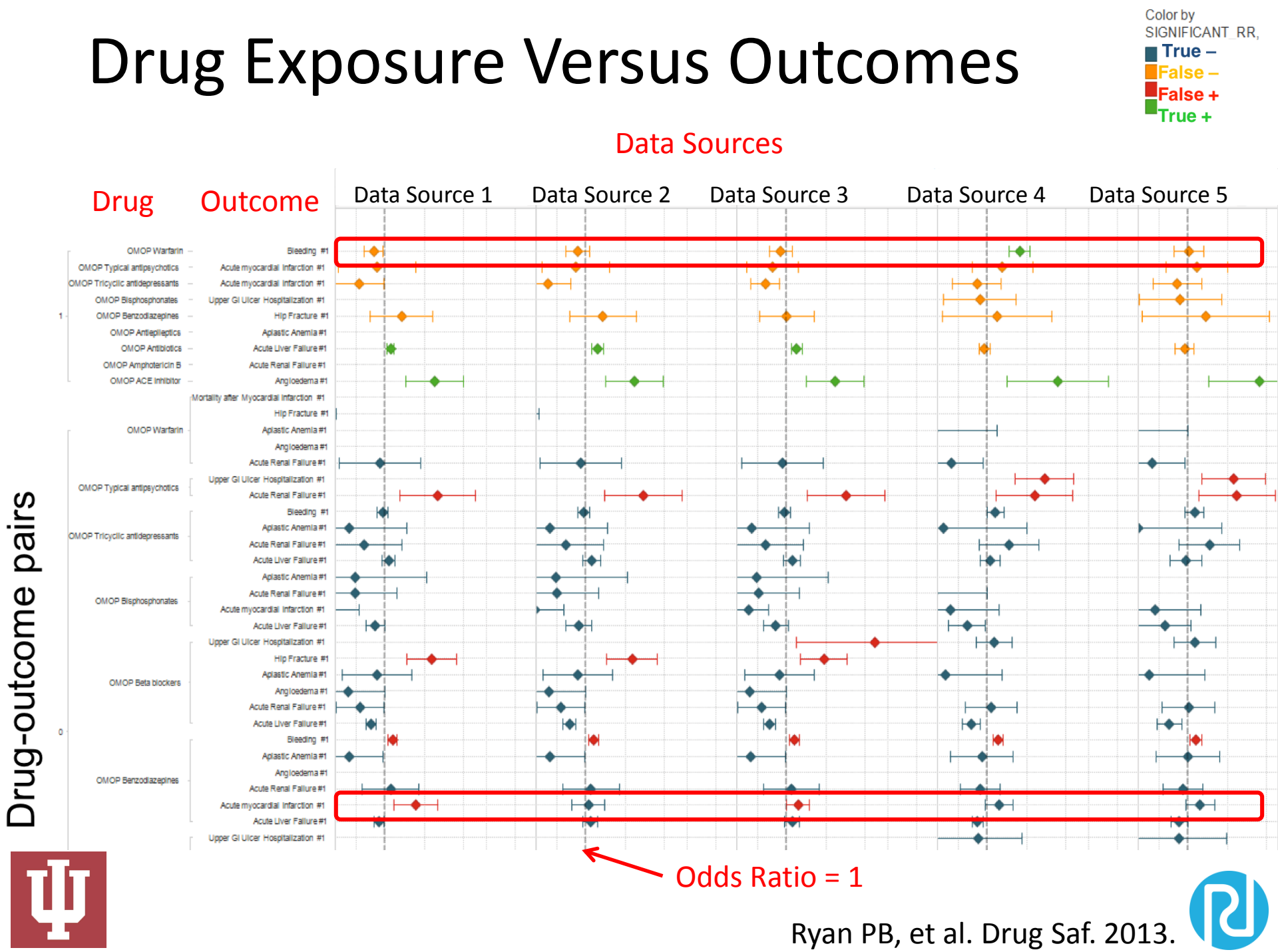
Criteria to combine Venn diagram zones	N in query	Sensitivity (%)	Sensitivity, SE (%)	PPV (%)	PPV, SE (%)
Any CHF	66 942	94.3	1.3	42.8	1.5
Any dx of 428	64 832	90.9	1.3	42.5	1.5
Any dx of CHF and BNP >500 pg/mL	21 801	50.8	1.8	70.7	2.5
1^0 dx of any CHF	19 339	54.8	1.9	86.0	2.2
1^0 dx of 428	16 724	47.6	1.7	86.3	2.5
1^0 dx of any CHF and BNP >500 pg/mL	11 298	33.5	1.3	90.0	2.1
1^0 dx of 428 and BNP >500 pg/mL	9662	28.8	1.1	90.4	2.4
1^0 dx of 428 and BNP >500 pg/mL and echocardiogram	5678	16.2	0.8	86.6	3.5
1^0 dx of any CHF or BNP >500 pg/mL	29 587	71.4	2.1	73.3	2.2
1^0 dx of 428 or BNP >500 pg/mL	28 863	69.6	2.1	73.2	2.2
High BNP, no ICD-9 diagnosis for CHF					
Zone X: no ICD-9 dx of 428, but BNP >500 pg/mL	12 149	N/A	N/A	14.3	3.5

BNP, B-natriuretic peptide; PPV, positive predictive value.

Rosenman MR, et al. JAMIA. 2014

OMOP

Drug Exposure Versus Outcomes

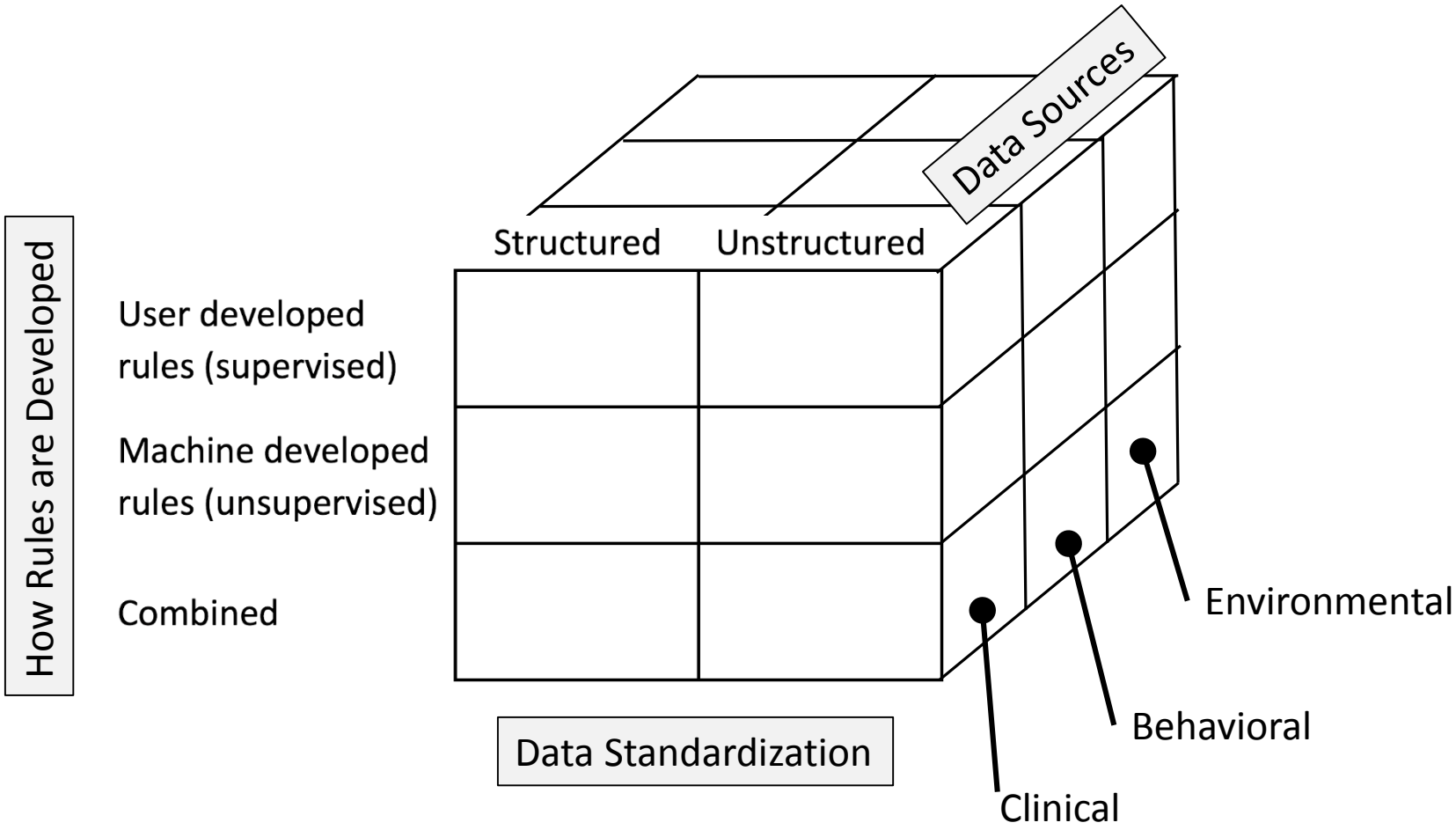


Ryan PB, et al. Drug Saf. 2013.

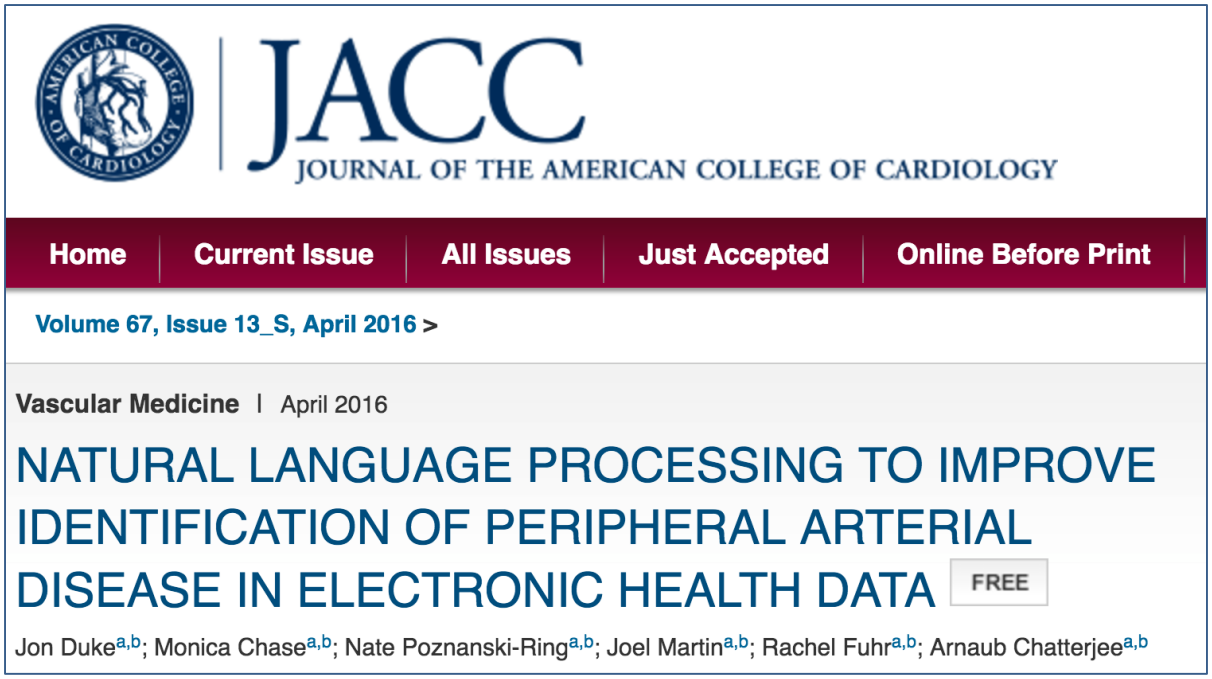
Framing Features for the Future of Phenotyping

- How rules are developed (discovered)?
  - Human, Machine, Both
- How data is standardized
  - Structured, Unstructured
- Sources of data
  - Clinical, Behavioral, Environmental

Framework for the Future of Phenotyping



Structured and Unstructured Data: Peripheral Artery Disease



JACC
JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY

Home | Current Issue | All Issues | Just Accepted | Online Before Print

Volume 67, Issue 13_S, April 2016 >

Vascular Medicine | April 2016

NATURAL LANGUAGE PROCESSING TO IMPROVE IDENTIFICATION OF PERIPHERAL ARTERIAL DISEASE IN ELECTRONIC HEALTH DATA FREE

Jon Duke[a,b]; Monica Chase[a,b]; Nate Poznanski-Ring[a,b]; Joel Martin[a,b]; Rachel Fuhr[a,b]; Arnaub Chatterjee[a,b]

"… identification of PAD patients using NLP was significantly more robust (4-fold) compared to use of structured data alone"

Machine Learning Breaks Bottleneck In Gathering Valuable Information About Cancer

Kevin Murnane, CONTRIBUTOR

I write about technology, science and video games **FULL BIO**

Opinions expressed by Forbes Contributors are their own.

Credit: Getty Images

Machine learning's ability to produce actionable results from unstructured data is clearly demonstrated in a study published in

Example: Identifying Cancer from Free-Text

Journal of Biomedical Informatics 60 (2016) 145–152

Contents lists available at ScienceDirect

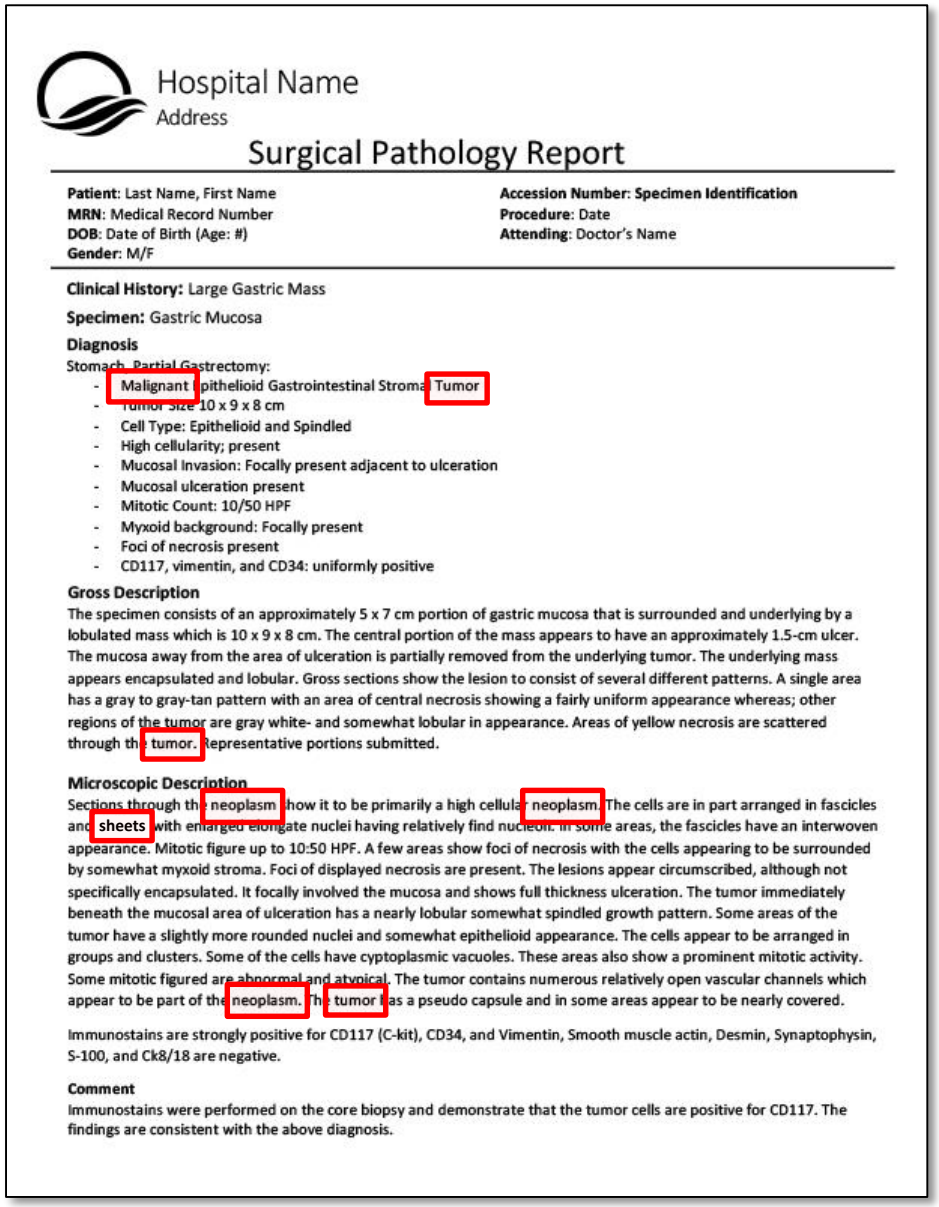
Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection

Suranga N. Kasthurirathne [a,*], Brian E. Dixon [b,c], Judy Gichoya [d], Huiping Xu [c], Yuni Xia [d], Burke Mamlin [b,d], Shaun J. Grannis [b,d]

Machine Learning Example: Cancer Phenotype



Hospital Name
Address

Surgical Pathology Report

Patient: Last Name, First Name
MRN: Medical Record Number
DOB: Date of Birth (Age: #)
Gender: M/F

Accession Number: Specimen Identification
Procedure: Date
Attending: Doctor's Name

Clinical History: Large Gastric Mass

Specimen: Gastric Mucosa

Diagnosis

Stomach, Partial Gastrectomy:

- Malignant Epithelioid Gastrointestinal Stromal Tumor
- Tumor Size 10 x 9 x 8 cm
- Cell Type: Epithelioid and Spindled
- High cellularity; present
- Mucosal Invasion: Focally present adjacent to ulceration
- Mucosal ulceration present
- Mitotic Count: 10/50 HPF
- Myxoid background: Focally present
- Foci of necrosis present
- CD117, vimentin, and CD34: uniformly positive

Gross Description

The specimen consists of an approximately 5 x 7 cm portion of gastric mucosa that is surrounded and underlying by a lobulated mass which is 10 x 9 x 8 cm. The central portion of the mass appears to have an approximately 1.5-cm ulcer. The mucosa away from the area of ulceration is partially removed from the underlying tumor. The underlying mass appears encapsulated and lobular. Gross sections show the lesion to consist of several different patterns. A single area has a gray to gray-tan pattern with an area of central necrosis showing a fairly uniform appearance whereas; other regions of the tumor are gray white- and somewhat lobular in appearance. Areas of yellow necrosis are scattered through the tumor. Representative portions submitted.

Microscopic Description

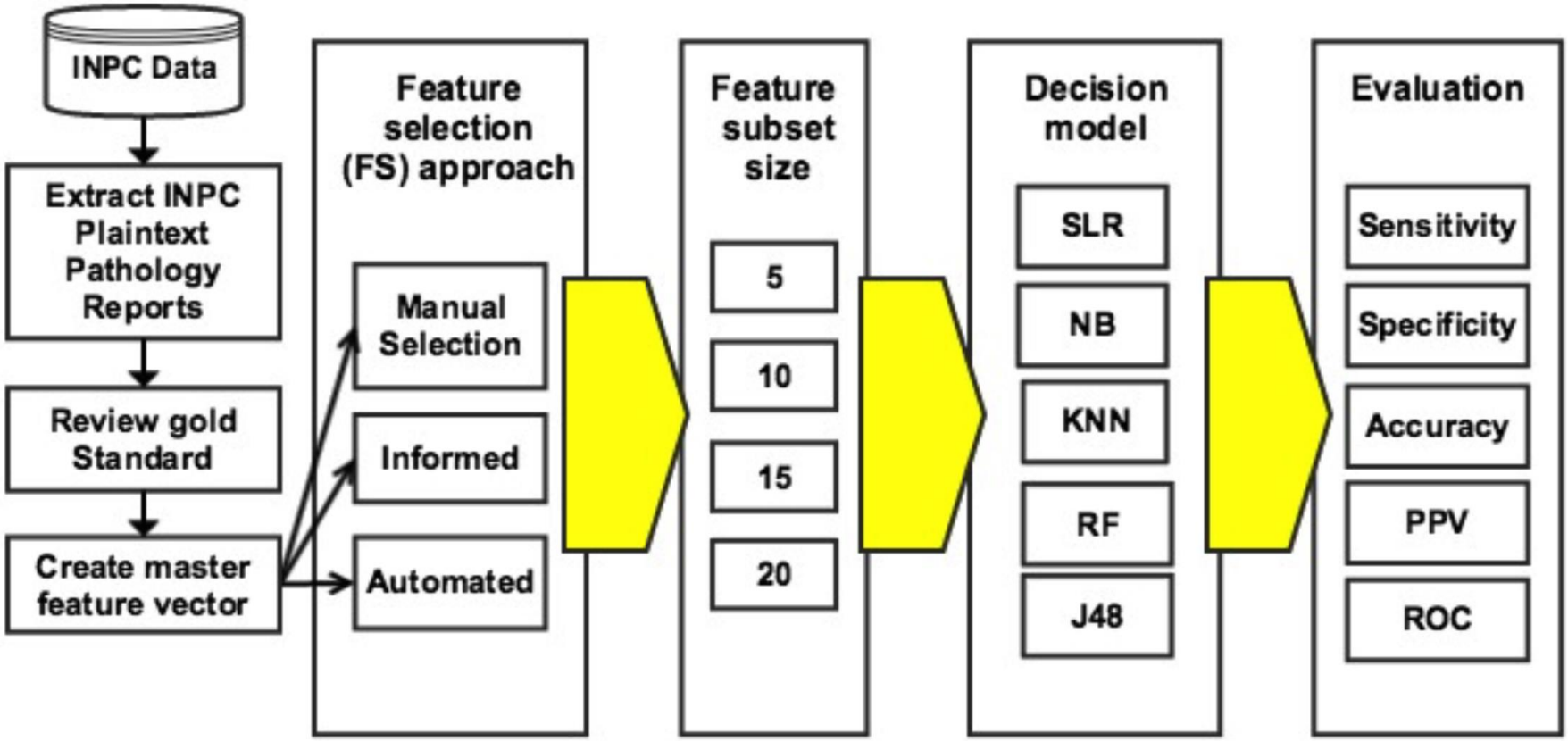
Sections through the neoplasm show it to be primarily a high cellular neoplasm. The cells are in part arranged in fascicles and sheets with enlarged elongate nuclei having relatively find nucleoli. In some areas, the fascicles have an interwoven appearance. Mitotic figure up to 10:50 HPF. A few areas show foci of necrosis with the cells appearing to be surrounded by somewhat myxoid stroma. Foci of displayed necrosis are present. The lesions appear circumscribed, although not specifically encapsulated. It focally involved the mucosa and shows full thickness ulceration. The tumor immediately beneath the mucosal area of ulceration has a nearly lobular somewhat spindled growth pattern. Some areas of the tumor have a slightly more rounded nuclei and somewhat epithelioid appearance. The cells appear to be arranged in groups and clusters. Some of the cells have cytoplasmic vacuoles. These areas also show a prominent mitotic activity. Some mitotic figured are abnormal and atypical. The tumor contains numerous relatively open vascular channels which appear to be part of the neoplasm. The tumor has a pseudo capsule and in some areas appear to be nearly covered.

Immunostains are strongly positive for CD117 (C-kit), CD34, and Vimentin, Smooth muscle actin, Desmin, Synaptophysin, S-100, and Ck8/18 are negative.

Comment

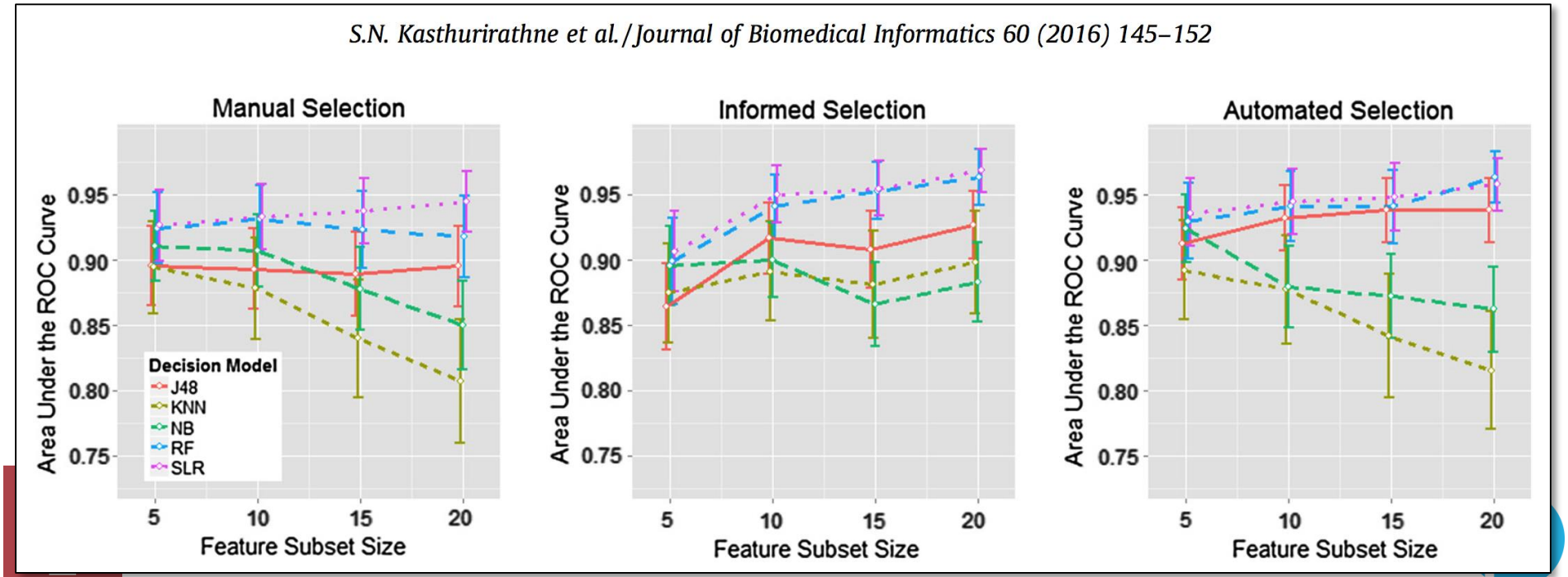
Immunostains were performed on the core biopsy and demonstrate that the tumor cells are positive for CD117. The findings are consistent with the above diagnosis.

Overall Process

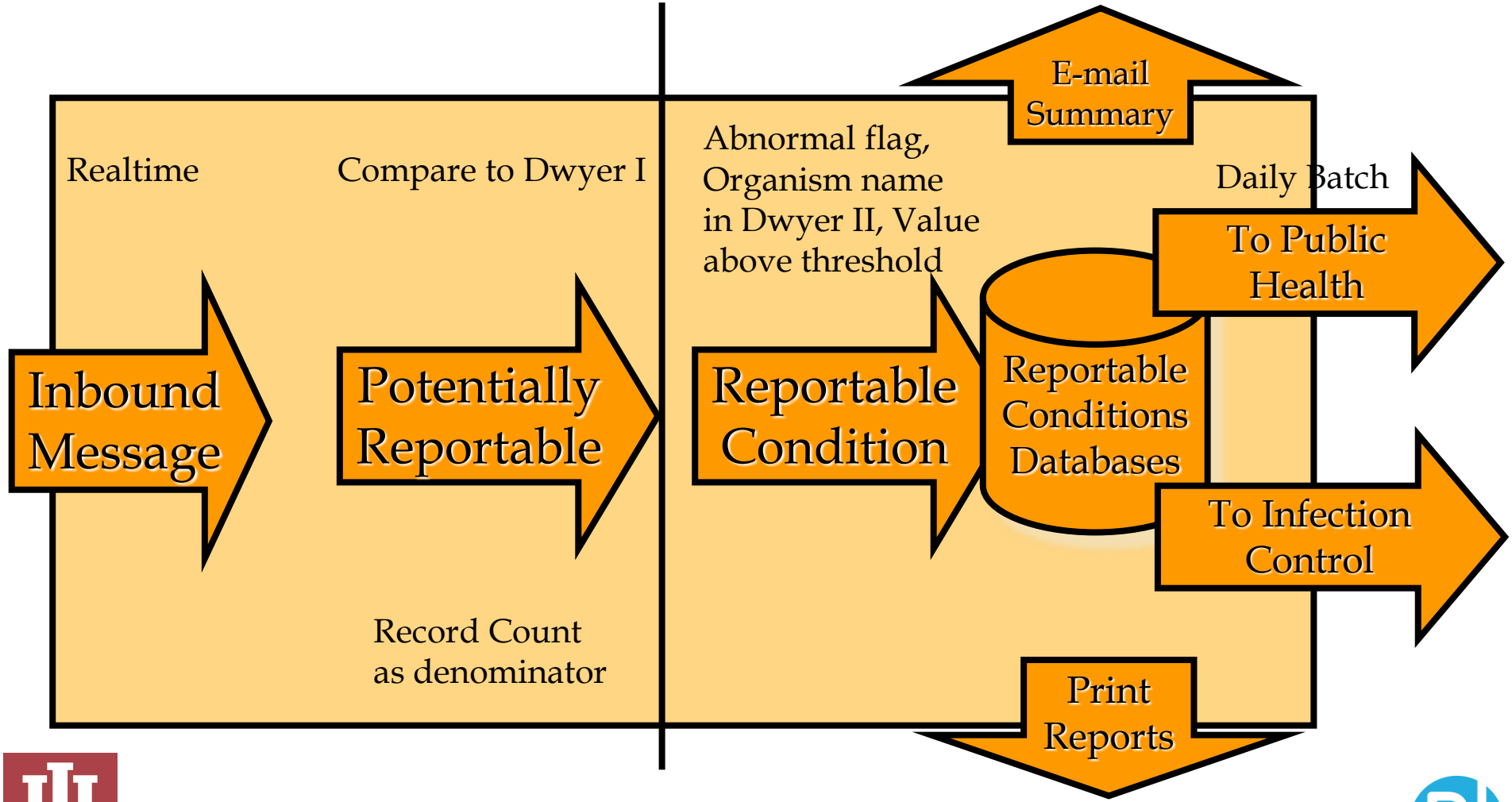


Results

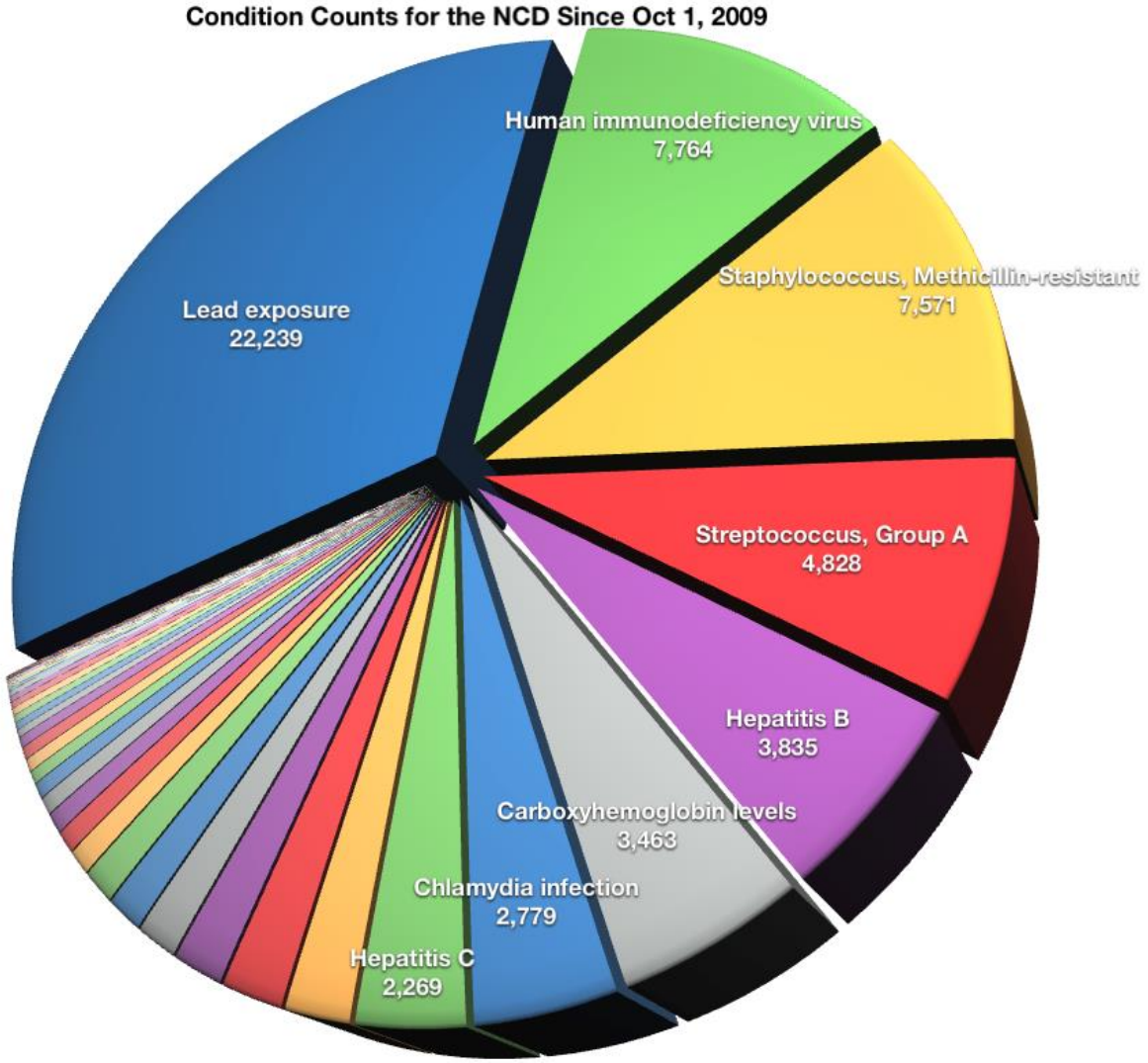
- Sensitivity: 85-90%
- Positive Predictive Value: 95-97%
- Specificity: 97-99%
- Random Forest and Logistic Regression exhibited highest AUC



System Overview: Notifiable Condition Detector



Condition	Count
Lead exposure	22239
Human immunodeficiency virus	7764
Staphylococcus, Methicillin-resistant	7571
Streptococcus, Group A	4828
Hepatitis B	3835
Carboxyhemoglobin levels	3463
Chlamydia infection	2779
Hepatitis C	2269
Sickle cell disease	1418
Escherichia coli O157:H7 infection	1393
Trichomoniasis	1190
Gonorrhea	997
Influenza	903
Measles	832
Herpes simplex type 1	600
Syphilis	528
Hepatitis A	488
Coccidioidomycosis	419
Chickenpox	369
Cytomegalovirus	308
Genital warts	308
Herpes simplex type 2	290
Fifth's disease	255
Legionellosis	179
Salmonellosis, non-typhoid	175
Mumps	174
Tuberculosis	165
Herpes simplex	158
Pertussis	137
Streptococcus pneumoniae	116
Histoplasmosis	73
Lyme disease	71
Toxoplasmosis	48
Streptococcus, Group B	38
Shigellosis	25
Giardiasis	23



- Lead exposure
- Hepatitis B
- Sickle cell disease
- Influenza
- Hepatitis A
- Genital warts
- Salmonellosis, non-typhoid
- Pertussis
- Toxoplasmosis
- Human T-lymphotrophic virus
- Rotavirus
- Blastomycosis
- Human immunodeficiency virus
- Carboxyhemoglobin levels
- Escherichia coli O157:H7 infection
- Measles
- Coccidioidomycosis
- Herpes simplex type 2
- Mumps
- Streptococcus pneumoniae
- Streptococcus, Group B
- Cryptococcosis
- Campylobacteriosis
- Hepatitis D
- Staphylococcus, Methicillin-resistant
- Chlamydia infection
- Trichomoniasis
- Herpes simplex type 1
- Chickenpox
- Fifth's disease
- Tuberculosis
- Histoplasmosis
- Shigellosis
- Q fever
- Bartonellosis
- Tularemia
- Streptococcus, Group A
- Hepatitis C
- Gonorrhea
- Syphilis
- Cytomegalovirus
- Legionellosis
- Herpes simplex
- Lyme disease
- Giardiasis
- Rubella
- Arbovirus

TABLE 2—Reporting Timeliness for Traditional Methods and Electronic Laboratory Reporting, by Condition: Indianapolis, Ind, 2001

Condition	Spontaneous Reporting (Public Health Department)	
	Average Lag Time, Days	No. of Cases
Campylobacteriosis	0.0	1
Chlamydial infection	10.0	363
Cryptosporidiosis	0.0	1
Escherichia coli O157:H7 infection	–1.0	1
Giardiasis	3.5	2
Hepatitis A	4.0	4
Hepatitis B	2.2	17
Hepatitis C	5.5	157
Histoplasmosis	–8.5	4
Salmonellosis, nontyphoid	–1.0	3
Shigellosis	0	24
Streptococcus: group A	1.2	5
Streptococcus: group B	20.0	2
Syphilis	4.4	17

Note. Negative values indicate that electronic laboratory reports were received later than were spontaneous reports on average. Only conditions for which timeliness could be calculated are included.

Timeliness

ELR identified cases 7.9 days earlier than did spontaneous reporting.

Salmonella Phenotype Example

```
OBX|1|CE|SDES^SPECIMEN DESCRIPTION^MIDAM|1|BLUDC^BLOOD(C) CENTRL LINE
   DRAW^SQS||||||F|||200809090810|^|
OBX|2|CE|SREQ^SPECIAL REQUESTS^MIDAM|1|OONLY^Aerobic bottle only
   received - unable to determine presence or absence
   of^SQD||||||F|||200809090810|^|
NTE|1|| strict anaerobic organisms.
OBX|3|CE|CULT^CULTURE^MIDAM|1|ENT^ENTEROCOCCUS SPECIES
   ^SQMO||||||F|||200809090810|^|
NTE|1||METHICILLIN RESISTANT STAPH AUREUS
NTE|2||PRESUMPTIVE
NTE|3||VANCOMYCIN RESISTANT ENTEROCOCCUS
NTE|4||GPC CALLED 9/10/08 1009AM AT 555-3244 TO MARY SMITH RN,
NTE|5||MRSA CALLED TO 555-3244 TO MARY SMITH, RN.  09/14/08  1044
```

“Stop” Words

```
OBX|1|CE|SDES^SPECIMEN DESCRIPTION^MIDAM|1|BLUDC^BLOOD(C) CENTRL LINE
   DRAW^SQS||||||F|||200809090810|^|
OBX|2|CE|SREQ^SPECIAL REQUESTS^MIDAM|1|OONLY^Aerobic bottle only
   received - unable to determine presence or absence
   of^SQD||||||F|||200809090810|^|
NTE|1|| strict anaerobic organisms.
OBX|3|CE|CULT^CULTURE^MIDAM|1|ENT^ENTEROCOCCUS SPECIES
   ^SQMO||||||F|||200809090810|^|
NTE|1||METHICILLIN RESISTANT STAPH AUREUS
NTE|2||PRESUMPTIVE
NTE|3||VANCOMYCIN RESISTANT ENTEROCOCCUS
NTE|4||GPC CALLED 9/10/08 1009AM AT 555-3244 TO MARY SMITH RN,
NTE|5||MRSA CALLED TO 555-3244 TO MARY SMITH, RN.  09/14/08  1044
```

“Stemmed” Words

```
OBX|1|CE|SDES^SPECIMEN DESCRIPTION^MIDAM|1|BLUDC^BLOOD(C) CENTRL LINE
   DRAW^SQS||||||F|||200809090810|^|
OBX|2|CE|SREQ^SPECIAL REQUESTS^MIDAM|1|OONLY^Aerobic bottle only
   received - unable to determine presence or absence
   of^SQD||||||F|||200809090810|^|
NTE|1|| strict anaerobic organisms.
OBX|3|CE|CULT^CULTURE^MIDAM|1|ENT^ENTEROCOCCUS SPECIES
   ^SQMO||||||F|||200809090810|^|
NTE|1||METHICILLIN RESISTANT STAPH AUREUS
NTE|2||PRESUMPTIVE
NTE|3||VANCOMYCIN RESISTANT ENTEROCOCCUS
NTE|4||GPC CALLED 9/10/08 1009AM AT 555-3244 TO MARY SMITH RN,
NTE|5||MRSA CALLED TO 555-3244 TO MARY SMITH, RN.  09/14/08  1044
```

Feature Vector

[SDES,SPECIMEN,DESCRIPTION,MIDAM,BLUDC,BLOOD(C),CENTRL,LINE,DRAW,SQS,
SREQ,SPECIAL,REQUESTS,MIDAM,OONLY,Aerobic,bottle,only,received,
unable,to,determine,presence,or,absence,of,SQD,strict,anaerobic,
organisms,CULT,CULTURE,MIDAM,ENT,ENTEROCOCCUS,SPECIES,SQMO,
METHICILLIN,RESISTANT,STAPH,AUREUS,PRESUMPTIVE,VANCOMYCIN,RESISTANT,
ENTEROCOCCUS,GPC,CALLED,AT,TO,MARY,SMITH,RN,MRSA,CALLED,TO,TO,MARY,
SMITH,RN]

Lower Case

[sdes,specimen,description,midam,bludc,blood(c),centrl,line,draw,sqs,
sreq,special,requests,midam,oonly,aerobic,bottle,only,received,
unable,to,determine,presence,or,absence,of,sqd,strict,anaerobic,
organisms,cult,culture,midam,ent,enterococcus,species,sqmo,
methicillin,resistant,staph,aureus,presumptive,vancomycin,resistant,
enterococcus,gpc,called,at,to,mary,smith,rn,mrsa,called,to,to,mary,
smith,rn]

Stop/Stemmed Removed

```
[sdes,specimen,description,midam,bludc,blood(c),centrl,line,draw,sqs,
sreq,special,request,midam,oonly,aerobic,bottle,only,receiv,
unable,determine,presence,absence,sqd,strict,anaerobic,
organism,cult,culture,midam,ent,enterococcus,species,sqmo,
methicillin,resistant,staph,aureus,presumptive,vancomycin,resistant,
enterococcus,gpc,call,mary,smith,rn,mrsa,call,mary,
smith,rn]
```

Sorted

[absence,aerobic,anaerobic,aureus,blood(c),bludc,bottle,call,call,
centrl,cult,culture,description,determine,draw,ent,enterococcus,
enterococcus,gpc,line,mary,mary,methicillin,midam,midam,midam,mrsa,
only,oonly,organism,presence,presumptive,receiv,request,resistant,
resistant,rn,rn,sdes,smith,smith,special,species,specimen,sqd,sqmo,
sqs,sreq,staph,strict,unable,vancomycin]

Tokenized

[a,b,c,d,e,f,g,h,h,i,j,k,l,m,n,o,p,p,q,r,s,s,t,u,u,u,v,w,x,y,z,aa,ab,
ac,ad,ad,ae,ae,af,ag,ag,ah,ai,aj,ak,al,am,an,ao,ap,aq,ar]

[a,b,c,d,e,f,g,2*h,i,j,k,l,m,n,o,2*p,q,r,2*s,t,3*u,v,w,x,y,z,aa,ab,ac
,2*ad,2*ae,af,2*ag,ah,ai,aj,ak,al,am,an,ao,ap,aq,ar]

[1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,2,1,1,2,1,3,1,1,1,1,1,1,1,1,2,2,1,2,1,
1,1,1,1,1,1,1,1,1,1,1]

[a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z,aa,ab,ac,ad,ae,af,ag,ah,ai,aj,ak,al,am,an,ao,ap,aq,ar]
[1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,2,1,1,2,1,3,1,1,1,1,1, 1, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Salmonella Use Case

- 286 Salmonella Lab Results from NCD
- 174 Positive / 112 Negative
- Using Information gain, selected top 4, 10, 9, and 38 tokens for analysis
- These tokens were used as inputs into Random Forest and Logistic Regression models

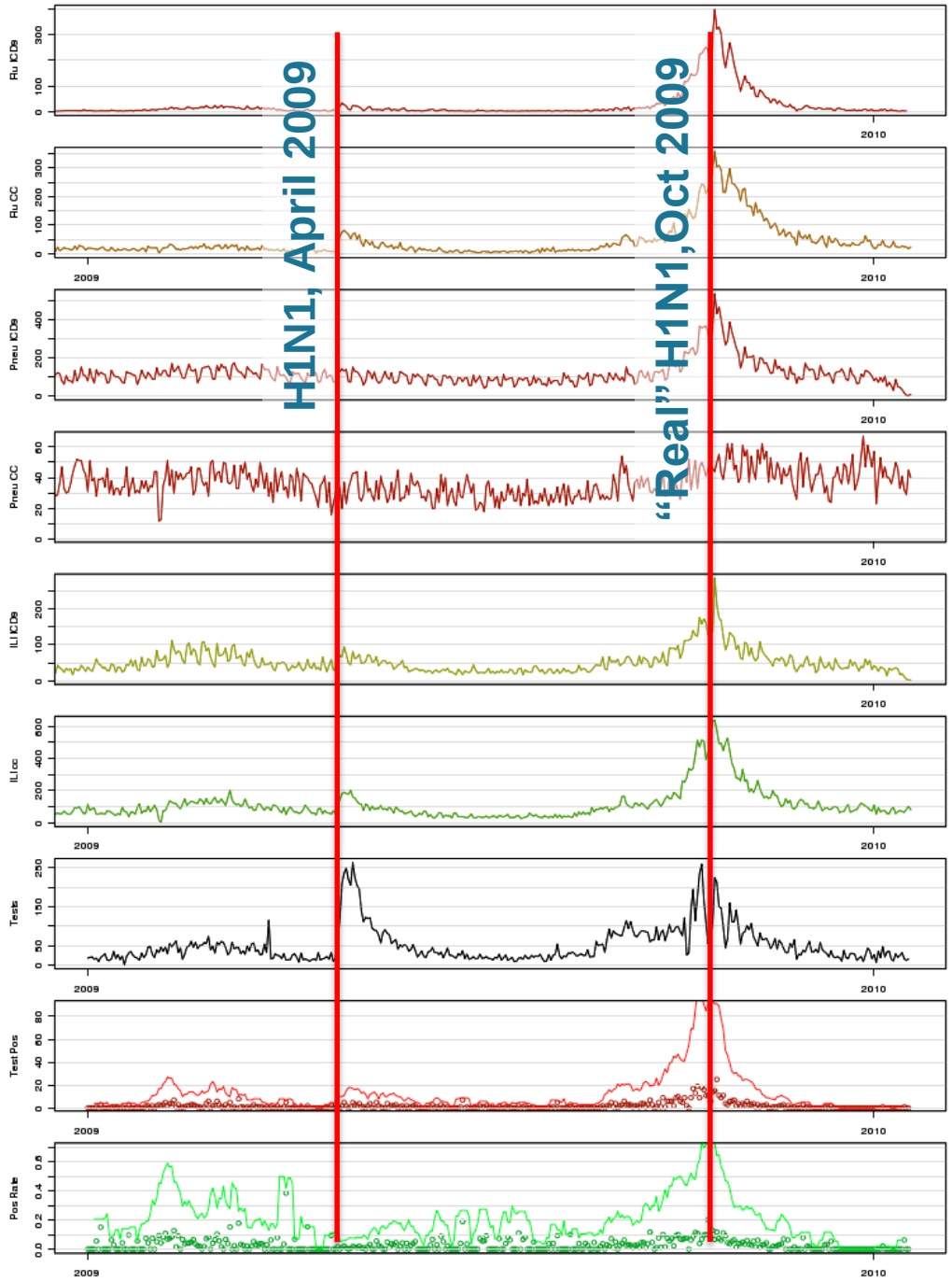
Salmonella Results[†]

Table 1: Salmonella detection accuracy stratified by decision model and the number of tokens used.

Model accuracy	**Classifier:** *Logistic*			**Classifier:** *Random forest*		
Tokens used	Sensitivity	specificity	ROC	Sensitivity	specificity	ROC
First 4	0.951	0.926	0.927	0.951	0.926	0.928
First 10	0.962	0.941	0.962	0.965	0.946	0.961
First 19	0.955	0.928	0.979	0.958	0.935	0.984
First 38	0.944	0.913	0.970	0.962	0.941	0.979

[†] Kirbiyik U, Lai PT, Dixon BE, Grannis SJ, Kasthurirathne SN. "Evaluating the Accuracy of Automated Notifiable Condition Detection in Free-Text Electronic Laboratory Report Results Using Contemporary Text Mining and Machine Learning Methods" presented at the AMIA 2015 Fall Symposium, San Francisco, CA November 17, 2015.

H1N1 Population Phenotype



Flu ICD9 HIE

Flu CC PHESS

Pneumonia ICD9 HIE

Pneumonia CC PHESS

ILI ICD9 HIE

ILI CC PHESS

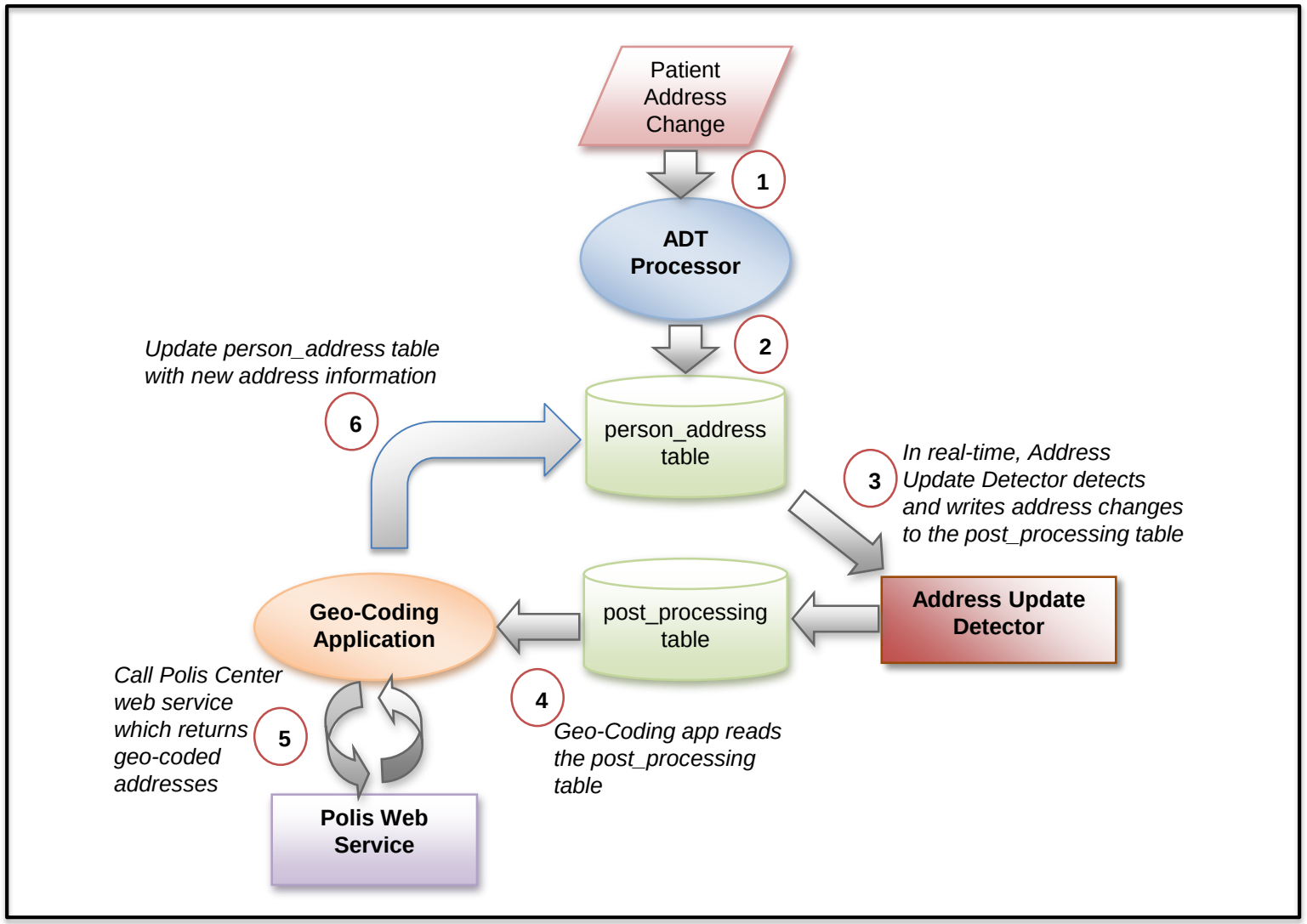
All Flu Tests HIE

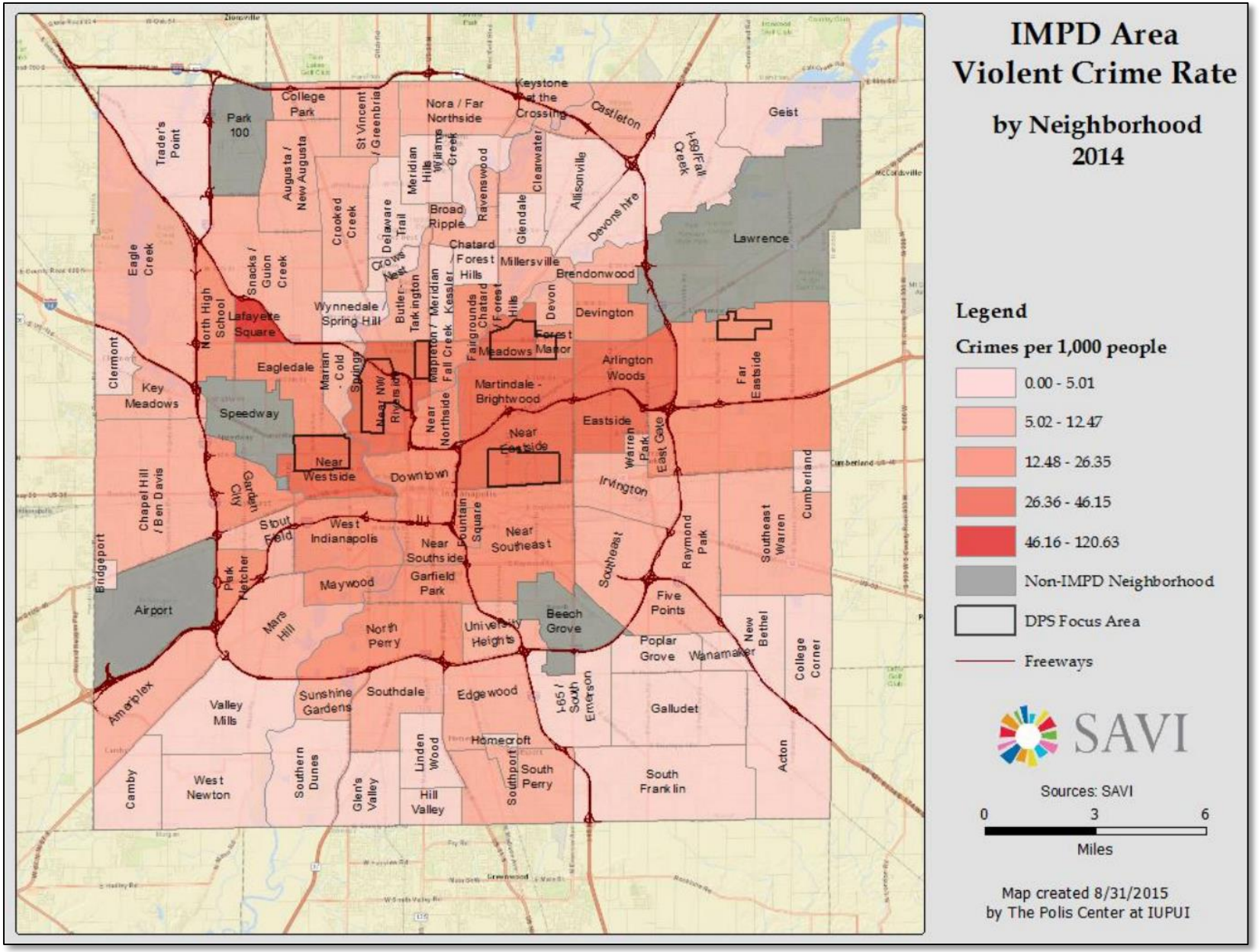
Positive Flu Tests NCD

Positive Rate ALL

Environmental – Linking Clinical Data to the Environment

Adding Geocoding to the INPC

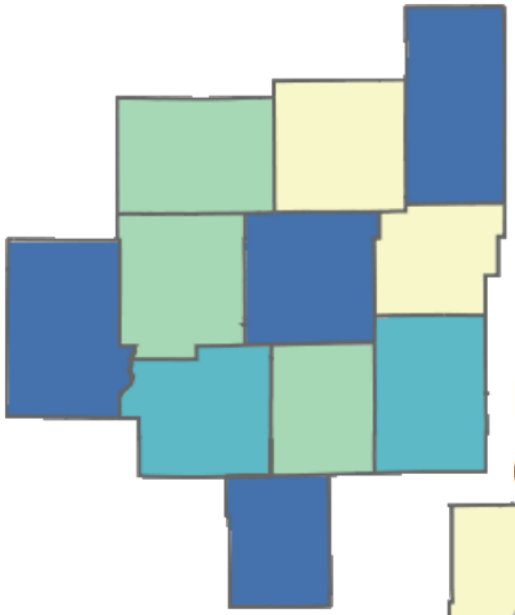




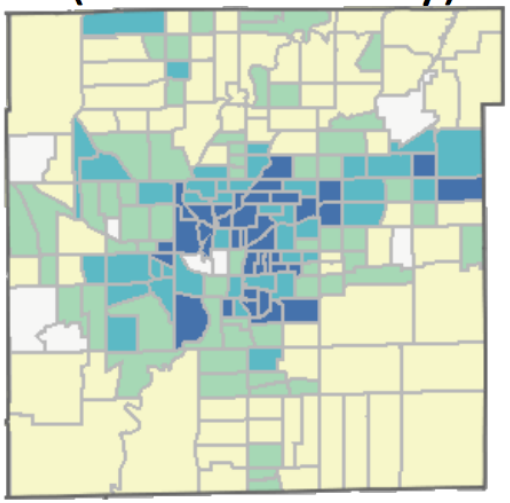
Source: www.savi.org

Families Living in Poverty

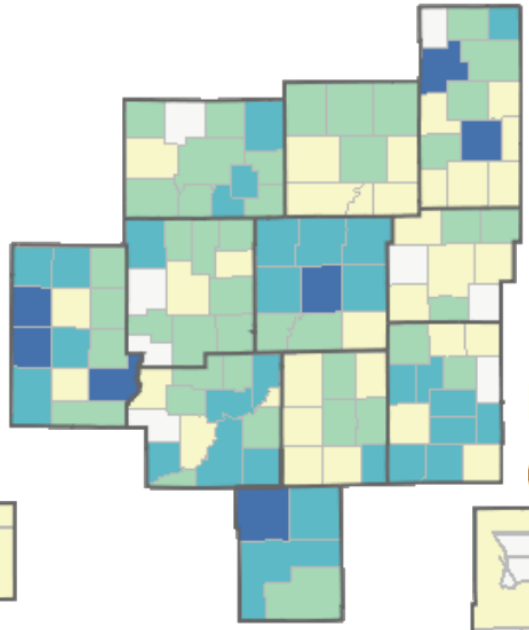
By Counties



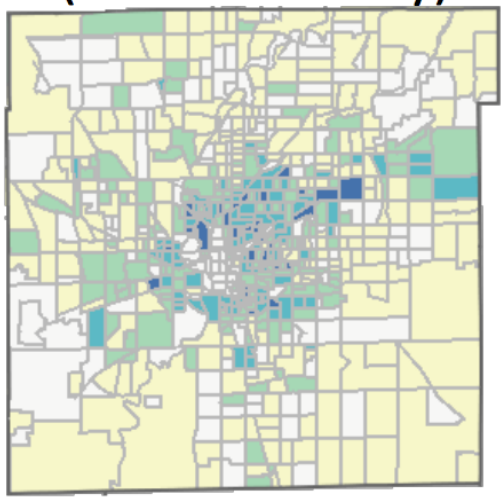
By Census Tracts (Marion County)



By Townships



By Block Groups (Marion County)

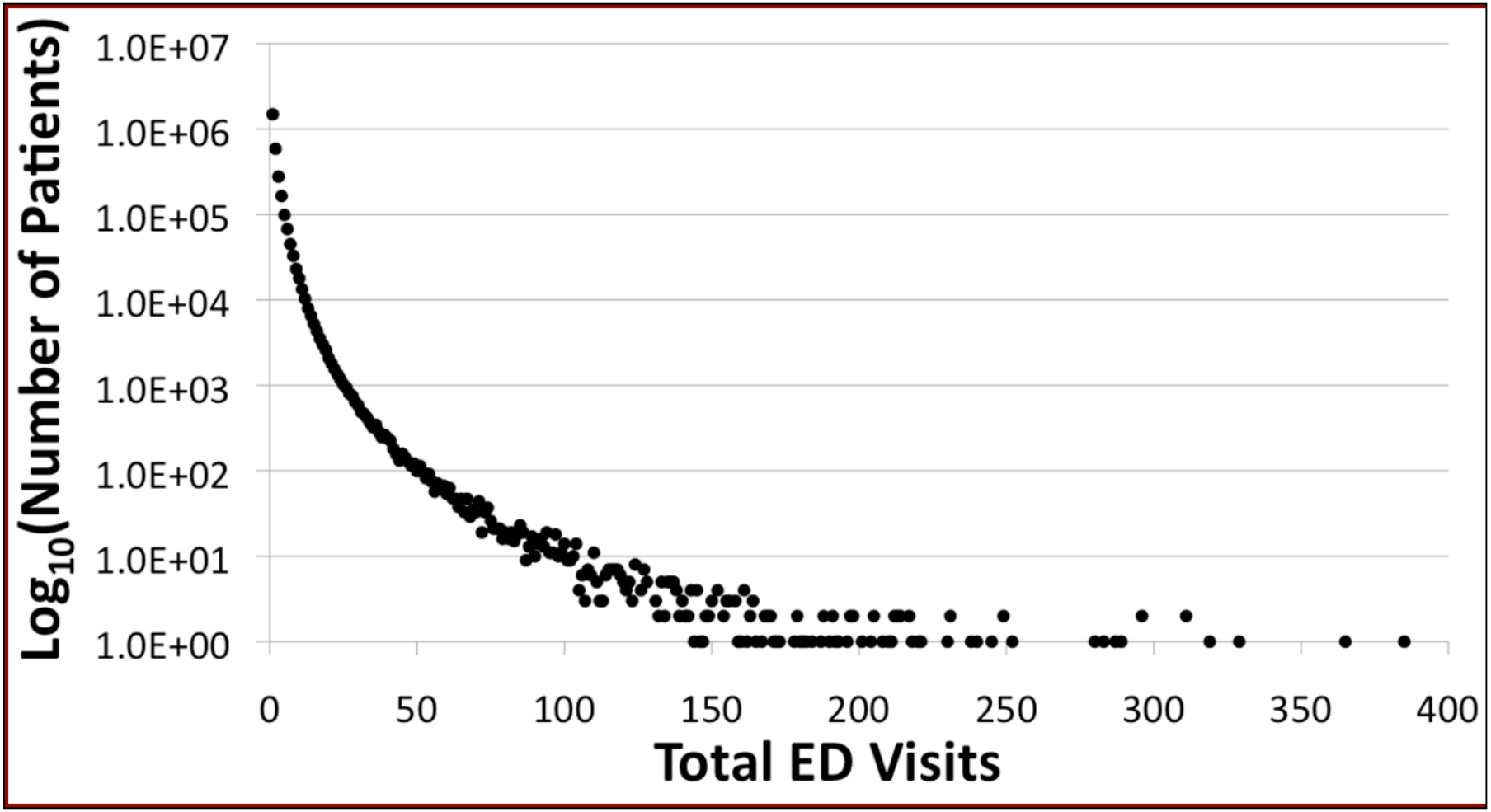


Source: SAVI Community Information System and 2000 Census

Behavioral

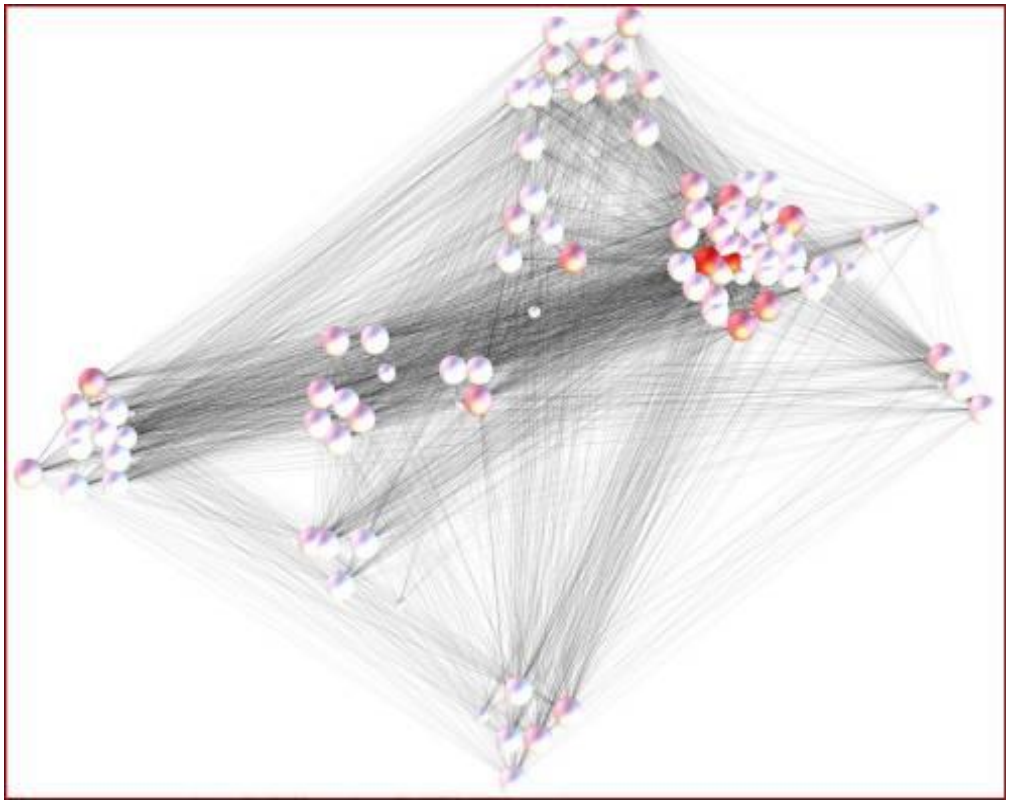
Results

A total of 96 emergency departments contributed data to this analysis.



Distribution of patients stratified by the total number of ED visits. Note that six patients visited the ED more than 300 times and a single patient accumulated 385 visits for the 3-year study period.

Leveraging Analytics to Develop Behavioral Phenotypes



- Patients receive healthcare from multiple providers and across organizations
- More than 40% of ED visits are for patients having data at multiple institutions

A network diagram illustrating the connectedness among Indiana EDs that participate in PHESS. Circular nodes represent EDs; node size indicates the visit volume; node color indicates the centrality of the ED. The gray edges connecting nodes indicate where patient crossover occurs. EDs that share proportionally larger number of patients are clustered together. While general clusters of “medical trading areas” emerge, the myriad gray edges clearly illustrate how interconnected all EDs are to one another.

The “high ED user” phenotype

Model A

$$\begin{aligned}\text{Log (Odds)} = {} & -5.57 + 0.28*\text{Age}(<5) + 1.09*\text{Age}(15\text{-}24) + 0.97*\text{Age}(25\text{-}44) + 0.71*\text{Age}(45\text{-}64) \\ & + 0.50*\text{Age}\ (>=65) + 0.32*\text{Female} + 0.34*\text{Distance}\ (<=5) - 0.24*\text{Distance}\ (>20) \\ & + 0.58*\text{Visits\_2008} - 0.0075*(\text{Visits\_2008})^2 + 0.08*\text{CC\_GI} - 0.10*\text{CC\_Skin} \\ & + 0.42*\text{CC\_RESP} + 0.11*\text{CC\_NEURO} + 0.001*\text{CC\_UDI} - 0.20*\text{CC\_ILI} + 0.08*\text{CC\_Pain} \\ & + 0.44*\text{CC\_Dental} - 0.16*\text{CC\_MUSC} + 0.07*\text{CC\_Lymphatic} + 0.23*\text{CC\_Alcohol} - 0.10*\text{CC\_Unclassified}\end{aligned}$$

Model B

$$\begin{aligned}\text{Log (Odds)} = {} & -7.93 + 0.08*\text{Age}(<5) + 1.77*\text{Age}(15\text{-}24) + 1.78*\text{Age}(25\text{-}44) + 1.47*\text{Age}(45\text{-}64) \\ & + 0.68*\text{Age}(>=65) + 0.27*\text{Female} + 0.33*\text{Distance}(<=5) - 0.069*\text{Distance}\ (>20) \\ & + 0.50*\text{Visits\_2008} - 0.0062*(\text{Visits\_2008})^2 + 0.12*\text{CC\_GI} - 0.42*\text{CC\_Skin} + 0.45*\text{CC\_RESP} \\ & + 0.15*\text{CC\_NEURO} - 0.10*\text{CC\_UDI} - 0.34*\text{CC\_ILI} + 0.21*\text{CC\_Pain} + 0.37*\text{CC\_Dental} \\ & - 0.26*\text{CC\_MUSC} - 0.08*\text{CC\_Lymphatic} + 0.48*\text{CC\_Alcohol} - 0.08*\text{CC\_Unclassified}\end{aligned}$$

Fig. 2

Equations for model predicting frequent emergency department (ED) use as defined as 8 or more visits (a) and model predicting frequent ED use as defined as 16 or more visits in the subsequent two years (b). Distance (<=5): the straight-line distances between geographic points from patients’ home to hospital less than 5 miles; Distance (>20): the straight-line distances between geographic points from patients’ home to hospital greater than 20 miles; CC: chief complaints; GI: gastrointestinal; RESP: respiratory; NEURO: neurological; UDI: undifferentiated infection; ILI: influenza-like illness and MUSC: musculoskeletal

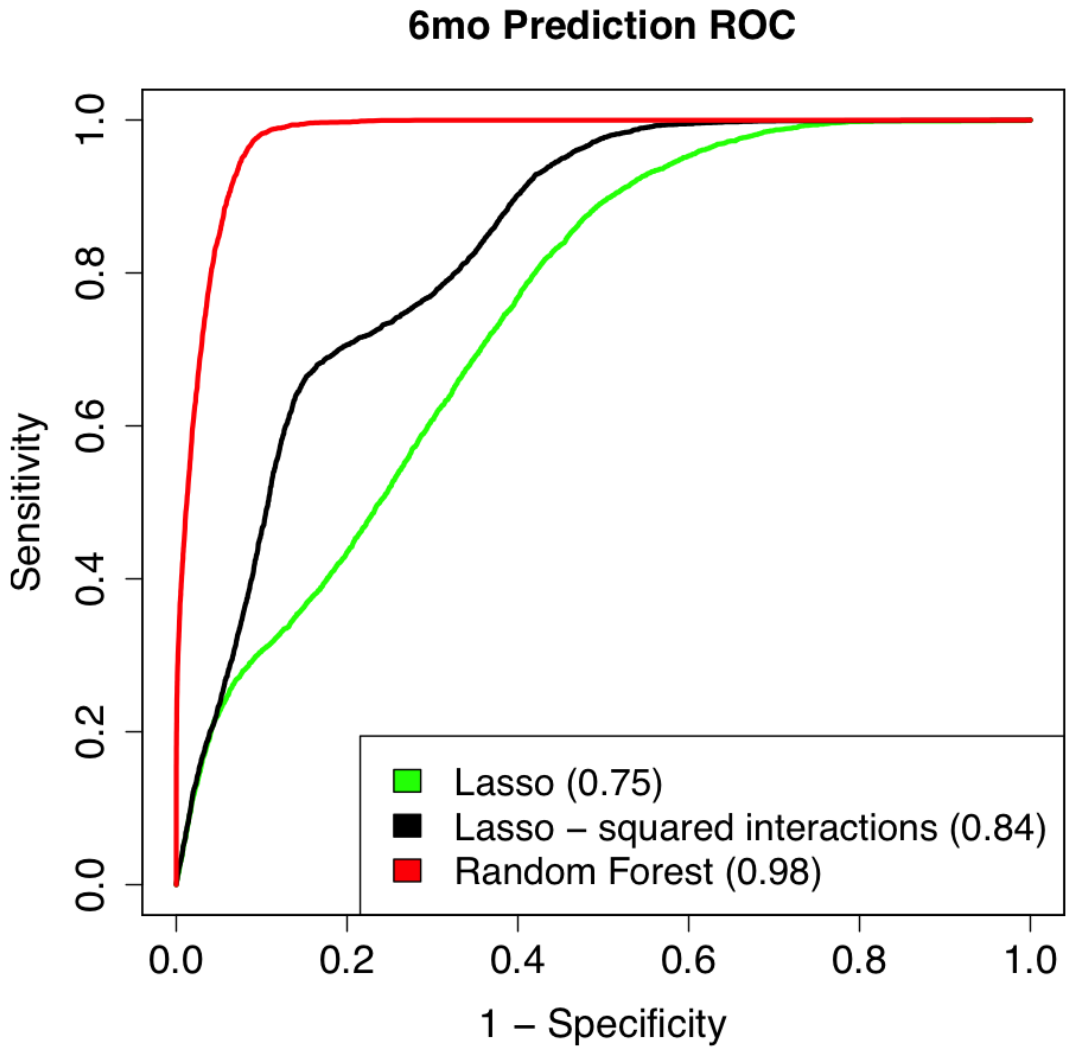
Source: Wu J, Grannis SJ, Xu H, Finnell JT. A practical method for predicting frequent use of emergency department care using routinely available electronic registration data. BMC Emerg Med. 2016 Feb 9;16:12. PubMed PMID: 26860825.

The Results

		Multivariable Logistic Regression Models								
Model No.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
No. of visits constituting 'frequent use'		> = 8	> = 9	> = 10	> = 11	> = 12	> = 13	> = 14	> = 15	> = 16
Area under ROC curve (AUC)		0.84	0.85	0.87	0.88	0.89	0.89	0.90	0.91	0.92
With sensitivity < =25 %, probability > 0.5										
PPV (%)		64.5	63.9	63.4	62.9	62.4	61.3	60.8	60.6	59.5
Specificity (%)		99.5	99.6	99.7	99.7	99.8	99.8	99.8	99.8	99.9
False positive patients	Total No.	5883	4923	4125	3447	2974	2610	2273	2007	1805
	> = 8 visits (No.)	0	565	843	1021	1071	1103	1077	1038	998
Adjusted PPV for patients with > =8 visits in subsequent two years (%)		64.5	68	70.9	73.9	75.9	77.7	79.4	81	81.9

Source: *Wu J, Grannis SJ, Xu H, Finnell JT. A practical method for predicting frequent use of emergency department care using routinely available electronic registration data. BMC Emerg Med. 2016 Feb 9;16:12. PubMed PMID: 26860825.*

Predicting ED High Utilizers



Putting it All Together

- Leveraging the academic, clinical, informatics and HIE infrastructure, our community is well-positioned to link the biologic (clinical/genetic), the environmental, and the behavioral.

Putting it All Together

Leveraging the academic, clinical, informatics and HIE infrastructure, our community is well-positioned to link the biologic, the environmental and the behavioral in support of Precision Medicine.