PheKnow-Cloud: A Tool for Evaluating High-Throughput Phenotype Candidates using Online Medical Literature

Jette Henderson[1], Ryan Bridges[2], Joyce C. Ho[3], Byron C. Wallace[4], Joydeep Ghosh[1]

[1]UT Austin, [2]Epic Systems, [3]Emory University, [3]Northeastern University

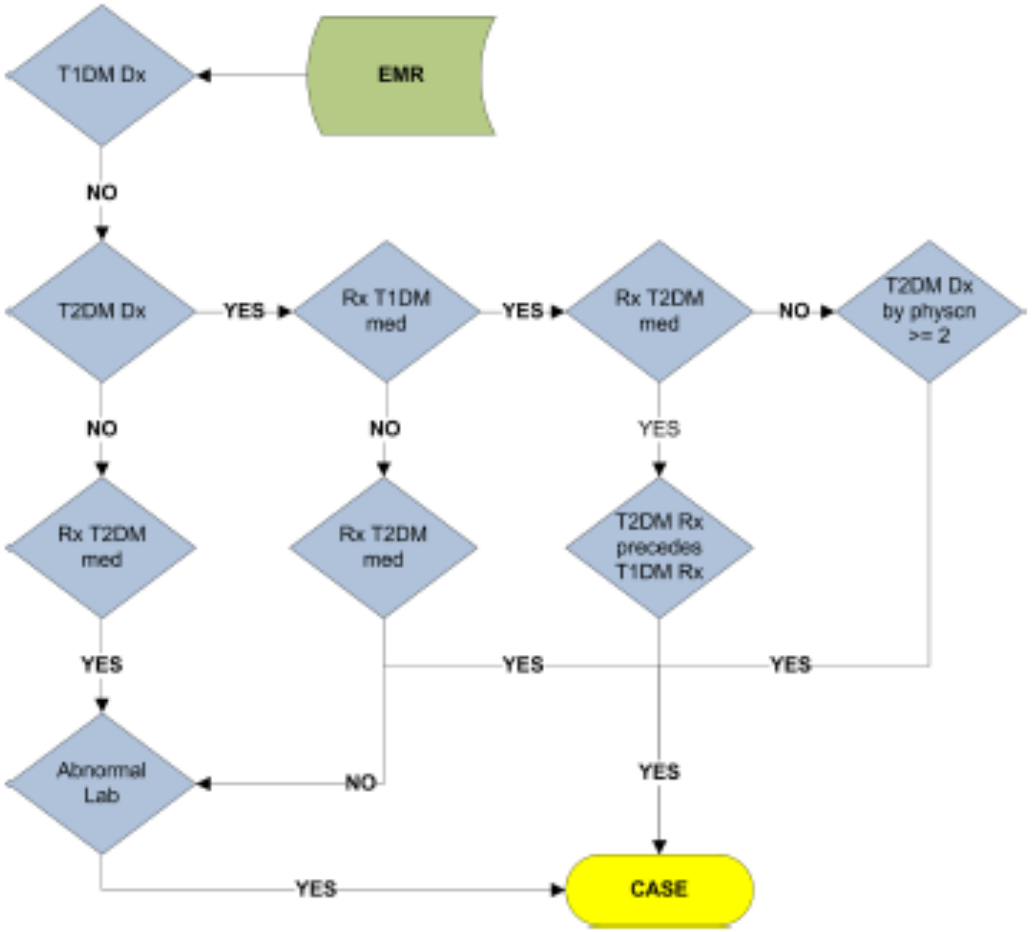
AMIA Joint Summits on Translational Sciences

Disclosure

- Neither my collaborators nor I have no relationships with commercial interests or conflicts of interests

Background: EHR-Based Phenotyping

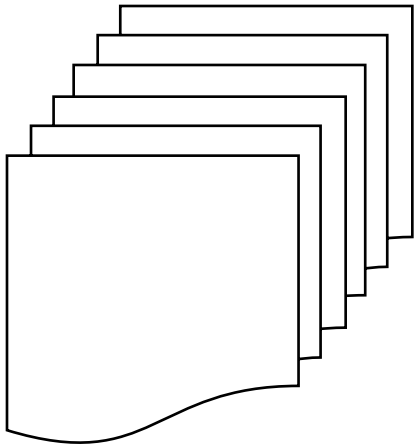
Figure 1: Algorithm for identifying T2DM cases in the EMR.



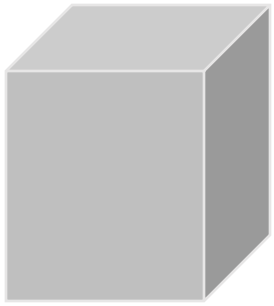
Manual Phenotype Extraction

- Laborious
- Time-consuming
- Requires domain expertise

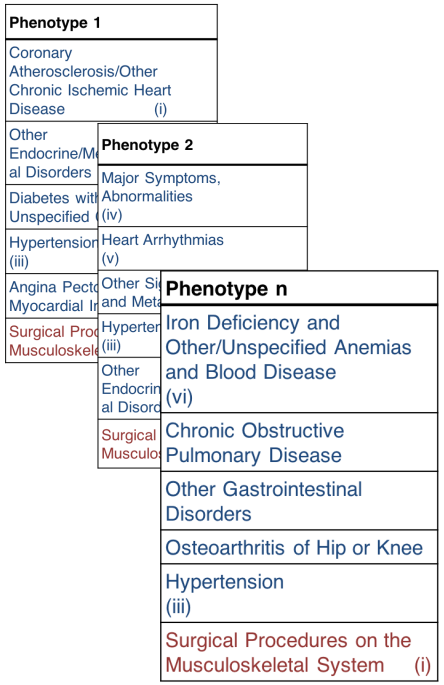
Motivation & Background



Raw Electronic Healthcare Record Data

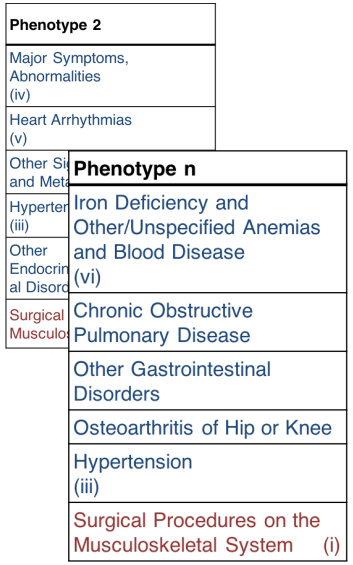


Automatic Phenotype Generation Process



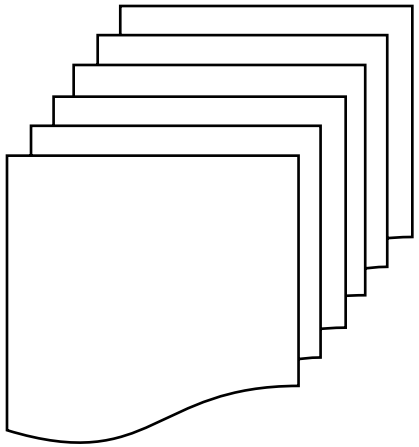
Candidate Phenotypes

Verification by Panel of Experts

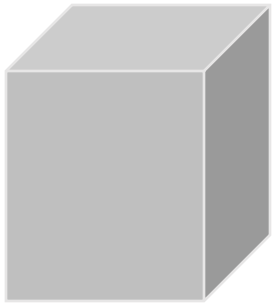


Verified Phenotypes

Goal



Raw Electronic Healthcare Record Data



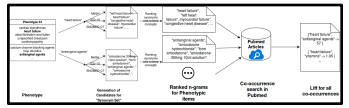
Automatic Phenotype Generation Process

Phenotype 1
Coronary Atherosclerosis/Other Chronic Ischemic Heart Disease (i)
Other Endocrine/M al Disorders
Diabetes wit Unspecified
Hypertensior (iii)
Angina Pecto Myocardial Ir
Surgical Proc Musculoskel

Phenotype 2
Major Symptoms, Abnormalities (iv)
Heart Arrhythmias (v)
Other Si and Meta
Hyperter (iii)
Other Endocrin al Disorc
Surgical Musculos

Phenotype n
Iron Deficiency and Other/Unspecified Anemias and Blood Disease (vi)
Chronic Obstructive Pulmonary Disease
Other Gastrointestinal Disorders
Osteoarthritis of Hip or Knee
Hypertension (iii)
Surgical Procedures on the Musculoskeletal System (i)

Candidate Phenotypes



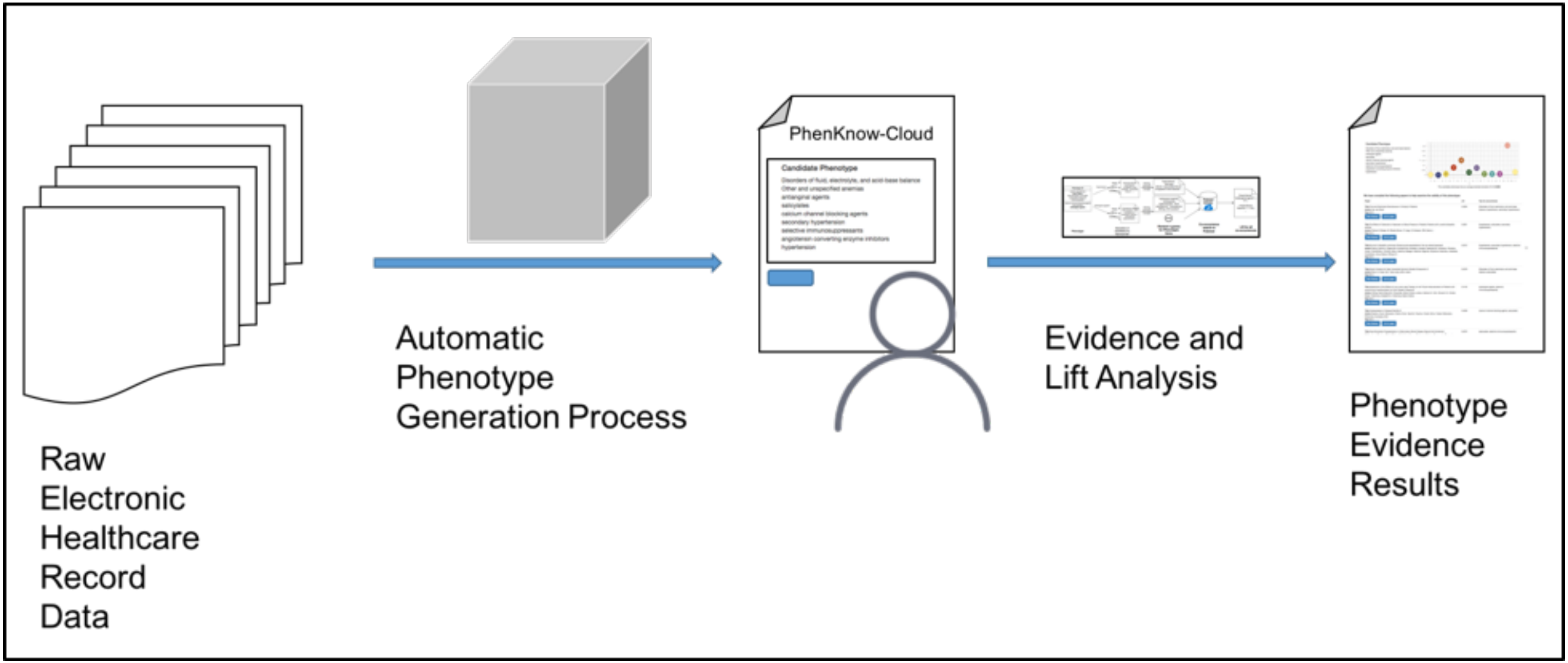
?

Automatic Verification

Phenotype 2
Major Symptoms, Abnormalities (iv)
Heart Arrhythmias (v)
Other Si and Meta
Hyperter (iii)
Other Endocrin al Disorc
Surgical Musculos

Phenotype n
Iron Deficiency and Other/Unspecified Anemias and Blood Disease (vi)
Chronic Obstructive Pulmonary Disease
Other Gastrointestinal Disorders
Osteoarthritis of Hip or Knee
Hypertension (iii)
Surgical Procedures on the Musculoskeletal System (i)

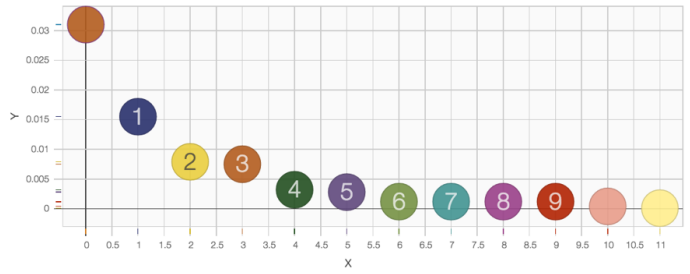
Verified Phenotypes


PhenKnow-Cloud
Candidate Phenotype
Raw
Electronic
Healthcare
Record
Data
Automatic
Phenotype
Generation Process
Evidence and
Lift Analysis
Phenotype
Evidence
Results

PheKnow-Cloud Result Interface

Candidate Phenotype

Disorders of fluid, electrolyte, and acid-base balance
Other and unspecified anemias
antianginal agents
salicylates
calcium channel blocking agents
secondary hypertension
selective immunosuppressants
angiotensin converting enzyme inhibitors
hypertension

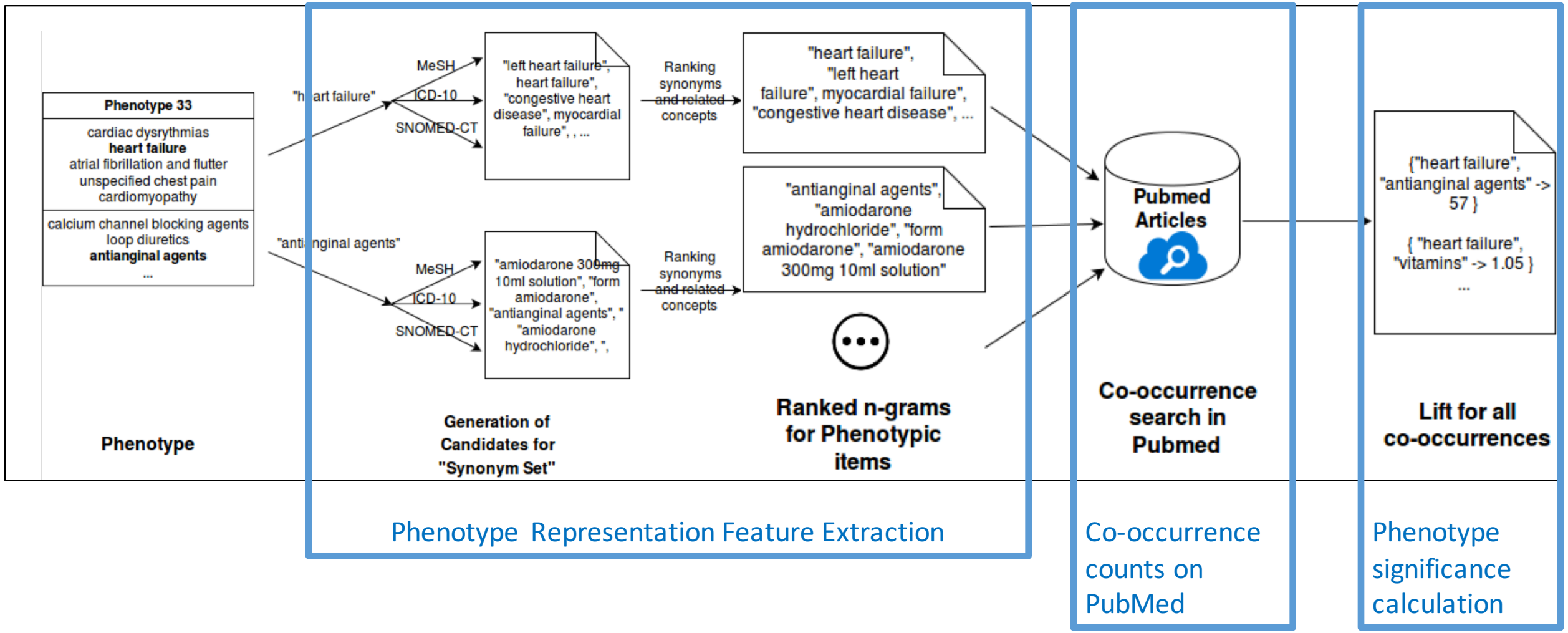


This candidate phenotype has an average standard deviation (above the median) lift of **0.0064**

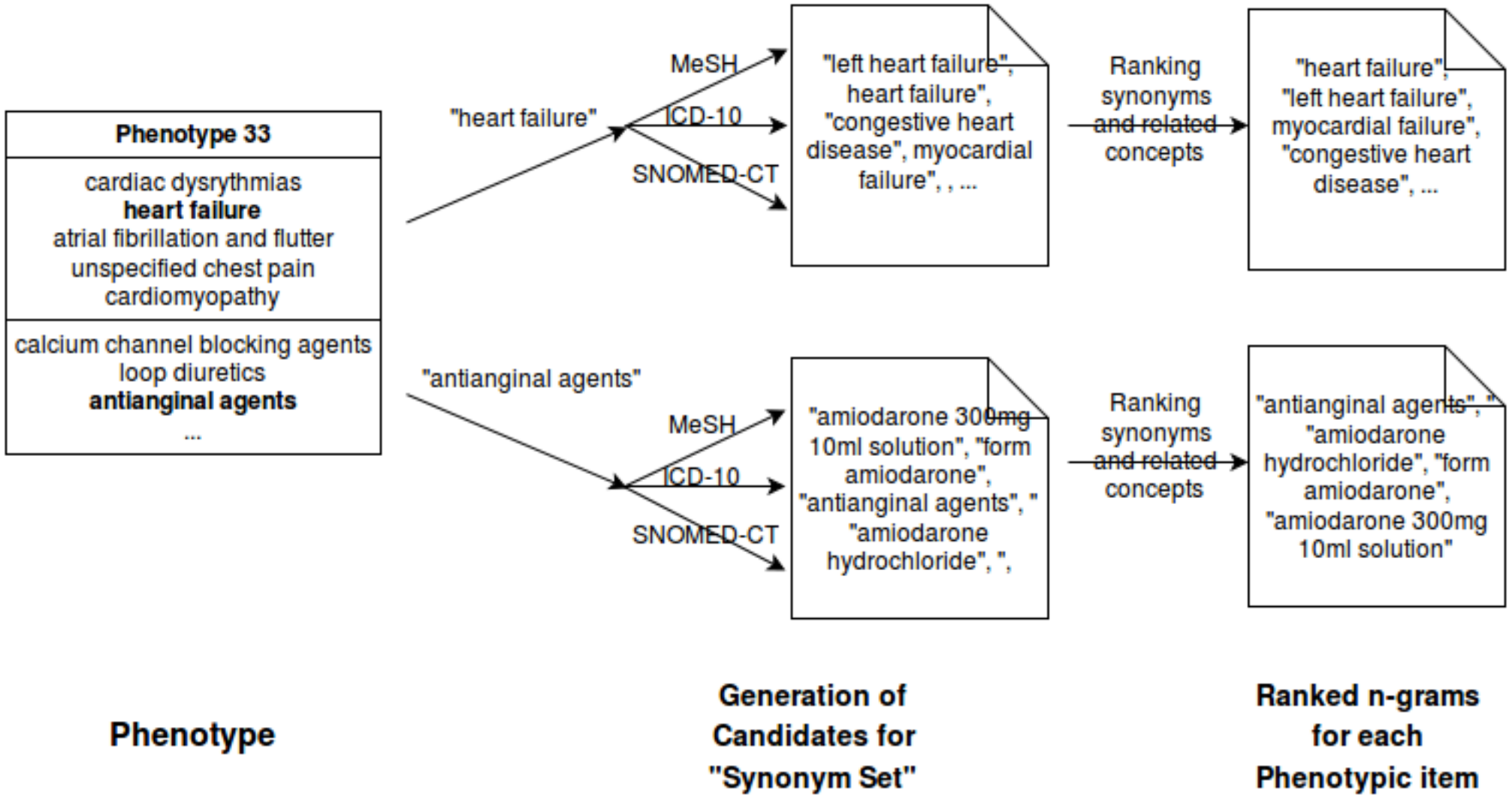
Table of Evidence

Index	Paper	Standard Deviations above Median Lift	Co-occurrence Tuples
0	**Title**:Unmet medical needs in lupus nephritis: solutions through evidence-based, personalized medicine **Author**:Anders, Hans-Joachim; Weidenbusch, Marc; Rovin, Brad **Year**:2015 View Abstract Link to paper	0.031	(calcium channel blocking agents, selective immunosuppressants)
1	**Title**:Assessment of the Effects of Low-Level Laser Therapy on the Thyroid Vascularization of Patients with Autoimmune Hypothyroidism by Color Doppler Ultrasound **Author**:Höfling, Danilo Bianchini; Chavantes, Maria Cristina; Juliano, Adriana G.; Cerri, Giovanni G.; Knobel, Meyer; Yoshimura, Elisabeth M.; Chammas, Maria Cristina **Year**:2012 View Abstract Link to paper	0.0155	(antianginal agents, selective immunosuppressants)
■ ■ ■	■ ■ ■	■ ■ ■	
10	**Title**:Fluid and Electrolyte Disturbances in Critically Ill Patients **Author**:Lee, Jay Wook **Year**:2010 View Abstract Link to paper	0.0004	(Disorders of fluid, electrolyte, and acid-base balance, hypertension, secondary hypertension)
11	**Title**:The Effects of Celecoxib or Naproxen on Blood Pressure in Pediatric Patients with Juvenile Idiopathic Arthritis **Author**:Falkner, B; Berger, M; Bhadra Brown, P; Iorga, D; Nickeson, RW; Zemel, L **Year**:2015 View Abstract Link to paper	0.0001	(hypertension, salicylates, secondary hypertension)

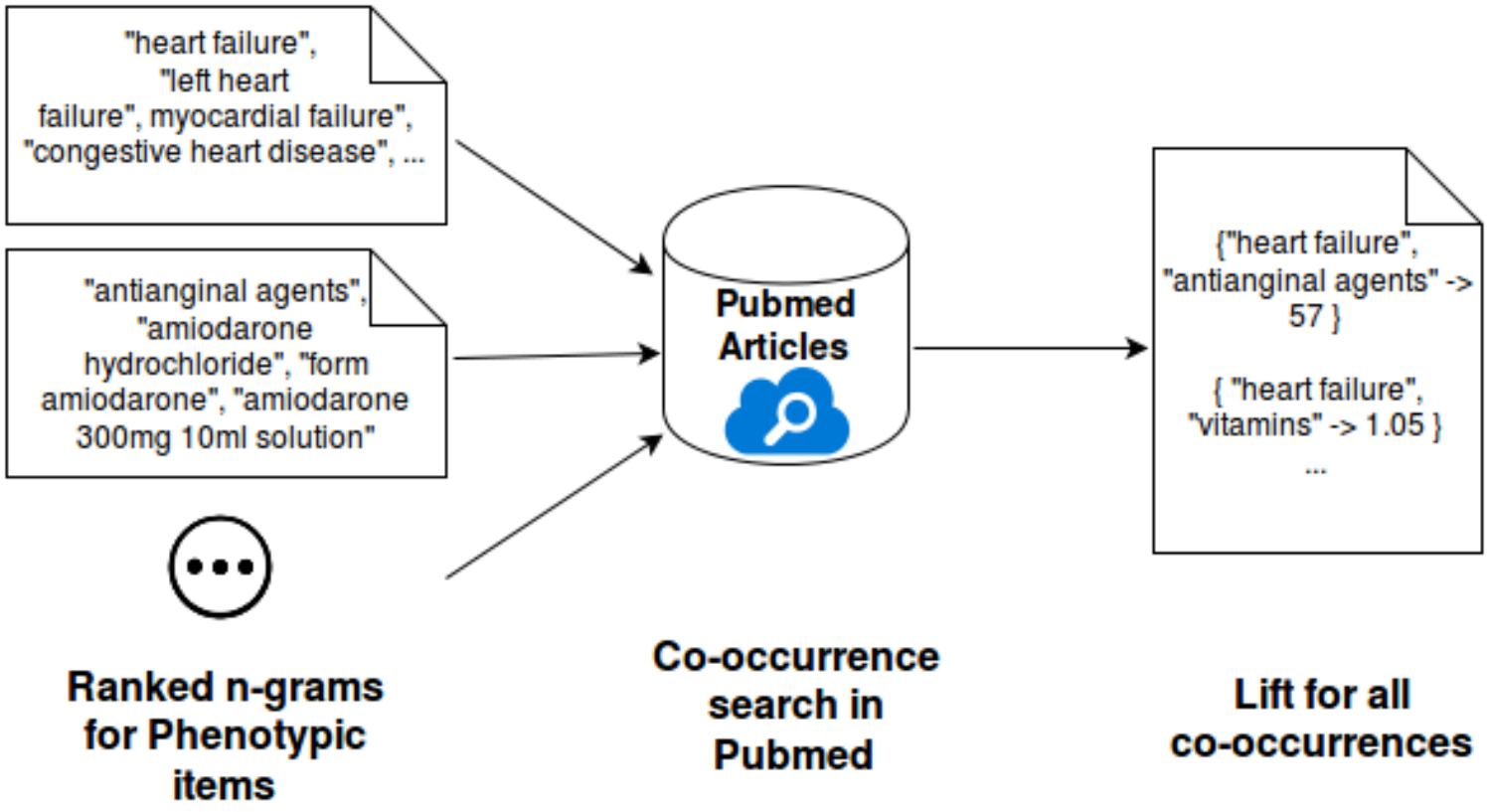
Phenotype Verification Process using Pubmed



Feature Extraction



Co-Occurrence Calculation Process



Given terms A, B, C

$$lift(A, B, C) = \frac{P(A \cap B \cap C)}{P(A) \cdot P(B) \cdot P(C)}$$

Significance Determination—Aggregating Lifts within Phenotypes

Phenotype
Rheumatoid arthritis and other inflammatory polyarthropathies
Other and unspecified disorders of joint
Osteoarthrosis and allied disorders
Osteoporosis
hypertension
Other and unspecified disorders of back
miscellaneous analgesics
antirheumatics
vitamins
cox-2 inhibitors
glucocorticoids
proton pump inhibitors
nutraceutical products

Feature Extraction

Phenotypic Term	Most Relevant Synonyms
hypertension	'regarding hypertension', 'hypertensive disorder systemic' ''ischemia due hypertension', etc.
osteoperosis	'osteoporosis postmenopausal', ''prevention osteoporosis', 'femur associated osteoporosis', etc
...	...
miscellaneous analgesics	'painful periods', ''abdominal pain finding', 'pain observable entity', etc
...	...

Co-occurrence and Lift Calculation

Co-occurrence set (represented by original phenotypic terms)	Cardinality	Standard Deviations Above Median
('hypertension','osteoporosis')	2	0.1169
('cox-2 inhibitors', 'proton pump inhibitors', 'vitamins')	3	0.0907
('cox-2 inhibitors', 'osteoporosis')	2	-0.0071
'osteoarthrosis and allied disorders', 'osteoporosis'	2	-0.0053
('osteoporosis', 'proton pump inhibitors')	2	0.0266
('osteoporosis', 'rheumatoid arthritis and other inflammatory polyarthropathies')	2	-0.0018

Average Standard Deviations Above Median : .0367

Experimental Set-up

Phenotype Data

- Random and curated phenotypes
- 80 annotated phenotypes generated by two different automatic phenotype generation algorithms
  - 14% -- clinically meaningful
  - 78% -- possibly significant
  - 8% not clinically meaningful

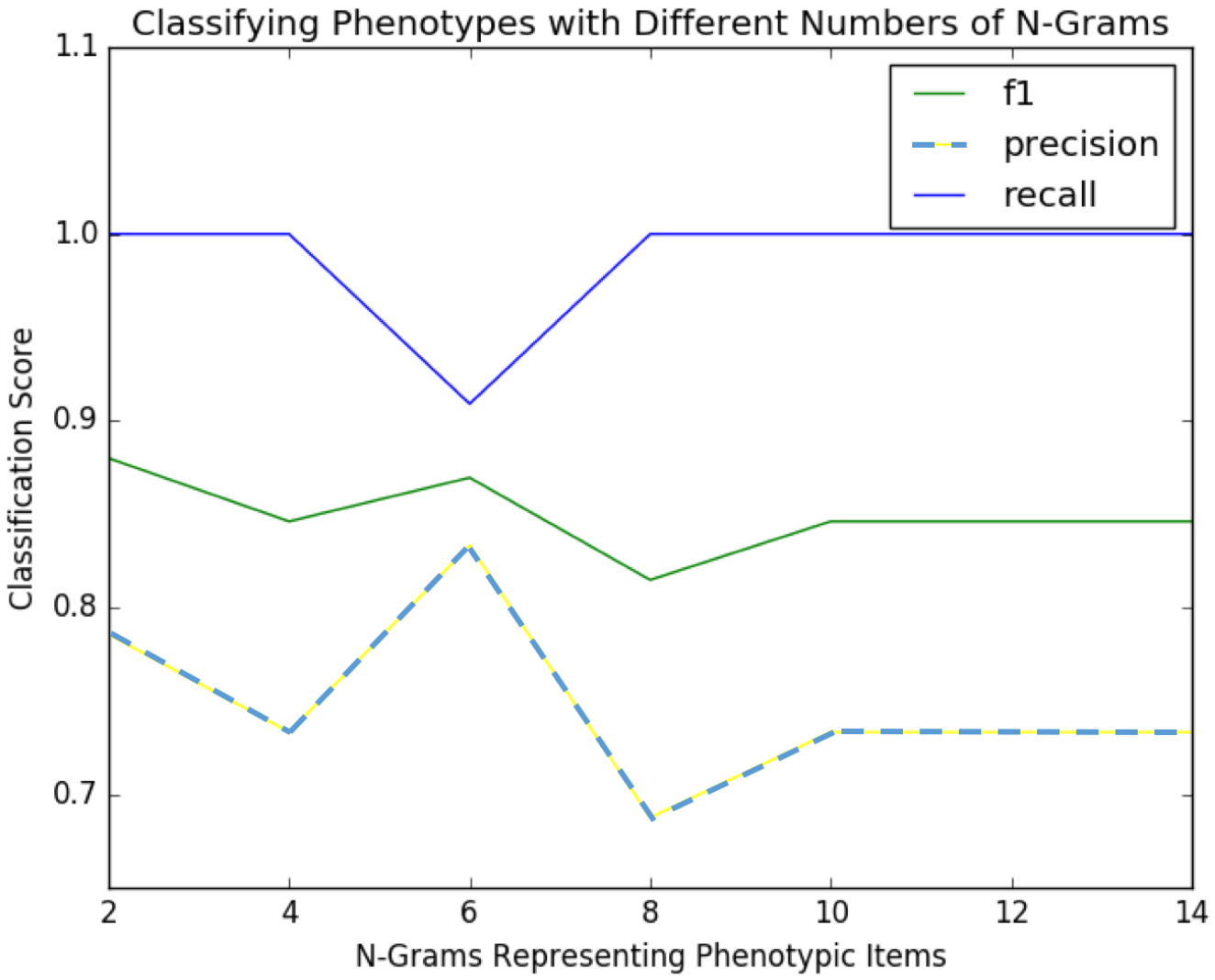
PubMed Data

- 25% of PubMed Open Access Subset

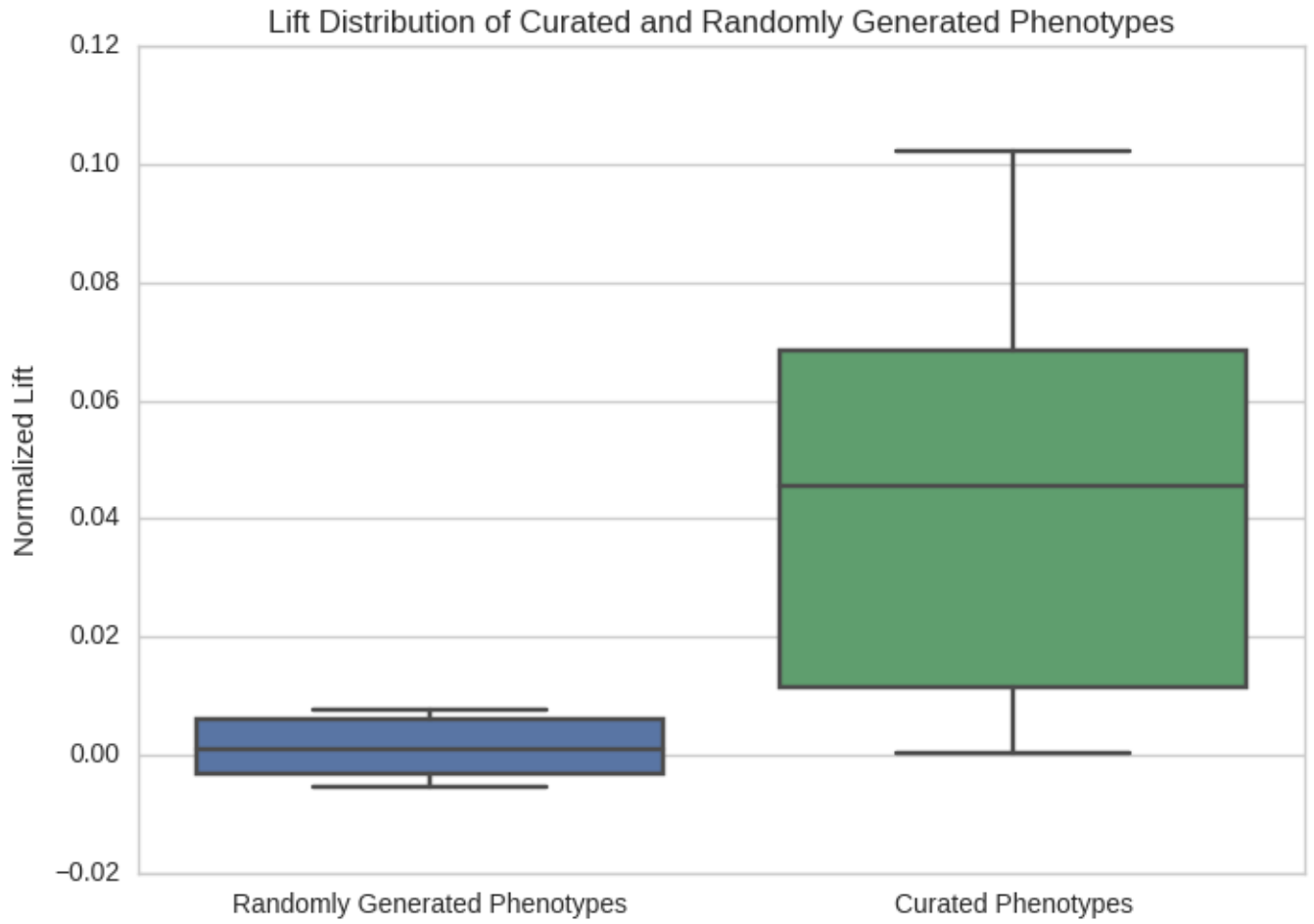
Method

1. Calculate lift
2. Determine “optimal” threshold that separates “significant” and ”not significant” phenotypes

Process Tuning—Phenotypic Item Synonym Set Size



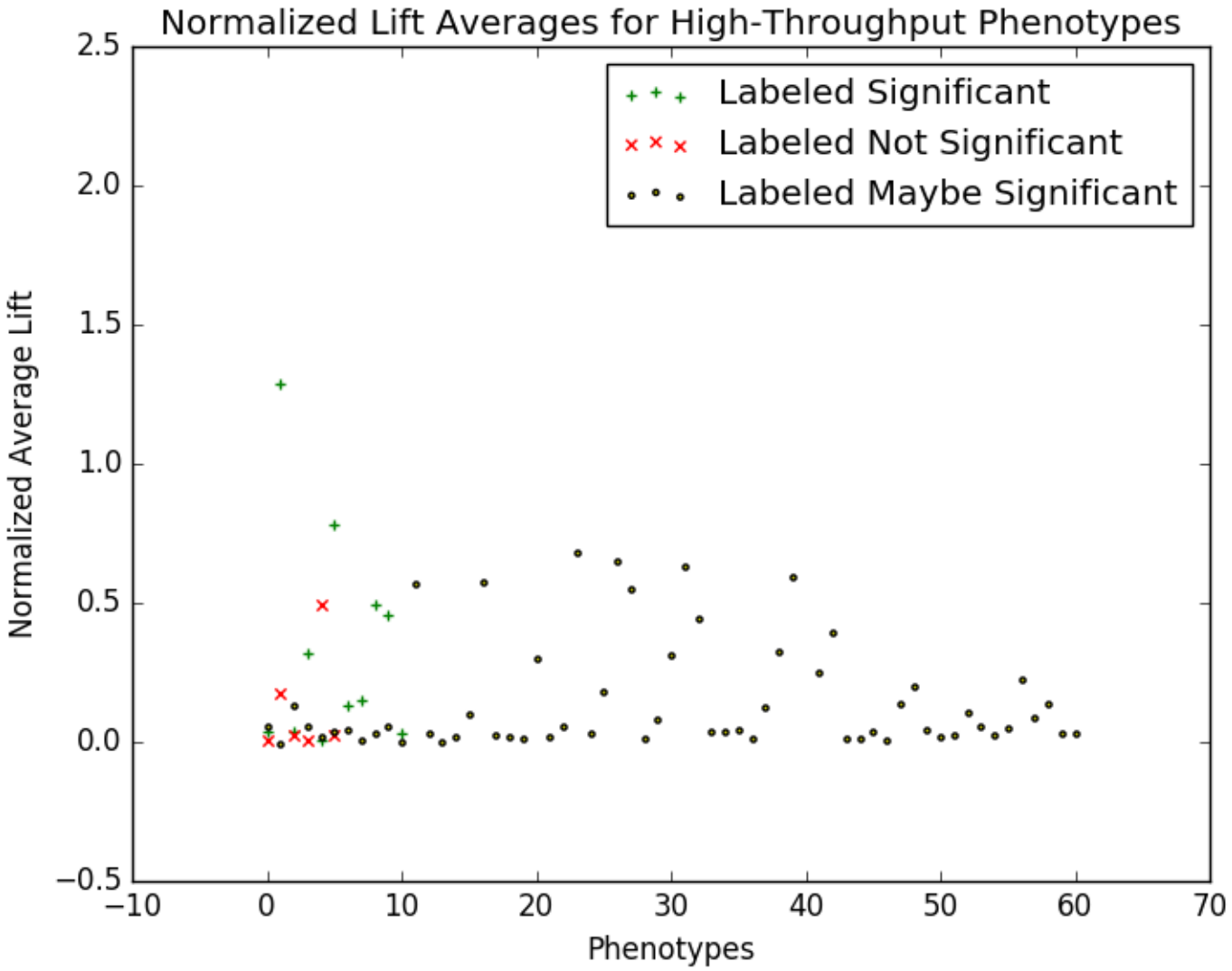
Results—Randomly Generated vs Curated Phenotypes



Classification Results

- 100% true negative classification
- 80% true positive classification
- F1 score of 0.89

Results—Automatically Generated Phenotypes



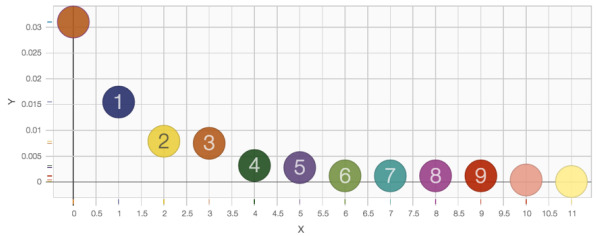
Classification Results

- Threshold = 0.028
- F1 score of 0.87

Demo

Candidate Phenotype

Disorders of fluid, electrolyte, and acid-base balance
Other and unspecified anemias
antianginal agents
salicylates
calcium channel blocking agents
secondary hypertension
selective immunosuppressants
angiotensin converting enzyme inhibitors
hypertension



This candidate phenotype has an average standard deviation (above the median) lift of **0.0064**

Table of Evidence

Index	Paper	Standard Deviations above Median Lift	Co-occurrence Tuples
0	**Title**:Unmet medical needs in lupus nephritis: solutions through evidence-based, personalized medicine **Author**:Anders, Hans-Joachim; Weidenbusch, Marc; Rovin, Brad **Year**:2015 View Abstract Link to paper	0.031	(calcium channel blocking agents, selective immunosuppressants)
1	**Title**:Assessment of the Effects of Low-Level Laser Therapy on the Thyroid Vascularization of Patients with Autoimmune Hypothyroidism by Color Doppler Ultrasound **Author**:Höfling, Danilo Bianchini; Chavantes, Maria Cristina; Juliano, Adriana G.; Cerri, Giovanni G.; Knobel, Meyer; Yoshimura, Elisabeth M.; Chammas, Maria Cristina **Year**:2012 View Abstract Link to paper	0.0155	(antianginal agents, selective immunosuppressants)
■ ■ ■ ■ ■ ■ ■ ■ ■			
10	**Title**:Fluid and Electrolyte Disturbances in Critically Ill Patients **Author**:Lee, Jay Wook **Year**:2010 View Abstract Link to paper	0.0004	(Disorders of fluid, electrolyte, and acid-base balance, hypertension, secondary hypertension)
11	**Title**:The Effects of Celecoxib or Naproxen on Blood Pressure in Pediatric Patients with Juvenile Idiopathic Arthritis **Author**:Falkner, B; Berger, M; Bhadra Brown, P; Iorga, D; Nickeson, RW; Zemel, L **Year**:2015 View Abstract Link to paper	0.0001	(hypertension, salicylates, secondary hypertension)

(hypertension, immunosuppressants)

ckj OXFORD

Leading European Nephrology

Clinical Kidney Journal, 2015, vol. 8, no. 5, 492–502

doi: 10.1093/ckj/sfv072
Advance Access Publication Date: 27 August 2015
CKJ Review

Genetic/metabolic risk stratification, combination of low-dose immunosuppressants with anti-inflammatory drugs, favor specific drugs over unselective immunosuppressants

CKJ REVIEW

Unmet medical needs in lupus nephritis: solutions through evidence-based, personalized medicine

Hans-Joachim Anders[1], Marc Weidenbusch[1], and Brad Rovin[2]

Clinical criteria
Male gender, older age, hypertension, increased SCr
Anti-snRNP, high SLE activity/anti-dsDNA, childhood-onset SLE, race, family history of diabetes and/or hypertension
Pre-term birth, birth weight, male gender, race (Afro-Americans, Hispanics), hypertension, kidney biopsy (LN Class III–VI, chronicity index/extent of scaring ≈ lost nephrons), SCr, failure to respond to induction therapy (proteinuria), number of flares, progressive fibrosis on re-biopsy

Review

ISSN 1738-5997 (Print) • ISSN 2092-9935 (Online)
Electrolyte Blood Press 8:72-81, 2010 • doi: 10.5049/EBP.2010.8.2.72

Fluid and Electrolyte Disturbances in Critically Ill Patients

Jay Wook Lee, M.D.

Disturbances in fluid and elect

(Disorders of fluid, electrolyte, and acid-ba balance, hypertension, secondary hypertension)

excretion. If it is appropriately low (i.e., 24-hour urine K^+ < 20 mEq/day or random urine K^+/creatinine < 15 mEq/g or 1.5 mEq/mmol), transcellular shift or extrarenal K^+ loss should be suspected. If urinary K^+ excretion is high, transtubular potassium gradient (TTKG), acid-base status, and the presence or absence of hypertension are helpful in differential diagnosis of hypokalemia due to renal potassium loss. A TTKG larger than 4 suggests that there is an increase in K^+ secretion into the cortical collecting duct, i.e., a high K^+ concentration in the cortical collecting duct.

Future Work

- Incorporate ”gold standard” phenotypes from PheKB and other sources
- Scale to whole PubMed Open Access Subset
- Speed up co-occurrence analysis
- Refine and automate phenotype classification process

References

M. R. Boland, Z. Shahn, D. Madigan, G. Hripcsak, and N. P. Tatonetti. Birth month affects lifetime disease risk: a phenome-wide method. Journal of the American Medical Informatics Association, page ocv046, 2015.

R. J. Carroll, A. E. Eyler, and J. C. Denny. Naive electronic health record phenotype identification for rheumatoid arthritis. In AMIA Annu Symp Proc, volume 2011, pages 189–96, 2011.

Y. Chen, R. J. Carroll, E. R. M. Hinz, A. Shah, A. E. Eyler, J. C. Denny, and H. Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association, 20(e2):e253–e259, 2013.

K. Dickersin. The existence of publication bias and risk factors for its occurrence. Jama, 263(10):1385–1389, 1990.

P. J. Easterbrook, R. Gopalan, J. Berlin, and D. R. Matthews. Publication bias in clinical research. The Lancet, 337(8746):867–872, 1991.

M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, W.-Q. Wei, S. J. Bielinski, C. G. Chute, C. L. Leibson, D. R. Crosslin, C. S. Carlson, K. M. Newton, W. A. Wolf, R. L. Chisholm, and W. L. Lowe. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association, 19(2):212–218, Mar. 2012.

J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. Journal of Biomedical Informatics, 52:199–211, Dec. 2014.

J. C. Ho, J. Ghosh, and J. Sun. Marble: High-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 115–124, 2014.

G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1):117–121, 2013.

C. Hu, P. Rai, C. Chen, M. Harding, and L. Carin. Scalable bayesian non-negative tensor factorization for massive count data. In Machine Learning and Knowledge Discovery in Databases, pages 53–70. Springer, 2015.

References

L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews genetics, 7(2):119–129, 2006.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv.org, Jan. 2013.

A. N´ev´eol, R. I. Dog˘an, and Z. Lu. Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction. Journal of Biomedical Informatics, 44(2):310–318, 2011.

NIH Health Care Systems Research Collaboratory. Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. July 2014.

S. Pletscher-Frankild, A. Palleja`, K. Tsafou, J. X. Binder, and L. J. Jensen. Diseases: Text mining and data integration of disease–gene associations. Methods, 74:83–89, 2015.

D. K. Rajpal, X. A. Qu, J. M. Freudenberg, and V. D. Kumar. Mining emerging biomedical literature for understanding disease associations in drug discovery. Biomedical Literature Mining, pages 171–206, 2014.

J. M. Stern and R. J. Simes. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. Bmj, 315(7109):640–645, 1997.

Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen, B. A. Malin, and J. Sun. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1265–1274. ACM, 2015.

H. Wasserman and J. Wang. An applied evaluation of snomed ct as a clinical vocabulary for the computerized diagnosis and problem list. In AMIA Annual Symposium Proceedings, volume 2003, page 699. American Medical Informatics Association, 2003.

S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. Journal of the American Medical Informatics Association, 22(5):993–1000, Apr. 2015.