

# Automated Phenotyping of Patient EMR Data: Feature Extraction and Selection

Rolando Acosta Nuñez, William Artman, Cassandra Burdziak

Dr. Tianxi Cai Laboratory  
Department of Biostatistics  
July 23<sup>rd</sup>, 2015

# Genetic Data

GATTACACCGTAAAATACATA  
GACGTAAGAGGGGGAATTCCA  
GATTAACAGGATTTGACAGGA  
TCAGGATCAGGATACCAGTAA  
ACGGATACCTACGATCAAGTT

# Computational Methods

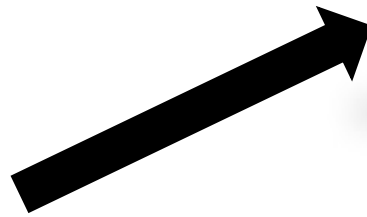


# Personalized Medicine

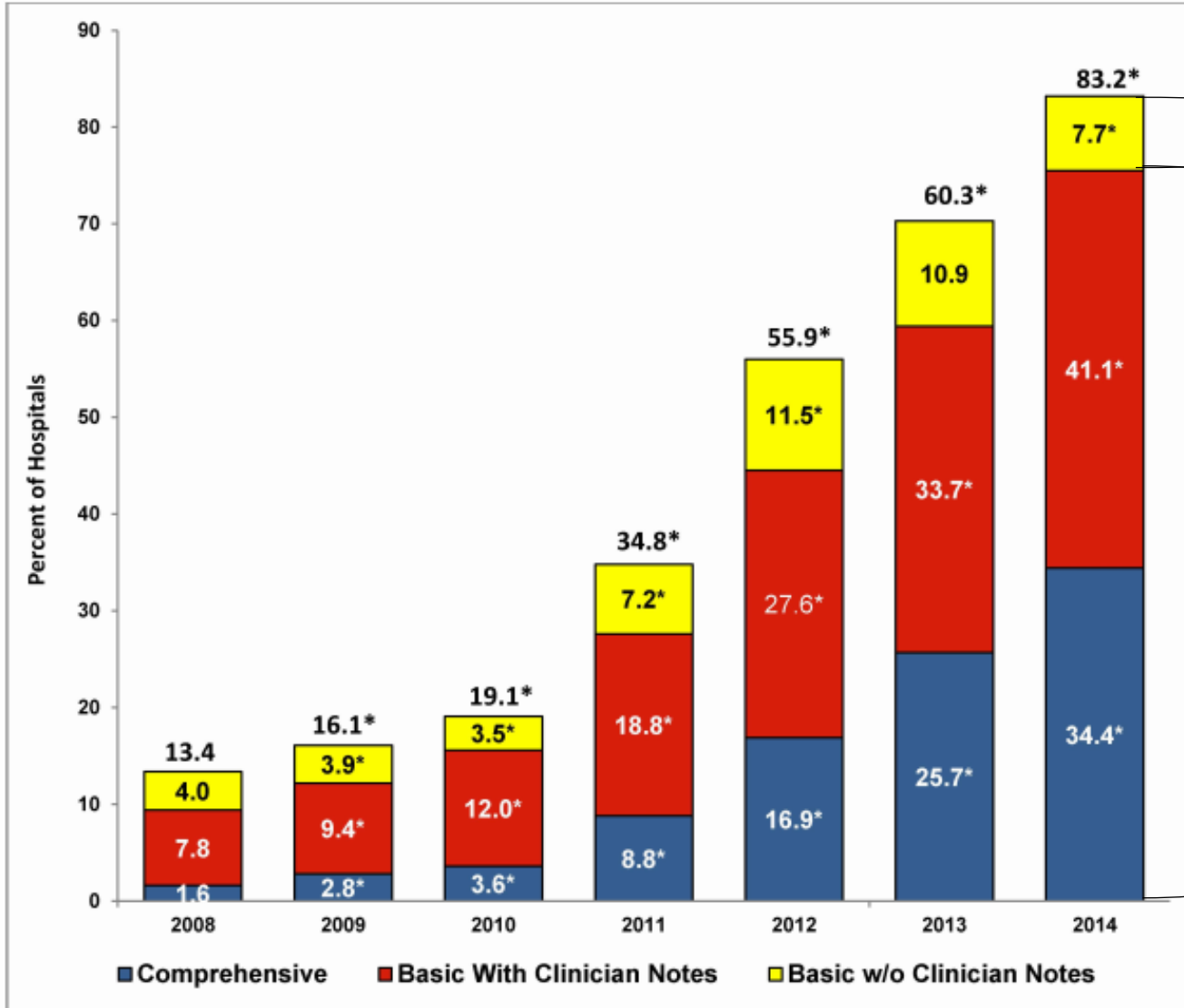


Gene-Disease  
Association

# Phenotype Data



# Percent of non-federal acute care hospitals with adoption of EHR Systems by level of functionality: 2008-2014



## Structured Data

"ICD-9: 714  
 BMI: 21.5  
 Medication: Methotrexate, 7.5mg  
 PO as a single weekly dose"

## Unstructured Data

"...bilateral joint pain and  
 stiffness. Physical exam  
 revealed nodules, swelling,  
 limited range of motion. X-ray  
 showed soft-tissue swelling and  
 early erosions. Blood tests  
 showed presence of rheumatoid  
 factor and elevated SED rate  
 ... "

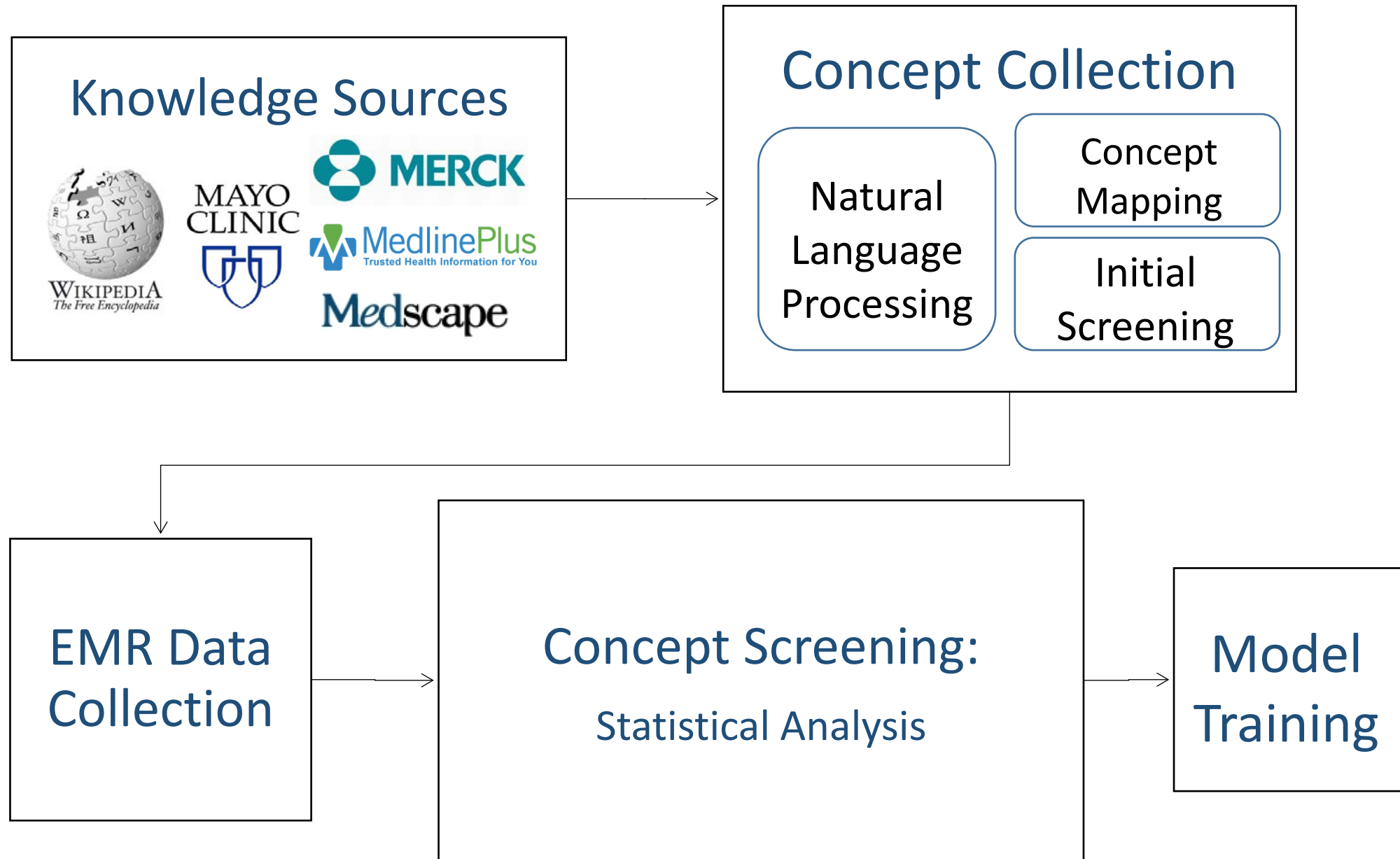
NOTES: Definitions of Basic EHR and Comprehensive EHR systems are reported in Table A1.

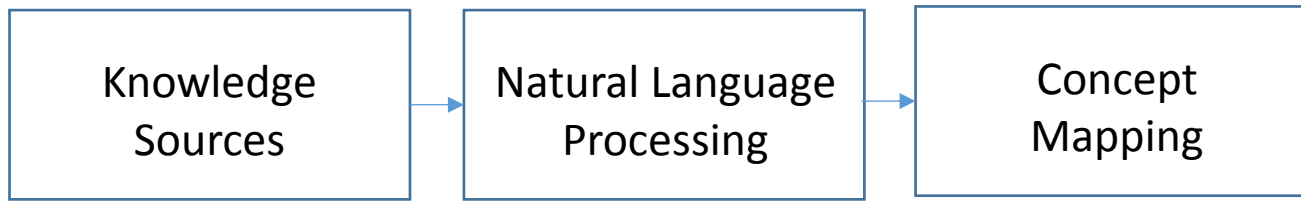
\*Significantly different from previous year ( $p < 0.05$ ).

SOURCE: ONC/AHA, AHA Annual Survey Information Technology Supplement.



develop disease classification model  
 for patient medical record notes

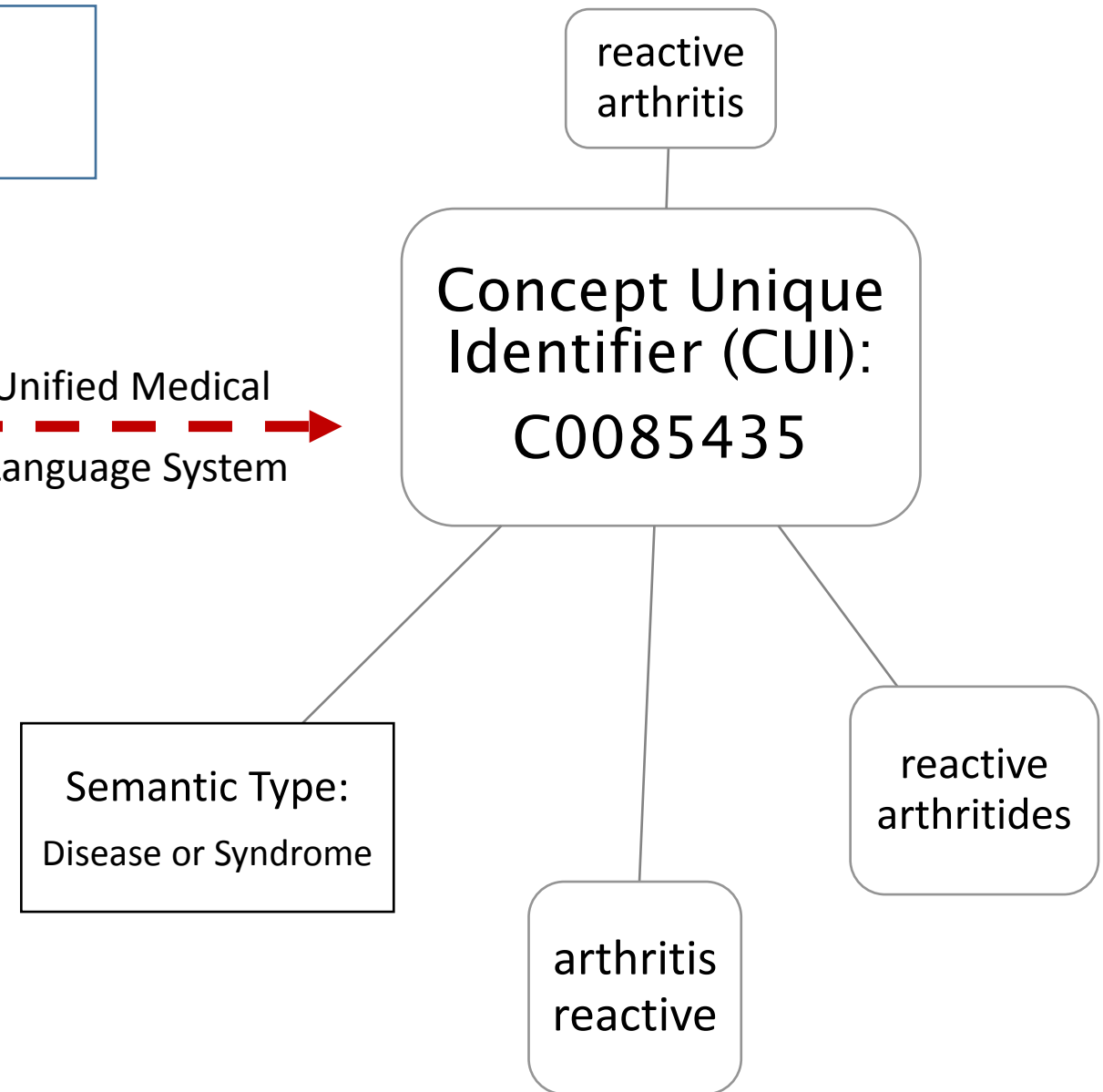


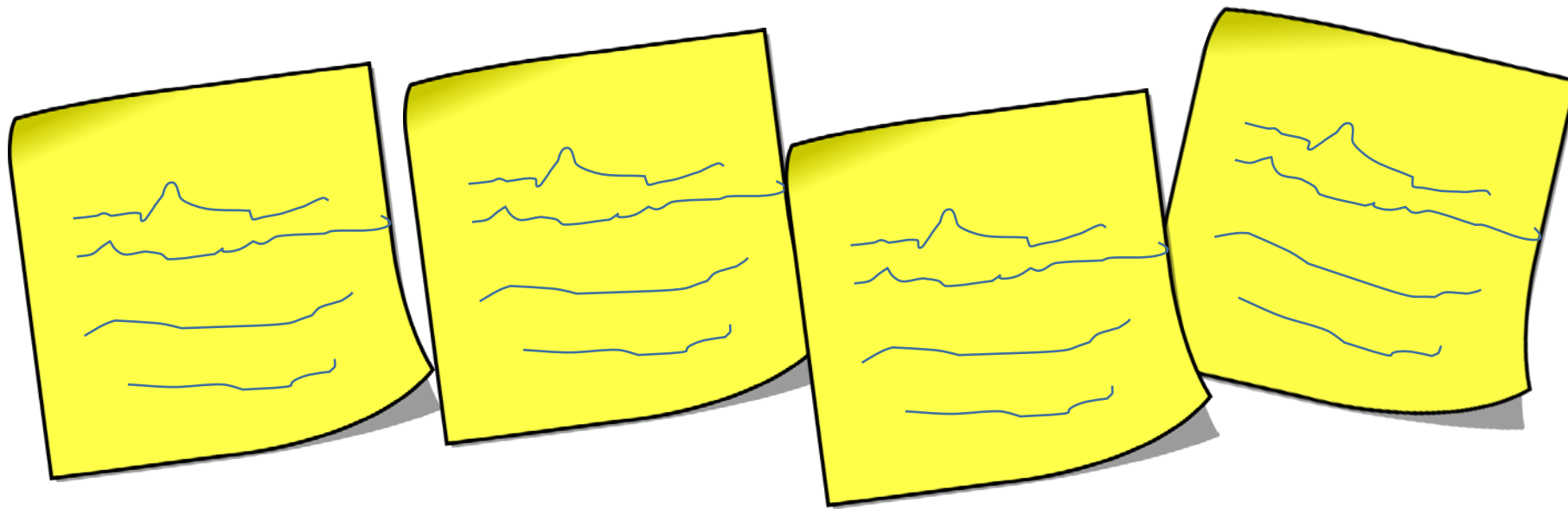


**Septic arthritis**, also known as infectious arthritis, may represent a direct invasion of joint space by various microorganisms, most commonly caused by a variety of bacteria. However, viruses, mycobacteria, and fungi have been implicated. Reactive arthritis is a sterile inflammatory process that usually results from an extra-articular infectious process. Bacteria are the most significant pathogens because of their rapidly destructive nature. For this reason, the current discussion concentrates on the bacterial septic arthritides. Failure to recognize and to appropriately treat septic arthritis results in significant rates of morbidity and may even lead to death ...

... Streptococcal species, such as Streptococcus viridans, S pneumoniae, and group B streptococci, account for 20% of cases. Aerobic gram-negative rods are involved in 20-25% of cases. Most of these infections occur in people who are very young, who are very old, who are diabetic, who are immunosuppressed, and who abuse intravenous drugs.

Unified Medical  
Language System →

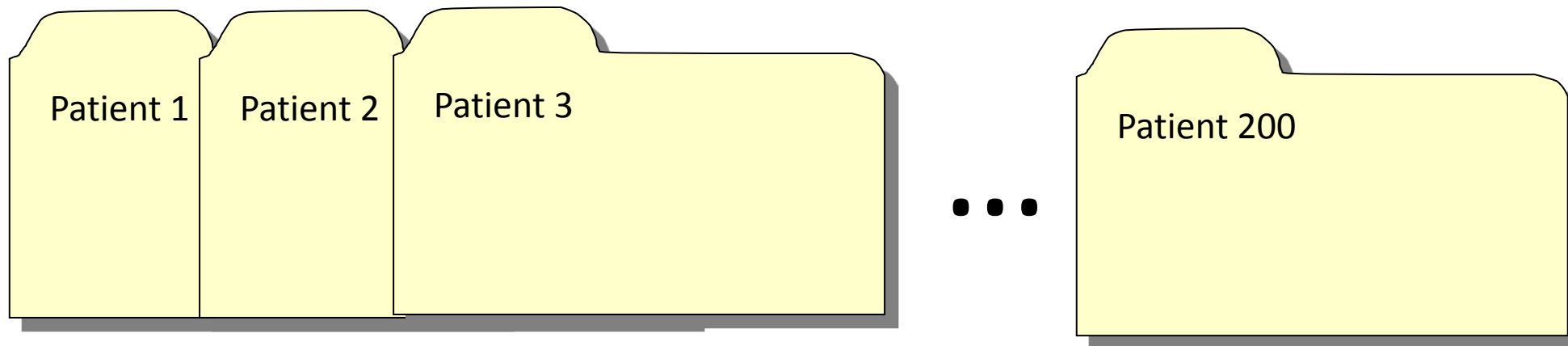




Note-Level  
CUI Screening:  
**non-main CUI**  
mentioned in >5%  
of notes including  
**main CUI**



Patient-Level  
CUI Screening:  
Spearman's Rank Correlation,  $r > 0$



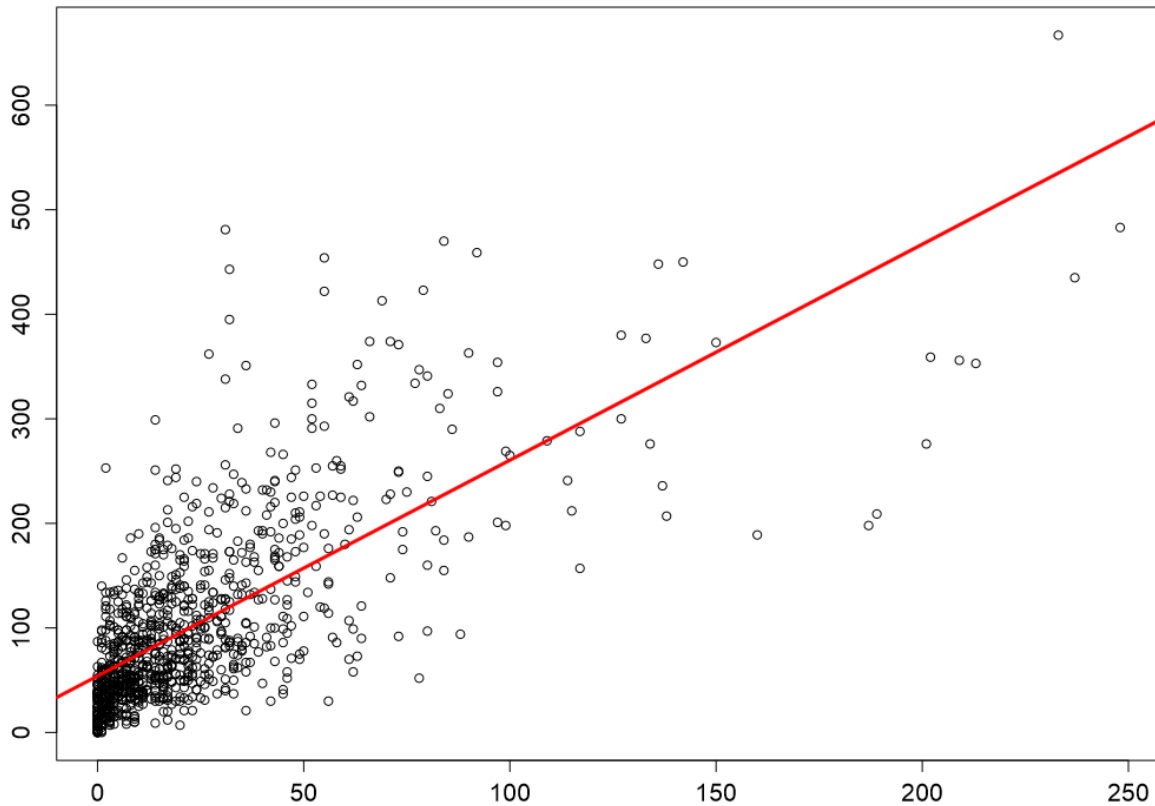
CUI	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Main CUI	92	16	44	368	144
C1247884	84	20	59	320	152
C0934556	122	89	153	0	167



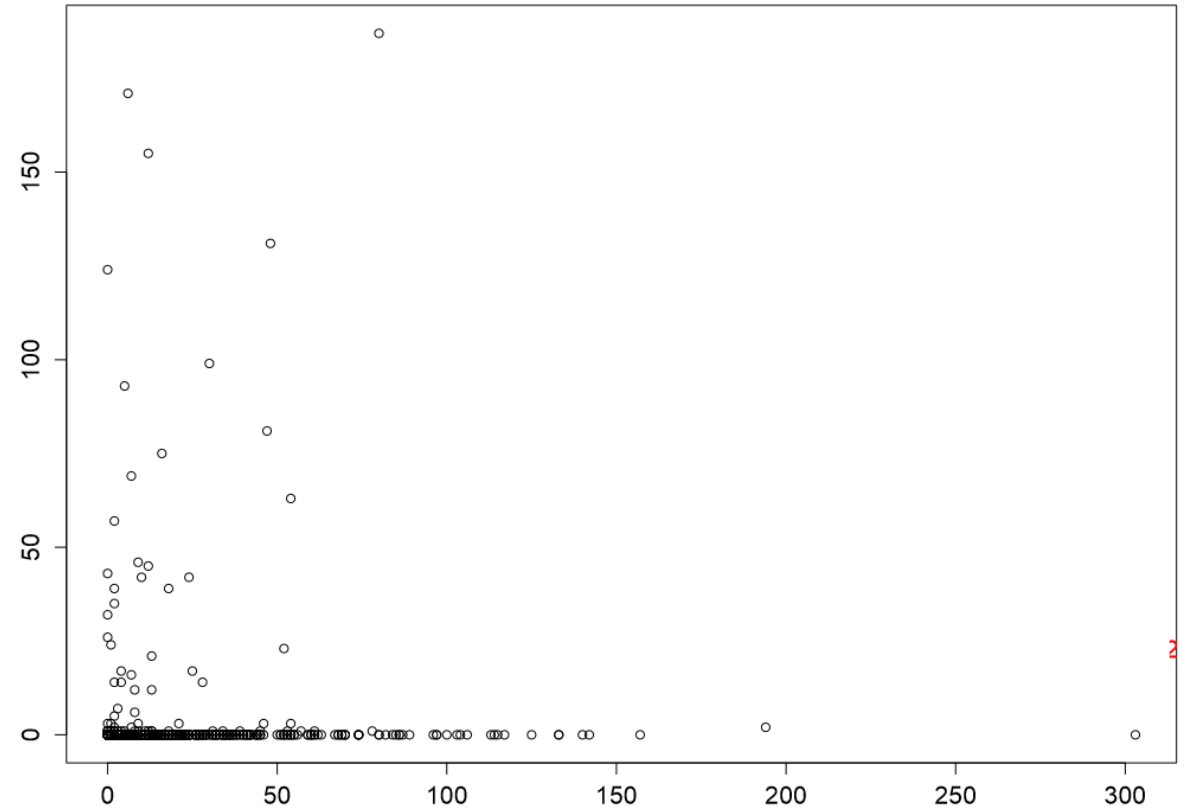
$r = 1, p = 0.017$

$r = -0.10, p = 0.95$

**Tenderness vs Migraine**



**Ibuprofen vs Giant Cell Arteritis**



# Multiple Testing Problem

$$\alpha = 0.05$$

Disease	Number of Tests
Obesity	129
Migraine	71
Septic Arthritis	100
Giant Cell Arteritis	93
Osteoarthritis	66

Multiple Testing  
→  
Correction  $\alpha_i'$



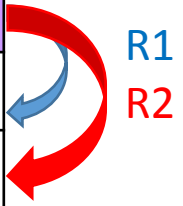
# Bonferroni Correction

$$\alpha' = \frac{0.05}{\# \text{ of tests}}$$

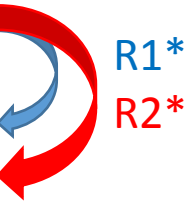
# Benjamini-Hochberg

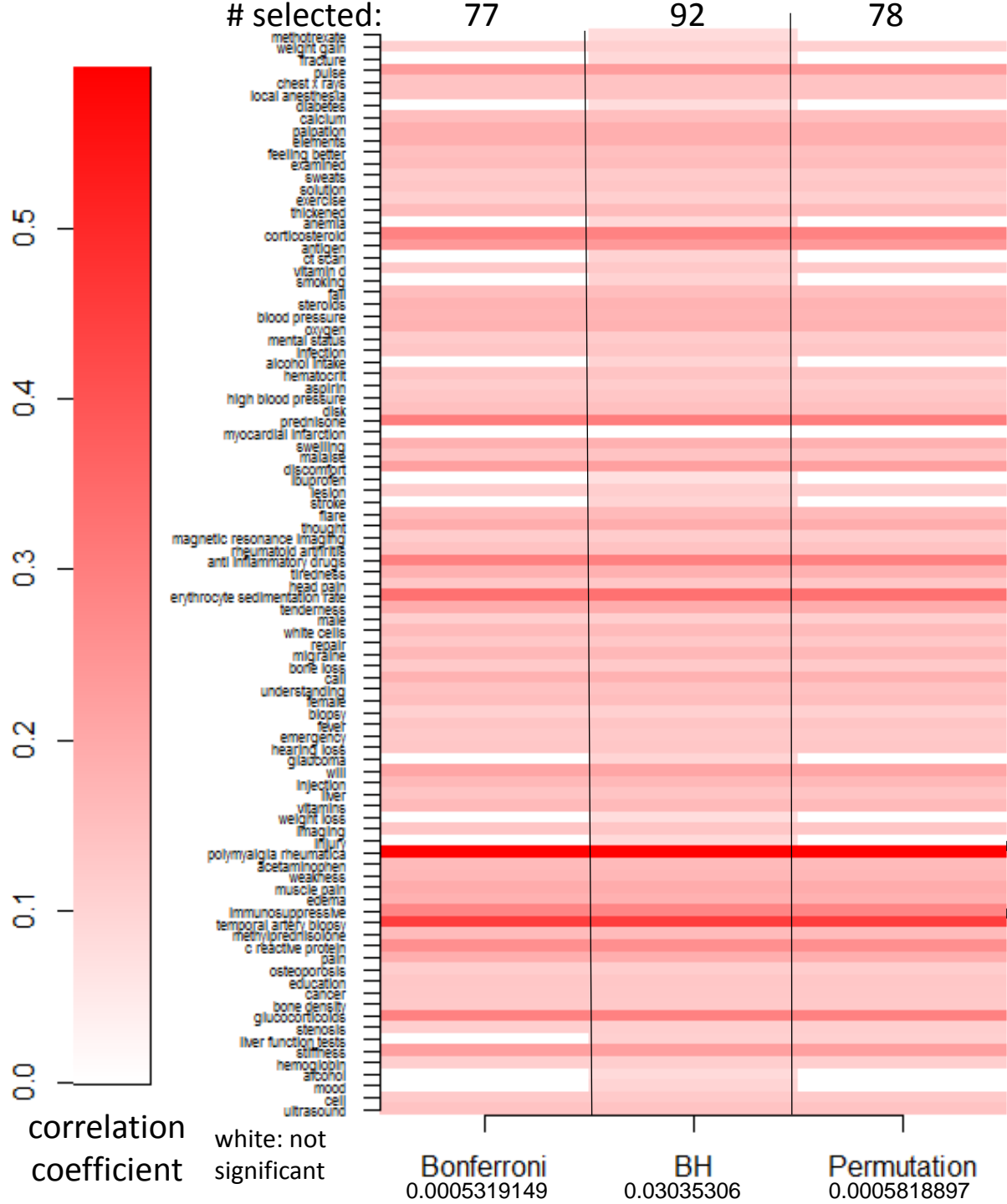
# Permutation Test

CUI	P1	P2	P3	P4	P5
Main CUI	92	16	44	368	144
C1247884	84	20	59	320	152
C0934556	122	89	153	0	167



CUI	P1	P2	P3	P4	P5
Main CUI	16	368	92	144	44
C1247884	84	20	59	320	152
C0934556	122	89	153	0	167





polymyalgia rheumatica  
 blood pressure corticosteroid  
 anti inflammatory drugs  
 thought glucocorticoids  
 injection will call oxygen weakness  
 pain palpation pulse tiredness  
 immunosuppressive  
 erythrocyte sedimentation rate  
 swelling steroids stiffness elements  
 migraine discomfort antigen  
 muscle pain edema tenderness  
 temporal artery biopsy  
 prednisone  
 c reactive protein

polymyalgia rheumatica  $r = 0.59$   
 temporal artery biopsy  $r = 0.45$

# Giant Cell Arteritis

rheumatoid arthritis  
infection  
high blood pressure  
diabetes  
stroke  
anemia  
myocardial infarction  
*Disease or Syndrome*

blood pressure  
weight gain  
swelling  
lesion  
fall  
sweats  
examined  
thickened  
mental status  
feeling better  
*Finding*  
discomfort  
head pain  
flare  
malaise  
tenderness  
tiredness  
*Sign or Symptom*

corticosteroid  
methotrexate  
ibuprofen  
vitamin d  
aspirin

anti inflammatory drugs  
prednisone  
*Pharmacologic Substance*

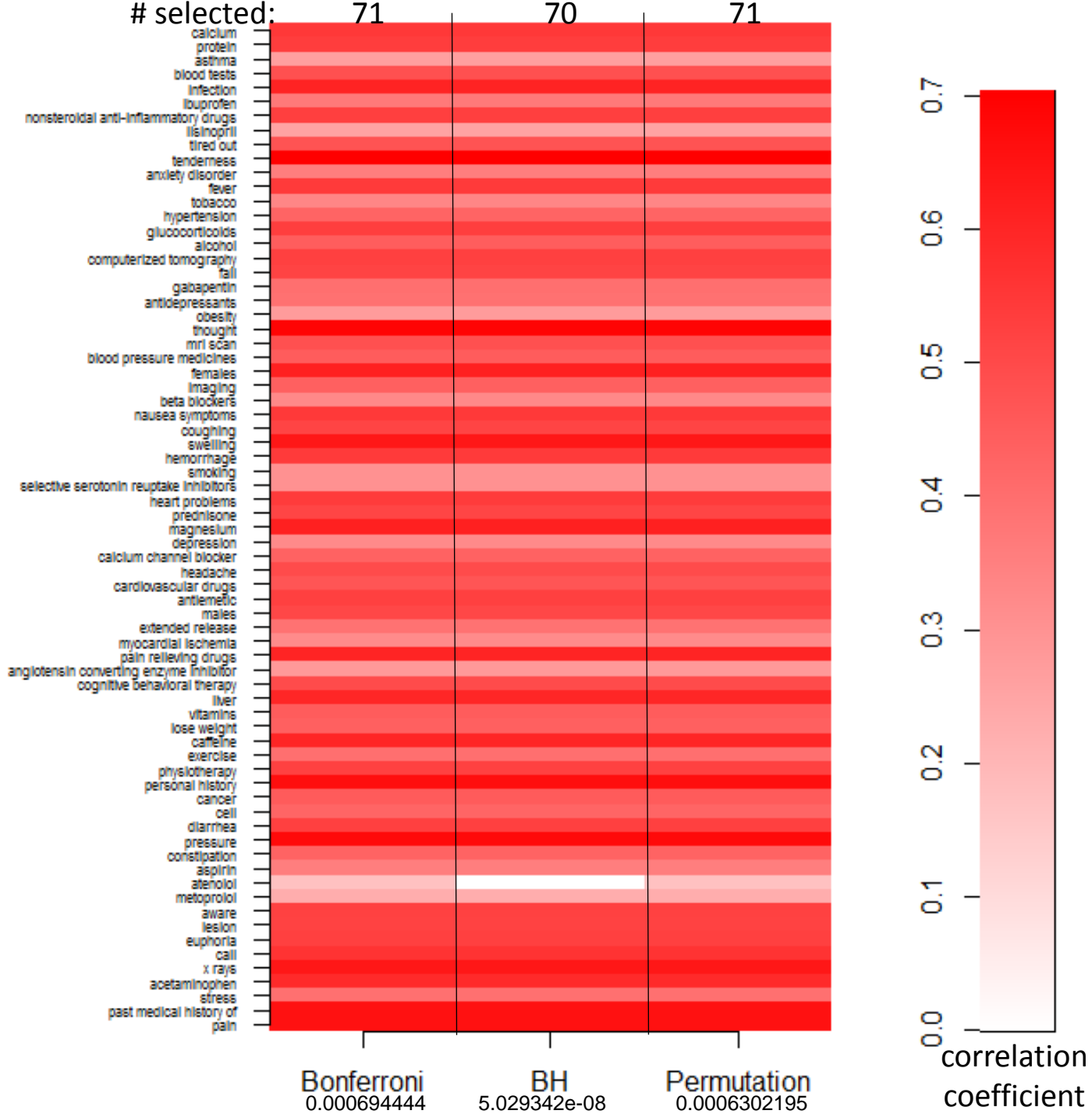
chest x rays  
ct scan  
palpation  
magnetic resonance imaging  
*Diagnostic Procedure*

# Giant Cell Arteritis

Highly Correlated Terms by Type

# Migraine

pain relieving drugs  
 personal history  
 computerized tomography acetaminophen  
 pressure tenderness  
 thought  
 hemorrhage  
 physiotherapy aware liver antileptic calcium  
 infection caffeine fever females  
 protein diarrhea  
 nonsteroidal anti-inflammatory drugs  
 call x rays euphoria heart problems  
 swelling magnesium pain  
 past medical history of  
 nausea symptoms



# Migraine

## Highly Correlated Terms by Type

computerized tomography  
imaging  
X rays  
mri scan  
**Diagnostic Procedure**

swelling  
males females  
fall  
past medical history of  
lose weight  
fever lesion  
personal history  
**Finding**

antiemetic  
liver prednisone  
caffeine  
calcium channel blocker  
cardiovascular drugs  
acetaminophen  
blood pressure medicines  
pain relieving drugs  
**Pharmacologic Substance**

nausea symptoms  
heart problems  
constipation headache  
coughing  
tired out diarrhea  
pain  
tenderness  
**Sign or Symptom**

adrenergic agonists  
metformin diuretics  
**phenylpropanolamine**  
cocaine  
stimulant caffeine  
dopamine reuptake inhibitors  
catechin  
pharmaceuticals  
**Pharmacologic Substance**

quetelet's index  
**body weight**  
electrocardiogram  
**Diagnostic Procedure**

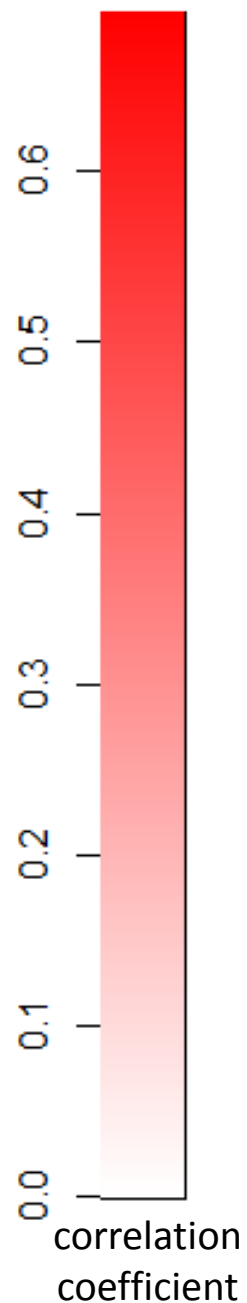
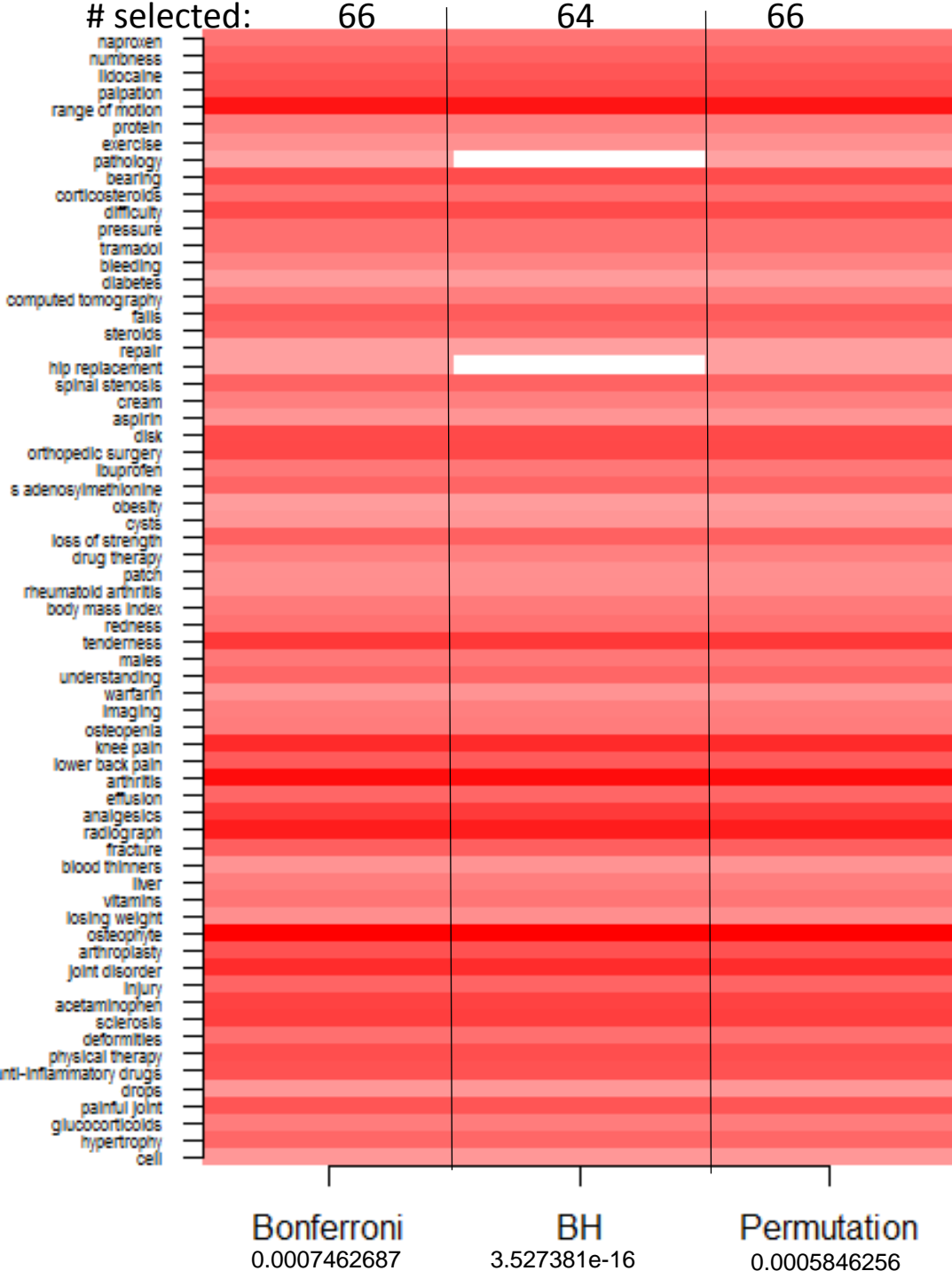
**lost weight**  
historical fast  
male stress grade 1  
**increase weight**  
education  
health history  
**Finding**

erectile dysfunction  
type 2 diabetes  
chronic obstructive pulmonary disease  
**diabetes**  
congestive heart failure  
**hypertension**  
multiple sclerosis  
**Disease or Syndrome**

glucose  
protein  
sodium  
trace elements  
**hemoglobin a1c**  
**Biologically Active Substance**

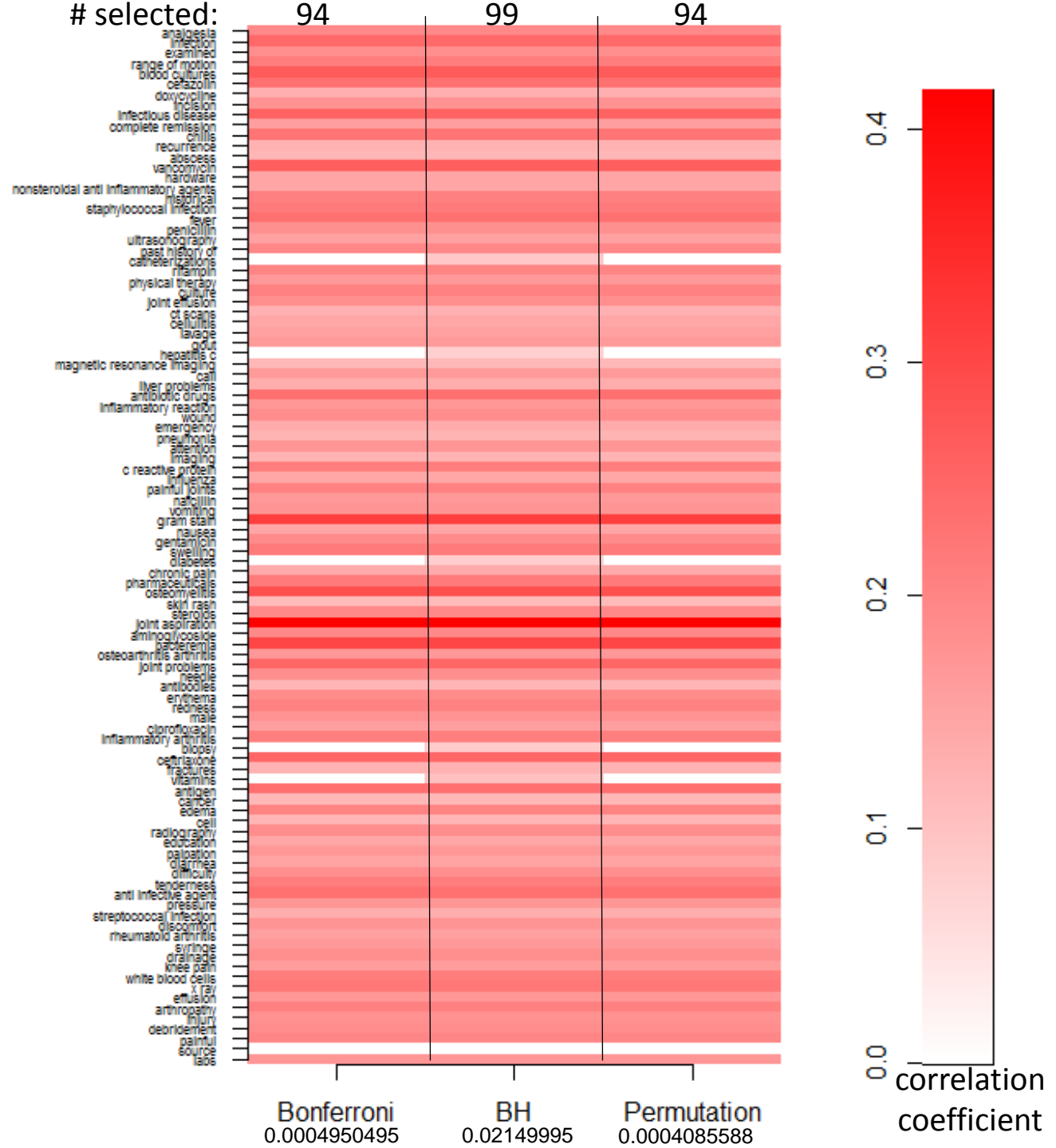
# Highly Correlated Terms by Type

# Obesity



# Osteoarthritis

# Septic Arthritis





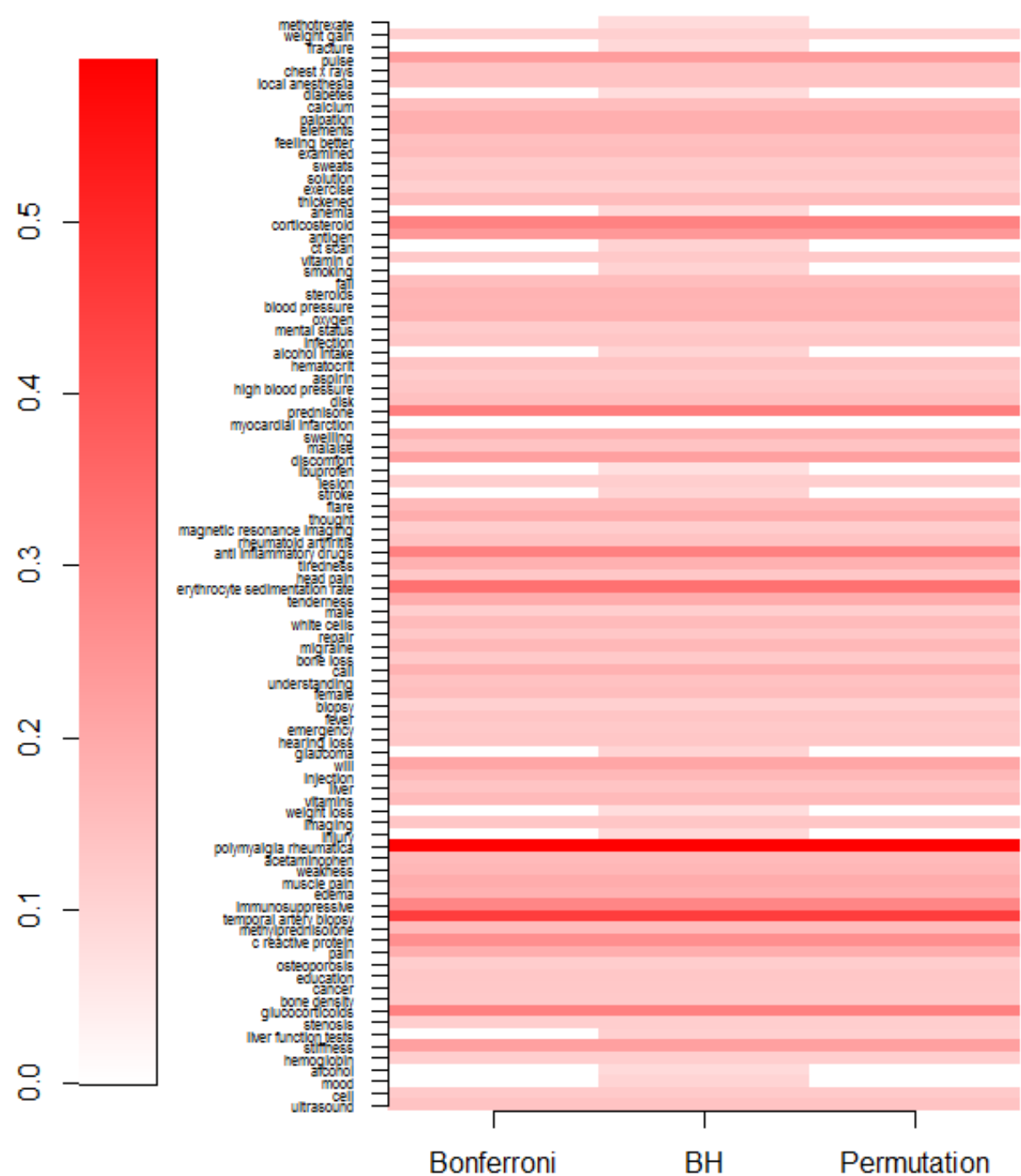
# Discussion

- Limitations

- Relying on quality of sources
- Diseases with multiple names which correspond to different CUIs in the UMLS

- Example: Giant Cell Arteritis and Temporal Arteritis

- Conclusion and Future Directions



# Acknowledgements

## **Cai Lab and Collaborators**

- Dr. Tianxi Cai
- Dr. Sheng Yu

## **Summer Program in Biostatistics and Computational Biology**

- Dr. Rebecca Betensky
- Tonia Smith
- Heather Mattie
- Eleanor Murray
- Joshua Barback