Efficient Development of Electronic Health Record Based Algorithms to Identify Rheumatoid Arthritis

Robert Carroll, Vanderbilt University

Rheumatoid Arthritis

- Prototypic chronic disease
- Inflammatory polyarthritis afflicting 1.3 million adults in the United States[1]
- 50% increased risk of premature mortality[2]
- Life expectancy is decreased by 3 to 10 years compared with the general population[2]
- Autoimmune disorder
- Primarily affects joints

1. Helmick CG, Felson DT, Lawrence RC, Gabriel S, Hirsch R, Kwoh CK, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. Arthritis Rheum. 2008 Jan;58(1):15-25.
2. Myasoedova E, Davis JM 3rd, Crowson CS, Gabriel SE. Epidemiology of rheumatoid arthritis: rheumatoid arthritis and mortality. Curr Rheumatol Rep. 2010 Oct;12(5):379-385.

How to find RA in the EHR

- Structured data
  - Billing Codes
  - Lab results
- Medications
  - Codified from the order entry system
  - Free text from other sources
- Free text data
  - Physician notes
  - Radiology reports

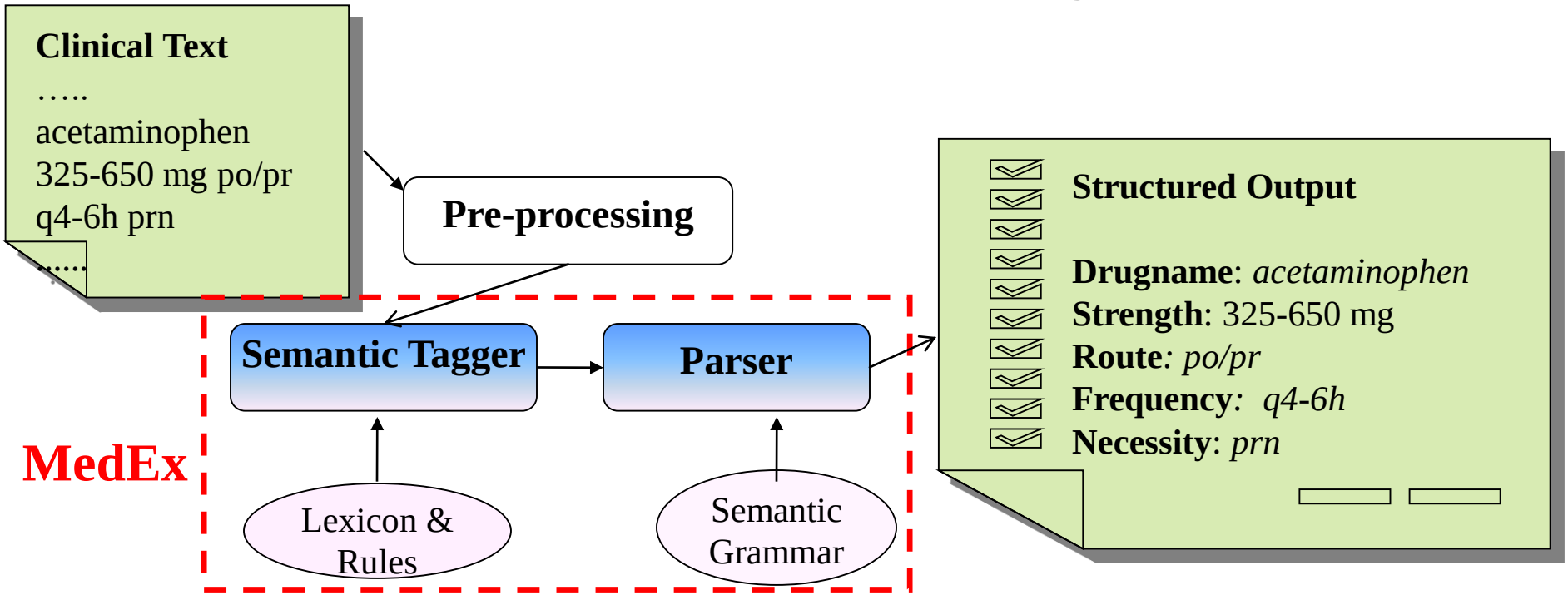
Billing data

- Why not just look for billing codes?
  - A VA database study showed that only 57% of individuals with one RA ICD-9 billing code were RA postive.[1]
  - Another study at Partners HealthCare showed that only 56% of individuals with three or more codes were RA positive.[2]

1. Singh JA, Holmgren AR, Noorbaloochi S. Accuracy of Veterans Administration databases for a diagnosis of rheumatoid arthritis. Arthritis Rheum 2004;51:952–7.
2. Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;**62**:1120–7.

MedEx

- Medex identifies medication references from clinical text.
- It additionally records the dose, route, and frequency information for the drug reference.



KnowledgeMap Concept Identifier (KMCI)

- Identifies UMLS concepts from text
- It uses context to help disambiguate words and phrases.
- Incorporates:
  - NegEx - identifies negation
  - SecTag - identifies note sections, aiding in context determination. It also helps separate out Family History from the rest of the patient's information.

How to find RA in the EHR

- Codified data
  - Billing Codes
  - Lab results
- Medications
  - Codified from the order entry system
  - Free text from other sources
- Free text data
  - Physician notes
  - Radiology reports

Phenotype algorithms

- We now have access to multimodal data.
- How can we incorporate it all?

- Deterministic Algorithm

Apply a Deterministic Algorithm

- Development cycle
  - Create a set of rules
  - Iterate:
    - Generate a set of cases and controls
    - Judge the algorithm
    - If performance has not reached threshold, modify the rules
- This method did replicate a known genetic association for RA.

Ritchie MD, Denny JC, Crawford DC, *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;**86**:560–72.

RA Deterministic Algorithm

ICD 9 codes (any of the below)

- 714 Rheumatoid arthritis and other inflammatory polyarthropathies
- 714.0 Rheumatoid arthritis
- 714.1 Felty's syndrome
- 714.2 Other rheumatoid arthritis with visceral or systemic involvement

AND

Medications (any of the below)

methotrexate [MTX][amethopterin] sulfasalazine [azulfidine]; Minocycline [minocin][solodyn]; hydroxychloroquine [Plaquenil]; adalimumab [Humira]; etanercept [Enbrel] infliximab [Remicade]; Gold [myochrysine]; azathioprine [Imuran]; rituximab [Rituxan] [MabThera]; anakinra [Kineret]; abatacept [Orencia]; leflunomide [Arava]

AND

Keywords (any of the below)

rheumatoid [rheum] [reumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

RA Deterministic Algorithm Exclusions

AND NOT
ICD 9 codes (any of the below)

- 714.30 Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
- 714.31 Polyarticular juvenile rheumatoid arthritis, acute
- 714.32 Pauciarticular juvenile rheumatoid arthritis
- 714.33 Monoarticular juvenile rheumatoid arthritis
- 695.4 Lupus erythematosus
- 710.0 Systemic lupus erythematosus
- 373.34 Discoid lupus erythematosus of eyelid
- 710.2 Sjogren's disease
- 710.3 Dermatomyositis
- 710.4 Polymyositis
- 555 Regional enteritis
- 555.0 Regional enteritis of small intestine
- 555.1 Regional enteritis of large intestine
- 555.2 Regional enteritis of small/large intestine
- 555.9 Regional enteritis of unspecified site
- 564.1 Irritable Bowel Syndrome
- 135 Sarcoidosis
- 696 Psoriasis and similar disorders
- 696.0 Psoriatic arthropathy
- 696.1 Other psoriasis and similar disorders excluding psoriatic arthropathy
- 696.8 Other psoriasis and similar disorders
- 099.3 Reiter's disease
- 716.8 Arthropathy, unspecified
- 274.0 Gouty arthropathy
- 358.0 myasthenia gravis
- 358.00 myasthenia gravis without acute exacerbation
- 358.01 myasthenia gravis with acute exacerbation
- 775.2 neonatal myasthenia gravis
- 719.3 Palindromic rheumatism
- 719.30 Palindromic rheumatism, site unspecified
- 719.31 Palindromic rheumatism involving shoulder region
- 719.32 Palindromic rheumatism involving upper arm
- 719.33 Palindromic rheumatism involving forearm
- 719.34 Palindromic rheumatism involving hand
- 719.35 Palindromic rheumatism involving pelvic region and thigh
- 719.36 Palindromic rheumatism involving lower leg
- 719.37 Palindromic rheumatism involving ankle and foot
- 719.38 Palindromic rheumatism involving other specified sites
- 719.39 Palindromic rheumatism involving multiple sites
- 720 Ankylosing spondylitis and other inflammatory spondylopathies
- 720.0 Ankylosing spondylitis
- 720.8 Other inflammatory spondylopathies
- 720.81 Inflammatory spondylopathies in diseases classified elsewhere
- 720.89 Other inflammatory spondylopathies
- 720.9 Unspecified inflammatory spondylopathy
- 721.2 Thoracic spondylosis without myelopathy
- 721.3 Lumbosacral spondylosis without myelopathy
- 729.0 Rheumatism, unspecified and fibrositis
- 340 Multiple sclerosis
- 341.9 Demyelinating disease of the central nervous system unspecified
- 323.9 transverse myelitis
- 710.1 Systemic sclerosis
- 245.2 Hashimoto's thyroiditis
- 242.0 Toxic diffuse goiter
- 443.0 Raynaud's syndrome

OR
Keywords (any of the below)

juvenile [juv] rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]
juvenile [juv] arthritis arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]
juvenile chronic arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]
juvenile [juv] RA; JRA
Inflammatory [inflamatory] [inflam] osteoarthritis [osteoarthrosis] [OA]
Reactive [psoriatic] arthritis [arthropathy] [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]

Phenotype algorithms

- Deterministic
  - Provides reasonable performance
  - Relies on the expert designer
- Machine Learning (ML)

Applying ML Algorithms

- Select attributes of interest, including related conditions or potential misdiagnoses
- Review subjects for training set
- Train a model and apply it to the record pool
- Select predictive positive patients and review
- Ensure performance meets standard

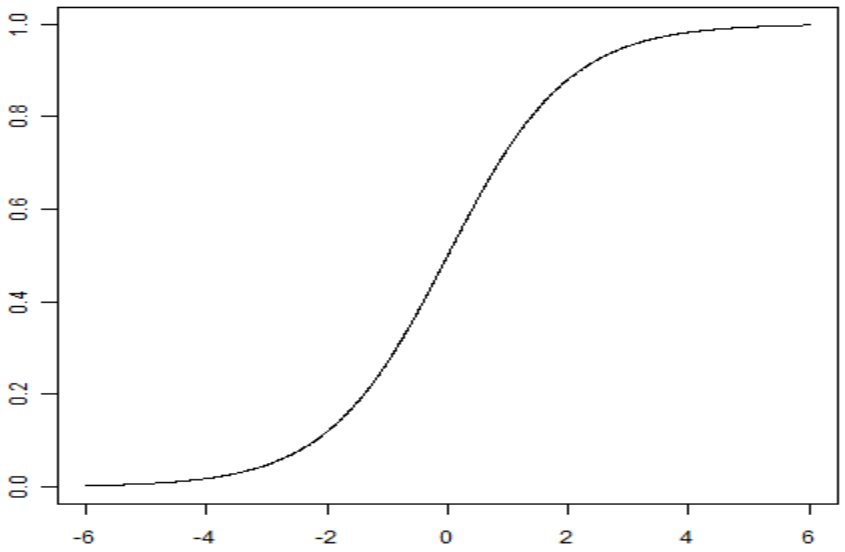
- This method has also been shown to be effective in genetic studies

Liao KP, Cai T, Gainer V, *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)* 2010;**62**:1120–7

Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am. J. Hum. Genet. 2011 Jan 7;88(1):57–69.

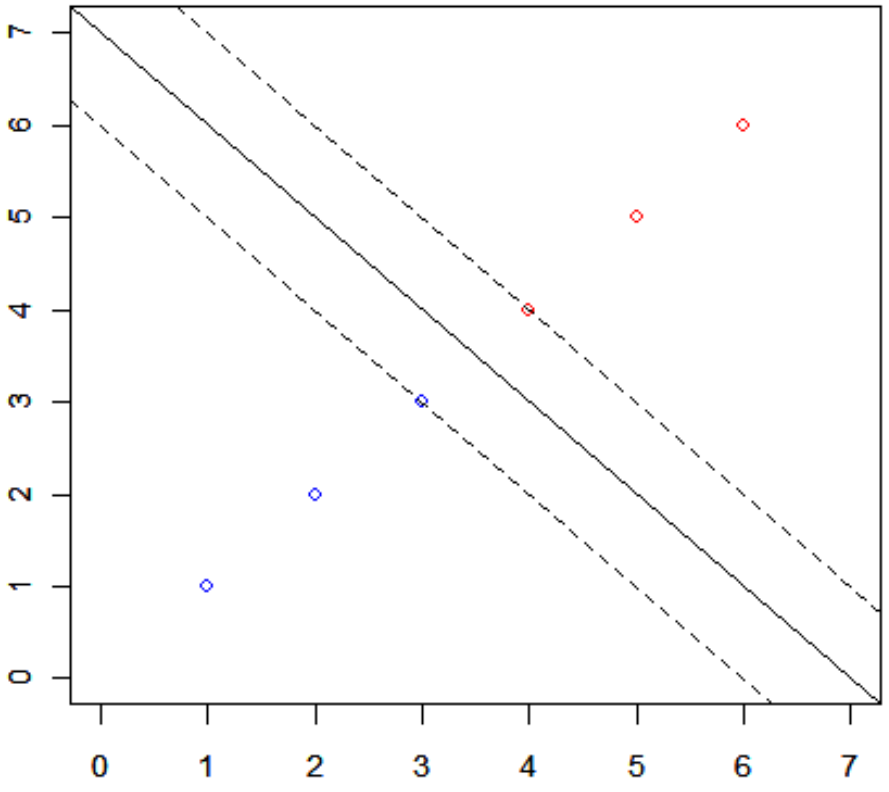
Logistic Regression

- Two-group classifier
- Finds the optimal weight for every attribute to best predict the class
  - Z = Intercept + Weight 1 * Attribute 1 + W2 * A2…
  - Maps the Z score onto a logistic curve
  - $F(Z)= 1/(1+e^{-z})$
  - Outputs a probability
  - Weights have some interpretation



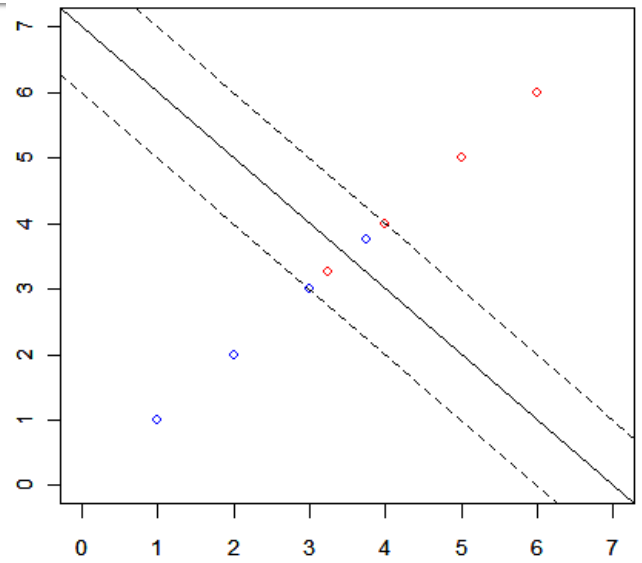
Support Vector Machine

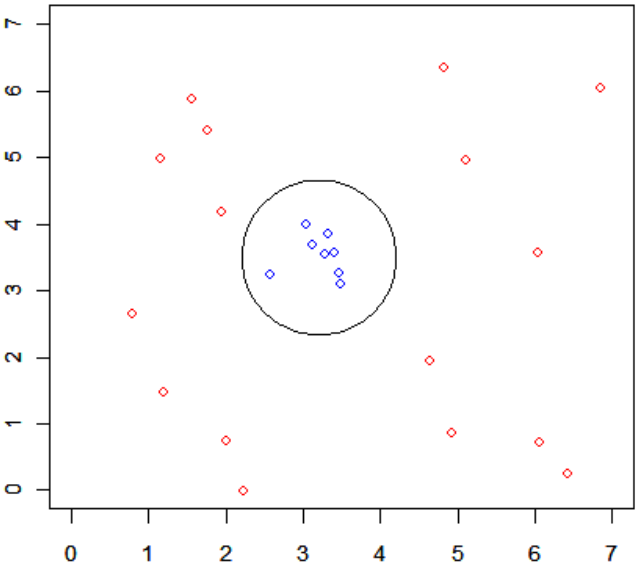
- Two-group classifier
- Finds the best hyperplane to separate the two classes
  - Each attribute becomes a dimension
  - In the case of two dimensions, the hyperplane is a line.
  - Maximizes the distance from the points to the hyperplane.



SVM Modifications

- Soft margins: Allows points to be within the margin or across the hyperplane, but with a penalty.
- Kernel methods: Maps the attributes to another dimension, allowing for interactions between the attributes



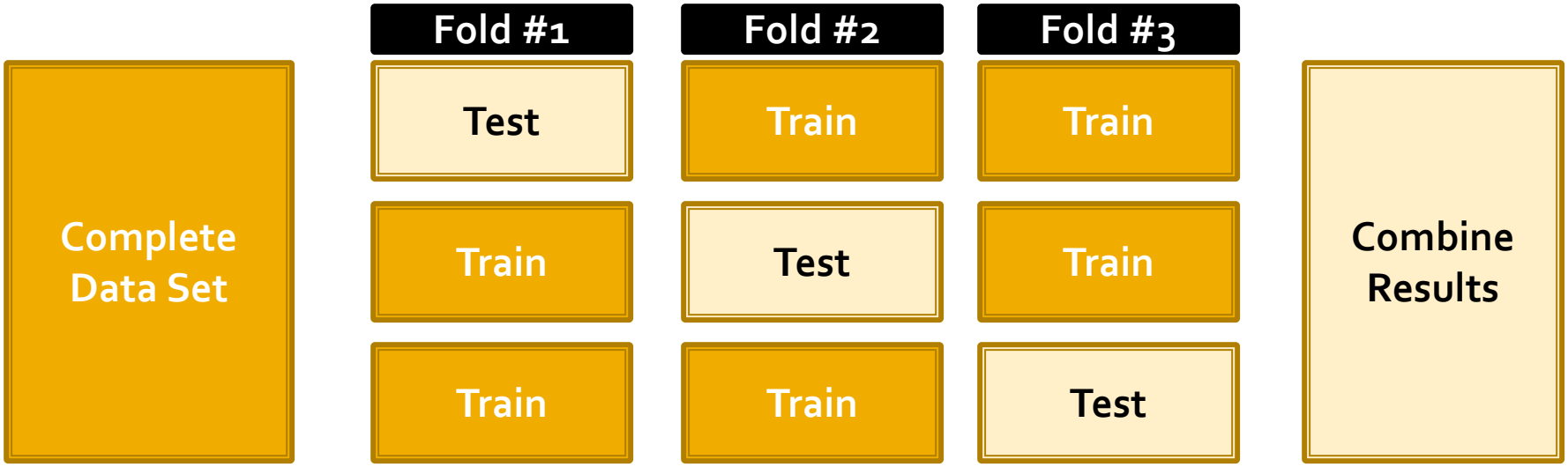


Overfitting

- Machine learning algorithms are designed to identify patterns.
- They can only learn what they are shown however.
  - Their behavior is not always well defined outside of the training space.
  - They can't tell the difference between erroneous patterns and clinically significant patterns.
- More attributes means more coincidental patterns, so choosing relevant attributes can help limit overfitting the training data.

Cross Validation

- Used to estimate more accurately the performance for algorithms being trained and tested on the same pool of data.
- Splits up the data set into groups.
  - 3-fold CV means three different groups
  - Train on two groups and test on the third
  - Repeat three times



Phenotype algorithms

- Deterministic
  - Provides reasonable performance
  - Relies on the expert designer
- Machine Learning (ML)
  - Allows for more complex relationships
  - Reliance on the perceived relevance of attributes
- The naïve approach

Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis

Hypothesis

A support vector machine (SVM) trained using a naïve data set can identify rheumatoid arthritis (RA) cases as well as one trained using an expert defined set of attributes.

Methods

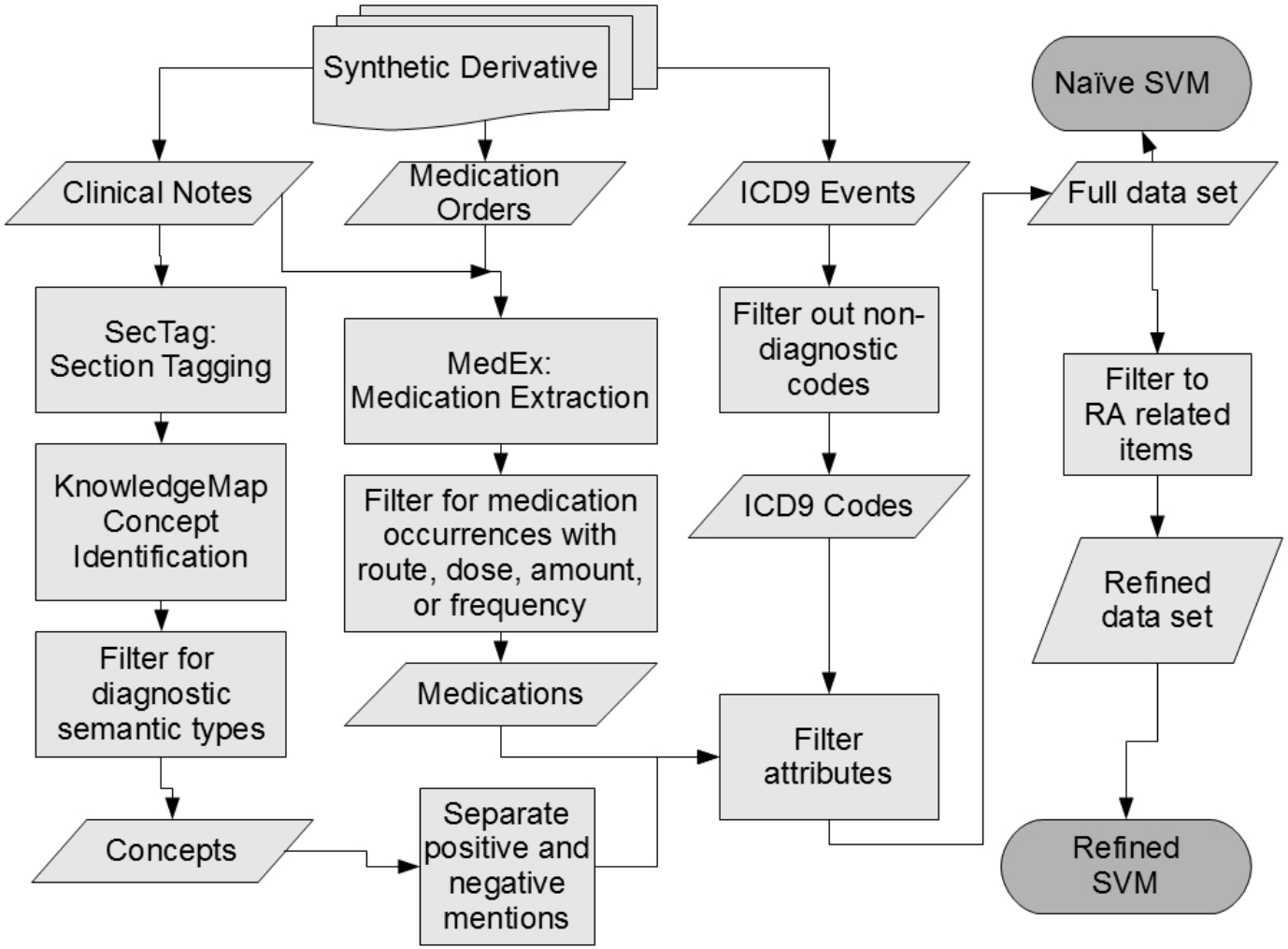
- Cohort
  - 376 records
  - Rheumatologist reviewed
  - Individuals with at least one RA ICD-9
- EHR-derived attributes
  - ICD-9 counts
  - Medication Entries
  - Unified Medical Language System (UMLS) Concepts

Our population

	Vanderbilt (n=376)	
	RA	**Non-RA**
Total	185 (49.2%)	191 (50.8%)
Age	52.9 ± 13.1	56.2 ± 16.5
Female	148 (80.0%)	141 (73.8%)
Ethnicity		
Caucasian	143 (77.3%)	155 (81.2%)
African American	14 (7.6%)	26 (13.6%)
Hispanic	1 (0.5%)	1 (0.5%)
Other	3 (1.6%)	2 (1.0%)
Unknown	24 (13.0%)	7 (3.7%)
Drugs		
Anti-TNF use	88 (47.6%)	26 (13.6%)
MTX	133 (71.9%)	63 (33.0%)
Codes		
RA	185 (100.0%)	191 (100.0%)
SLE	14 (7.6%)	32 (16.8%)
JRA	6 (3.2%)	8 (4.2%)
PsA	6 (3.2%)	14 (7.3%)
EHR Followup*	9.97 ± 4.06	9.06 ± 4.32

*Mean ± standard deviation in years

Attribute Creation



Methods: Evaluation

- Algorithms
  - Prior Published Deterministic[1]
  - Support Vector Machine
    - Naïve and Gaussian kernel
    - Refined and Gaussian kernel
- Method
  - 10 fold cross validation
  - Nested CV for gamma and cost parameters
  - Subset evaluations (e.g., ICD9 only, Meds only)

1. Ritchie MD, et al. Am. J. Hum. Genet. 2010;86(4):560-572

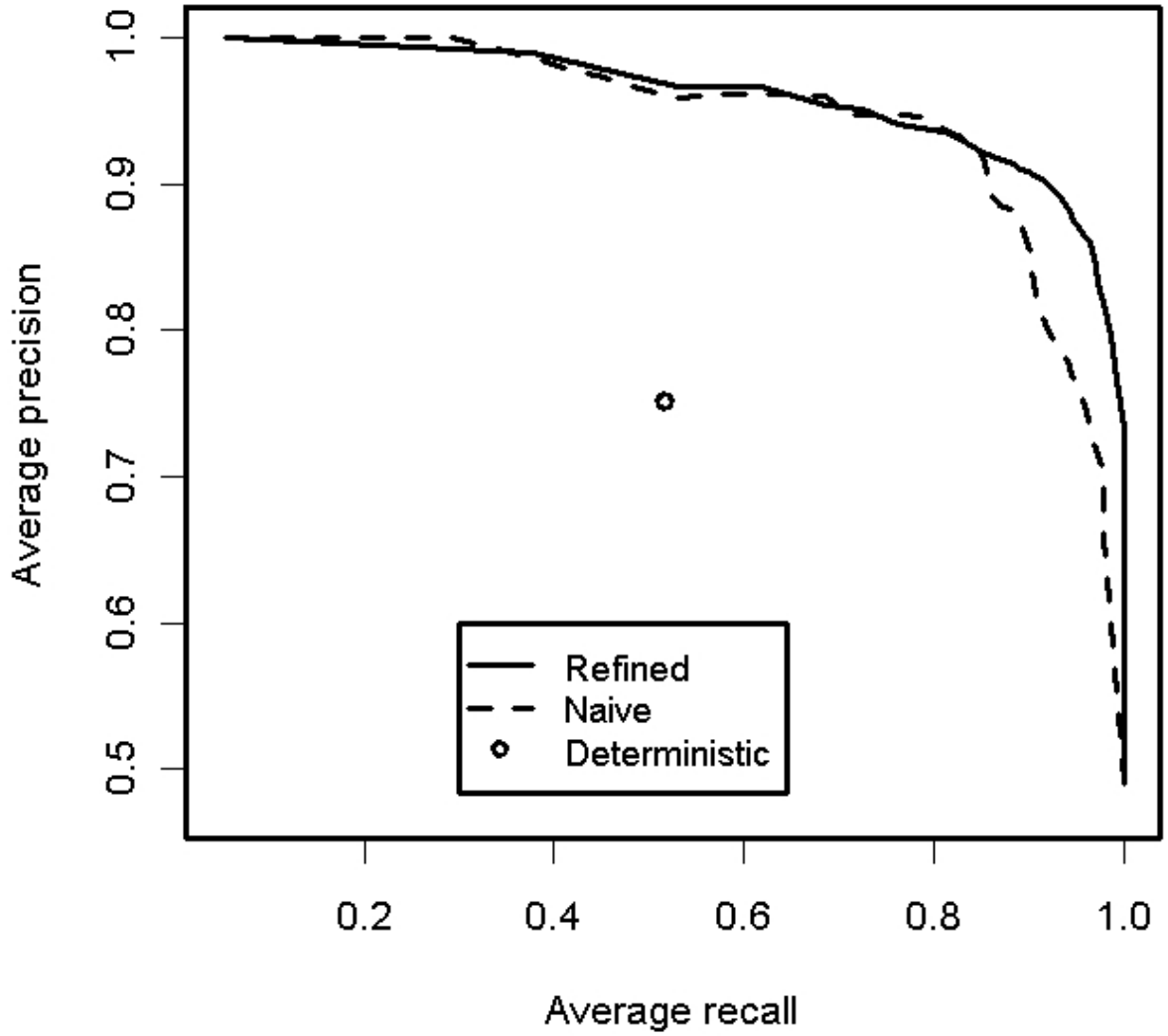
Measurements

- Recall (Sensitivity)
  - tp/(tp+fn)
- Precision (Positive Predictive Value):
  - tp/(tp+fp)
- Specificity
  - tn/(tn+fp)
- F-measure
  - 2*(precision*recall)/(precision+recall)
- Area under the receiver operator characteristic curve (AUC)

Model Performance

Naïve	Precision	Recall	F measure	AUC	Attributes
Full	**93.3 ± 0.5**	**79.7 ± 5.2**	**85.1 ± 3.7**	**94.2 ± 1.3**	**17110**
ICD-9	94.1 ± 0.2	87.1 ± 2.8	90.3 ± 1.6	95.6 ± 1.0	795
NLP	92.2 ± 0.6	68.2 ± 5.6	77.4 ± 4.1	90.4 ± 2.1	15171
Medication	88.9 ± 1.8	51.0 ± 5.4	63.5 ± 5.5	84.6 ± 2.6	1148
Refined	**Precision**	**Recall**	**F measure**	**AUC**	**Attributes**
Full	**93.7 ± 0.6**	**85.8 ± 5.7**	**88.6 ± 4.0**	**96.6 ± 1.1**	**59**
ICD-9	93.2 ± 0.5	78.1 ± 5.2	84.2 ± 3.5	95.5 ± 1.3	12
NLP	91.8 ± 1.0	68.8 ± 7.5	76.8 ± 5.7	89.5 ± 2.1	33
Medication	86.6 ± 1.6	40.5 ± 5.4	53.8 ± 5.2	83.3 ± 2.5	18
Deterministic	**Precision**	**Recall**	**F measure**	**AUC**	**Attributes**
Full	75.2 ± 2.5	51.6 ± 2.6	60.5 ± 2.6	N/A	N/A

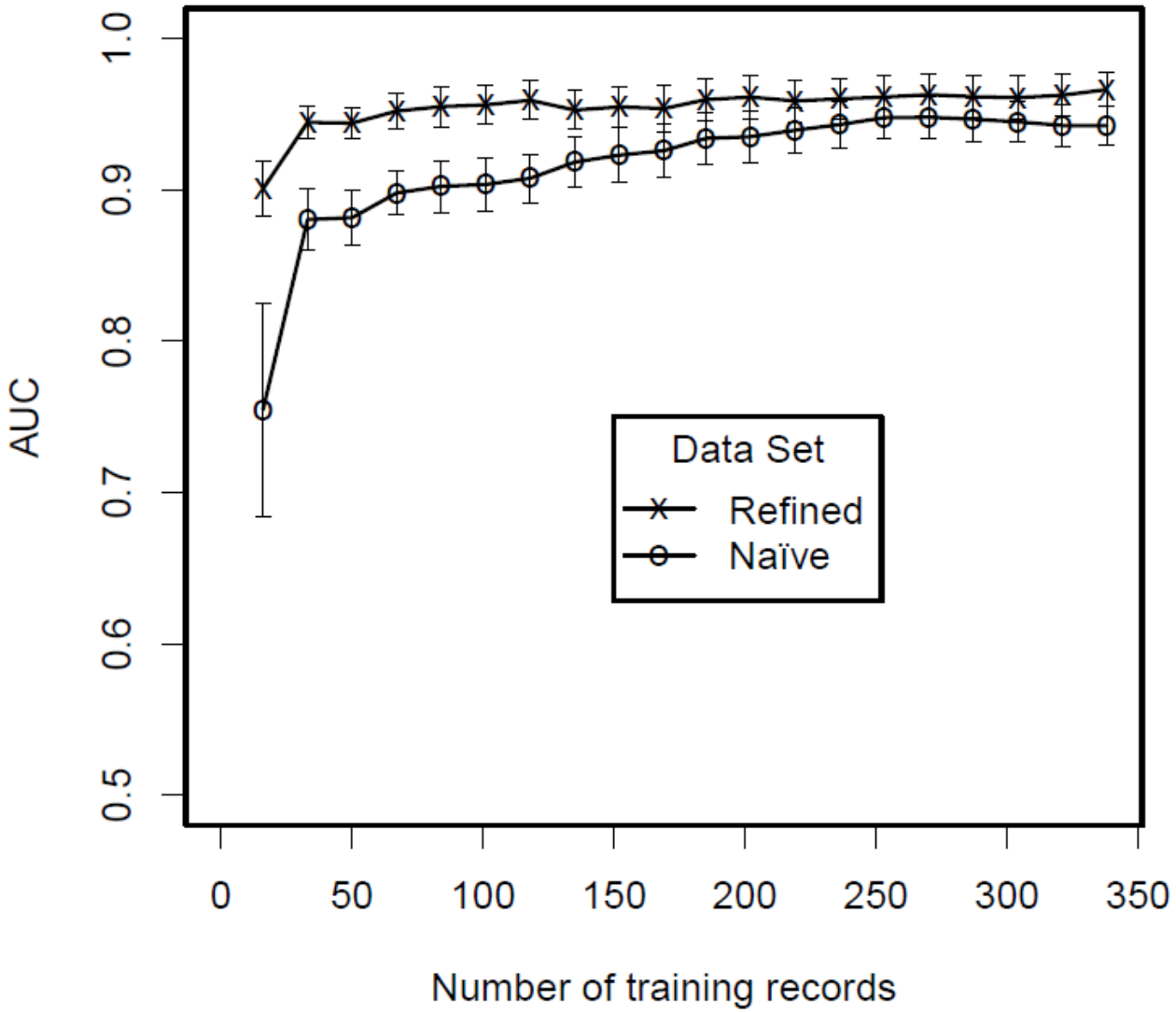
Precision Recall Curves



Evaluation: How many cases are needed to train?

- We used the same cross validation, but repeated it 20 times.
- Each time we increased the number of training samples available to the fold
- Approximated an iterative algorithm development approach

How many cases are required to train an accurate algorithm?



Discussion

- Both naïve and refined SVMs performed better than expert-designed algorithms
- Naïve SVMs can substitute for refined SVMs
- Naïve SVMs have some disadvantages
  - They require more training (100 cases vs. 20)
  - They have lower recall (6%)
- The NLP and medication methods had large standard error
- Medications alone had low recall and poor precision

Limitations

- RA is well represented by distinct ICD9 codes (though only 50% of cases were positive)
- RA is a chronic condition, yielding many relevant clinical encounters over time
- Only data from one institution was analyzed

Portability of an Algorithm to Identify Rheumatoid Arthritis in Electronic Health Records

Are phenotype algorithms portable to different EHRs?

- Previously published logistic regression model
  - Trained at Partners Healthcare
  - ICD9, Labs, Meds, NLP
  - Tested at Northwestern and Vanderbilt
- Is the disease signature the same across
  - Differing healthcare environments?
  - Differing EHR systems?
  - Differing information extraction techniques?

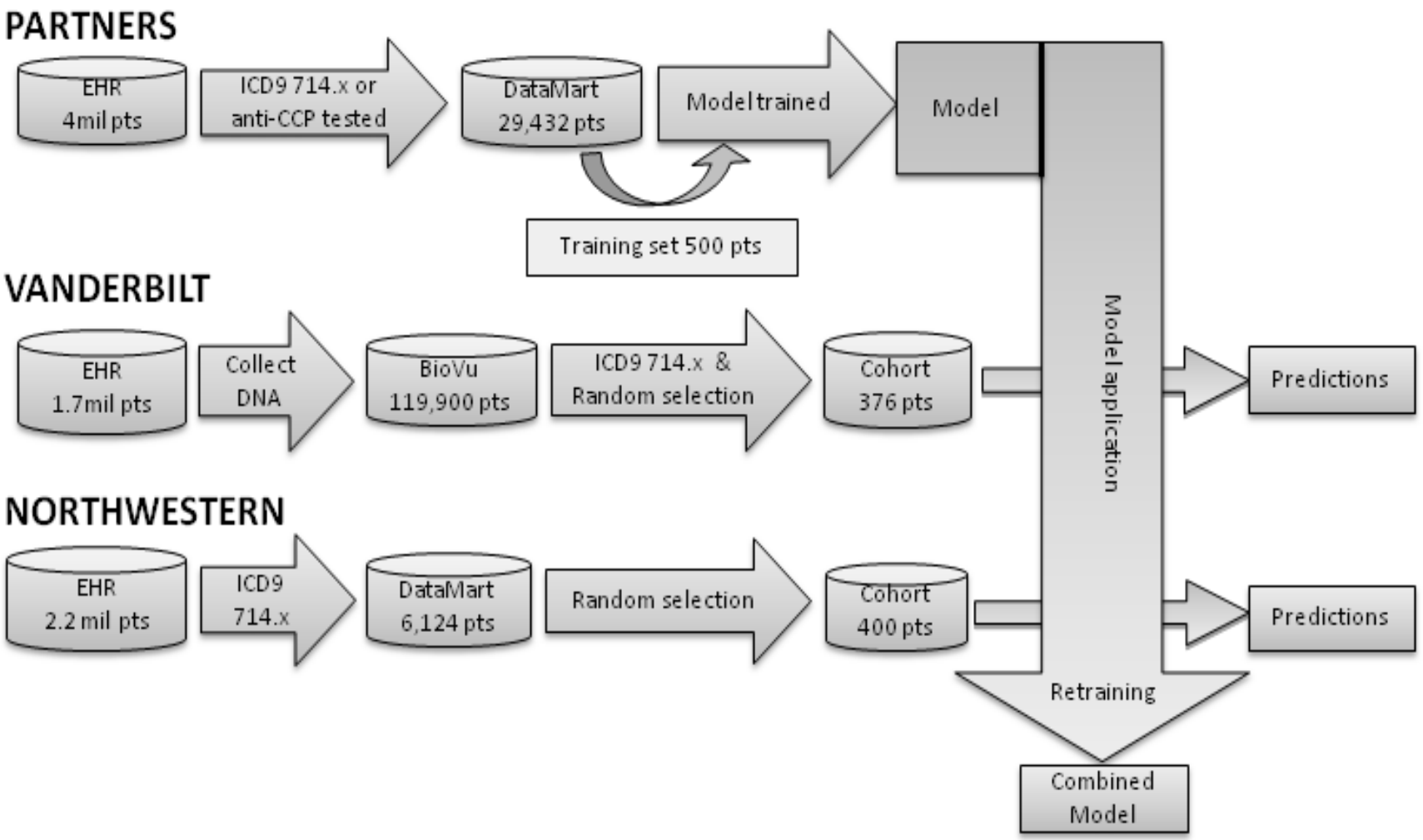
Hypothesis

- In spite of differing healthcare environments, EHR systems, and information extraction techniques, a logistic regression model trained at one institution will predict RA status at other institutions.

Overview of underlying differences

	Implementations by Institution		
	Partners Boston, MA	**Northwestern Chicago, IL**	**Vanderbilt Nashville, TN**
EHR system	Internally-developed	EpicCare (Outpatient) and Cerner PowerChart (Inpatient)	Internally-developed
Number of patients	4 million	2.2 million	1.7 million
Research EHR data	Enterprise Data Warehouse	Enterprise Data Warehouse	De-identified image of EHR (Synthetic Derivative)
Medication Source	Structured medication entries (inpatient and outpatient) and text queries	Structured outpatient medication entries and inpatient and outpatient text queries	NLP (MedEx) for outpatient medications and structured inpatient records
NLP system (disease concepts, lab results, medications, erosions)	HITEx	HITEx	KnowledgeMap Concept Identifier
NLP concept queries	Customized RegEx queries	Customized RegEx queries from Partners	Generic UMLS concepts, derived from KnowledgeMap web interface

Overview


PARTNERS
EHR
4mil pts
ICD9 714.x or
anti-CCP tested
DataMart
29,432 pts
Model trained
Model
Training set 500 pts
VANDERBILT
EHR
1.7mil pts
Collect
DNA
BioVu
119,900 pts
ICD9 714.x &
Random selection
Cohort
376 pts
Model application
Predictions
NORTHWESTERN
EHR
2.2 mil pts
ICD9
714.x
DataMart
6,124 pts
Random selection
Cohort
400 pts
Predictions
Retraining
Combined
Model

Methods: Partners RA algorithm

- Selected and grouped ICD-9 codes for four diseases: RA, Juvenile Rheumatoid Arthritis, Psoriatic Arthritis, and Systemic Lupus Erythematosus
- Chose three categories of RA medications.
- Uses some NLP data for each of the four diseases as well.

Methods

- Physician reviewed charts are the gold standard.
- Parameter selection and weight minimization using lasso.
- Five-fold cross validation.
- We reworked the “fact” normalization to use the more ubiquitous ICD-9 code count.

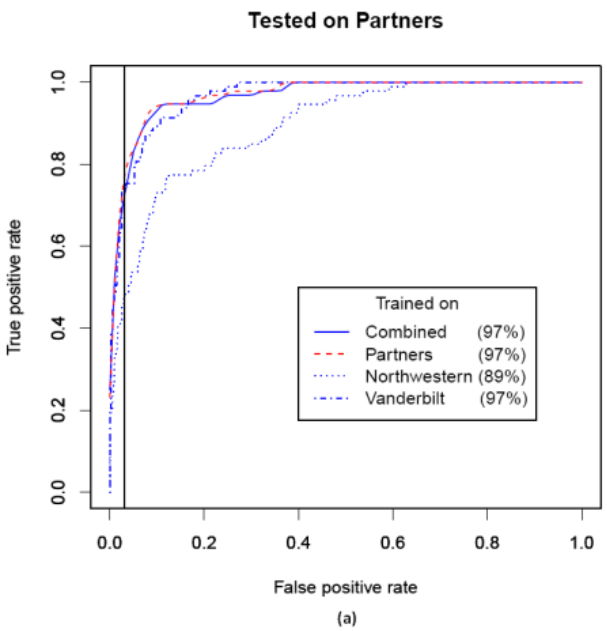
Patient Demographics

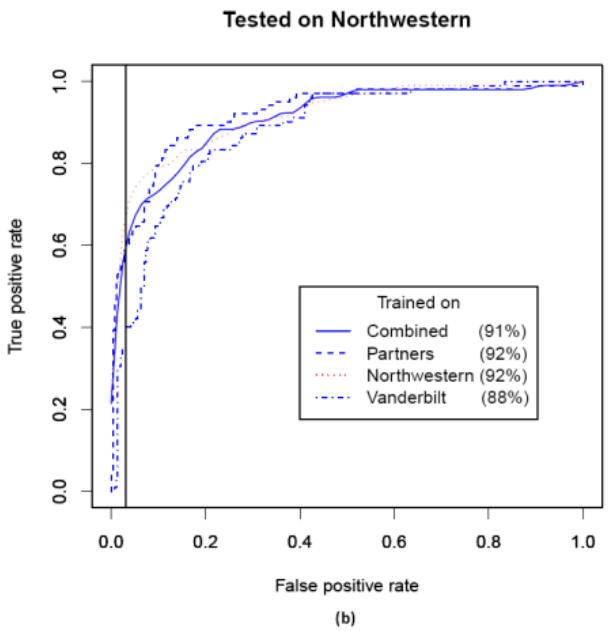
	Partners (n=500)		**Northwestern** (n=390)		**Vanderbilt** (n=376)	
	RA	**Non-RA**	**RA**	**Non-RA**	**RA**	**Non-RA**
Total	96 (19.2%)	404 (80.8%)	102 (26.2%)	288 (73.8%)	185 (49.2%)	191 (50.8%)
Age	60.7 ± 15.9	56.0 ± 18.6	54.3 ± 14.8	58.9 ± 16.8	52.9 ± 13.1	56.2 ± 16.5
Female	74 (77.1%)	303 (75.0%)	83 (81.4%)	209 (72.6%)	148 (80.0%)	141 (73.8%)
Ethnicity						
Caucasian	64 (66.7%)	286 (70.8%)	40 (39.2%)	120 (41.7%)	143 (77.3%)	155 (81.2%)
African American	3 (3.1%)	46 (11.4%)	18 (17.6%)	46 (16.0%)	14 (7.6%)	26 (13.6%)
Hispanic	2 (2.1%)	29 (7.2%)	6 (5.9%)	18 (6.3%)	1 (0.5%)	1 (0.5%)
Other	6 (6.3%)	7 (1.7%)	13 (12.7%)	44 (15.3%)	3 (1.6%)	2 (1.0%)
Unknown	21 (21.9%)	36 (8.9%)	25 (24.5%)	60 (20.8%)	24 (13.0%)	7 (3.7%)
Drugs						
Anti-TNF use	50 (52.1%)	50 (12.4%)	67 (65.7%)	37 (12.8%)	88 (47.6%)	26 (13.6%)
MTX	77 (80.2%)	105 (26.0%)	70 (68.6%)	61 (21.2%)	133 (71.9%)	63 (33.0%)
Codes						
RA	93 (96.9%)	329 (81.4%)	102 (100.0%)	283 (98.3%)	185 (100.0%)	191 (100.0%)
SLE	2 (2.1%)	37 (9.2%)	3 (2.9%)	22 (7.6%)	14 (7.6%)	32 (16.8%)
JRA	7 (7.3%)	28 (6.9%)	1 (1.0%)	18 (6.3%)	6 (3.2%)	8 (4.2%)
PsA	2 (2.1%)	21 (5.2%)	0 (0.0%)	12 (4.2%)	6 (3.2%)	14 (7.3%)
EHR Followup*	9.38 ± 6.77	10.14 ± 6.85	6.30 ± 4.69	6.05 ± 4.85	9.97 ± 4.06	9.06 ± 4.32

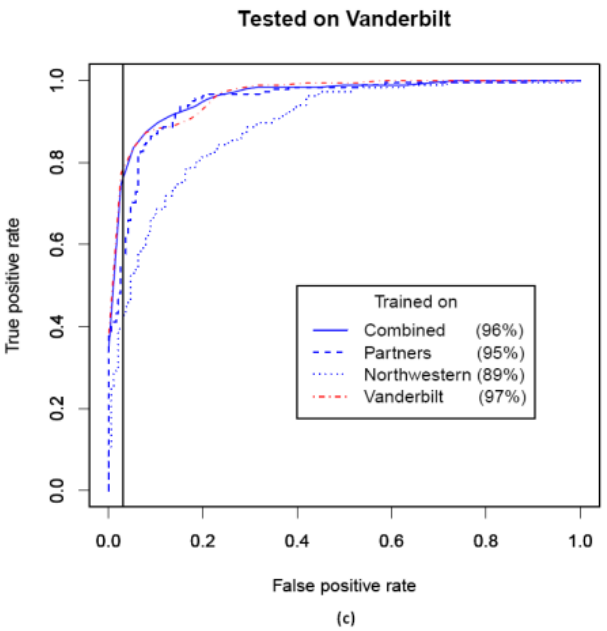
Results

	Testing Set											
	Partners			Northwestern			Vanderbilt			Average		
Algorithm	PPV	Sens	AUC	PPV	Sens	AUC	PPV	Sens	AUC	PPV	Sens	AUC
Published Algorithm	88%*	79%*	97%*	87%	60%	92%	95%	57%	95%	90%	65%	95%
Retrained with:												
Northwestern	79%	47%	89%	87%	73%	92%	93%	43%	89%	86%	54%	90%
Vanderbilt	85%	74%	97%	82%	40%	88%	97%	81%	97%	88%	65%	94%
Combined	86%	71%	97%	86%	65%	91%	97%	82%	96%	90%	72%	95%
ICD-9 Only:¥												
>1 RA code	22%	97%	N/A	26%	100%	N/A	49%	100%	N/A	33%	99%	N/A
>3 RA code	55%	81%	N/A	42%	87%	N/A	73%	98%	N/A	57%	89%	N/A
>Optimal	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Optimal Code Count		53			29			48			43.3	

ROC Curves


Tested on Partners
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Trained on
Combined (97%)
Partners (97%)
Northwestern (89%)
Vanderbilt (97%)
(a)


Tested on Northwestern
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Trained on
Combined (91%)
Partners (92%)
Northwestern (92%)
Vanderbilt (88%)
(b)


Tested on Vanderbilt
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Trained on
Combined (96%)
Partners (95%)
Northwestern (89%)
Vanderbilt (97%)
(c)

Discussion

- These results show that a previously published logistic regression method developed at one institution is portable.
- The published logistic regression model improved sensitivity by 22% and PPV by 6% compared to the optimal ICD-9 count threshold, demonstrating the added value of the more complex phenotyping algorithm.
- Northwestern has a shorter follow-up time for their patients.

Discussion- Normalization and Portability

- The original algorithm normalized an attribute by the number of "facts".
- Every piece of data available was a fact- from icd-9 codes, to lab results, and NLP results.
- This is hard to replicate, but normalizing by the record size is important.
- We selected a new proxy for record size, the number of ICD-9 codes.
- Using linear regression, we were able to approximate the number of facts and still apply the originally trained model.

Discussion- Betas

Attribute	Description	Original	Retrained		
			Combined	Vanderbilt	Northwestern
(Intercept)	Regression Intercept	-5.2088	-4.00186	-4.75161	-2.49521
age	Age of the patient	**-0.00096**	**-0.00426**	**0.010474**	**-0.00769**
sex	Binary: Female is 1	-0.10729	-0.15874	-0.10372	0
ICD-9 RA	Number of encounters with the specified billing code. Defined as sets of ICD-9s >7 days apart. Natural log transformed	0.639367	0.786732	1.036392	0.234517
ICD-9 PsA		0	-0.44454	0	-0.78851
ICD-9 SLE		-0.95937	-0.36747	-0.12847	-0.37016
ICD-9 JRA		-2.25118	-0.67657	-0.49168	0
Normalized ICD-9 RA	ICD-9 RA before transformation, normalized by the total ICD-9 counts.	**66.02406***	91.33932	123.3725	18.72591
Methotrexate	Binary variable. Denotes exposure to this medication from codified sources.	0	0	0	0.541961
Anti-TNF		0.958811	0.745813	0	0.818504
Other Medications		0	0	0	0.147411
RF Negative	Binary variable for the Rheumatoid Factor test.	0.850944	-0.42402	-0.32315	-0.58007
RF Positive		0	0.748071	0.795551	0
NLP RA	Natural log transformed count of the number of notes with the specified concept.	0.969956	0.733087	0.645841	0.914909
NLP SLE		-0.52562	-0.21839	-0.09926	0
NLP JRA		0	-1.00356	-1.27468	0
NLP PsA		-0.85581	-0.05785	0	0
Methotrexate	Binary variable. Denotes exposure to this medication from narrative sources.	0.631764	0.442066	0	0
Anti-TNF		0.520743	0.321728	0	0.908849
Other Medications		0.298111	0.479028	0	0.020784
Cyclic citrullinated Peptide	Binary variable. Denotes positive mention of this test in narrative sources.	1.312583	0.701096	0	0
NLP RF	Binary variable. Denotes positive mention of this test in narrative sources.	0	-0.28693	0	1.279047
Seropositive	Binary variable. Denotes positive mention of this term in narrative sources.	2.773642	1.04373	0	0.16429
Erosions	Binary variable. Denotes positive mention of this finding in narrative sources.	1.259249	0.540583	0.464227	0

- The normalized RA count weight was very big, but the value was also very small
- Weights are very different, and sometimes opposite, but the model still predicted well.
- Vanderbilt has no medication attributes that were selected by the model.

Partners algorithm: Lots of regular expressions!

nlpRF	RFPositive	(?i)(?m)(?s)(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(\+\s*|pos(itive|itivity)?\s+|elevated\s+|increased\s+)((test\s+)?for\s+)?(IgG\s+)?((anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\s+and\s+)?(\bRF|rheumatoid\s+factor(s)?)
nlpRF	RFPositive	(?i)(?m)(?s)(\bRF|rheumatoid\s+factor(s)?)(\s+and\s+(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*- ?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP))?\s+(is\s+|was\s+|are\s+|has\s+(also\s+)?been\s+)?(also\s+)?(found\s+)?(to\s+ be\s+)?(also\s+)?(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(present\b|\+|positive|positivity|pos\b|elevated|increased|greater\s+than\s+\d{2,})
nlpRF	RFPositive	(?i)(?m)(?s)(\bRF|rheumatoid\s+factor(s)?)(\s+and\s+(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*- ?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP))?\s*\+
nlpRF	RFPositive	(?i)(?m)(?s)(?<!(no\s{1,100}))(presence\s+of\s+circulating\s+|presence\s+of\s+|circulating\s+)(IgG\s+)?((anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\s+and\s+)?(\bRF|rheumatoid\s+factor(s)?)
nlpRF	RFStandalone	(?i)(?m)(?s)(\bRF\b|\brheumatoid\s+factor\b)
nlpccp	CCPStandalone	(?i)(?m)(?s)\b(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)\b
nlpccp	CCPPositive	(?i)(?m)(?s)(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(\+\s*|pos(itive|itivity)?\s+|elevated\s+|increased\s+)((test\s+)?for\s+)?(IgG\s+)?((\bRF|rheumatoid\s+factor(s)?)\s+and\s+)?(anti\s*- ?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*-?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)
nlpccp	CCPPositive	(?i)(?m)(?s)(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)(\s+and\s+(\bRF|rheumatoid\s+factor(s)?))?\s+(is\s+|was\s+|are\s+|has\s+(also \s+)?been\s+)?(also\s+)?(found\s+)?(to\s+be\s+)?(also\s+)?(borderline\s+|slightly\s+|strongly\s+|low\s+|weakly\s+|low\s*-?\s*titer\s+|high\s*- ?\s*titer\s+)?(present\b|\+|positive|positivity|pos\b|elevated|increased|greater\s+than\s+\d{2,})
nlpccp	CCPPositive	(?i)(?m)(?s)(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*-?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*- \s*)?CCP|anti\s*-?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*- ?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)(\s+and\s+(\bRF|rheumatoid\s+factor(s)?))?\s*\+
nlpccp	CCPPositive	(?i)(?m)(?s)(?<!(no\s{1,100}))(presence\s+of\s+circulating\s+|presence\s+of\s+|circulating\s+)(IgG\s+)?(anti\s*-?\s*CCP\s+antibod(ies|y)|anti\s*- ?\s*CCP|CCP\s+antibod(ies|y)|(anti\s*-?\s*)?cyclic\s+citrullinated\s+peptide(\s+antibod(ies|y))?|a(\s*-\s*)?CCP|anti\s*- ?\s*citrullinated\s+protein\s*/?\s*peptide\s+antibod(ies|y)|anti\s*-?\s*citrullinated\s+(protein|peptide)\s+antibod(ies|y)|ACPA|ACP\s+antibod(ies|y)|CCP)

Leveraging the UMLS

Attribute	CUI	Name
RA	C0003873	rheumatoid arthritis (diagnosis)
	C0019773	Felty's Syndrome [Disease/Finding]
	C0035450	Rheumatoid nodule (morphologic abnormality)
	C0263741	Extra-articular rheumatoid process (disorder)
	C0409628	Other rh.arthr.+visc/syst.dis.
	C0409629	Rheumatoid arthritis of interphalangeal joint of toe (disorder)
	C0409630	Rheumatoid arthritis of lesser metatarsophalangeal joint (disorder)
	C0409631	Rheumatoid arthritis of first metatarsophalangeal joint (disorder)
	C0409632	Rheumatoid arthr-oth tarsal jt
	C0409633	Rheumatoid arthritis of talonavicular joint (disorder)
	C0409634	Rheumatoid arthritis of subtalar joint (disorder)
	C0409635	Rheumatoid arthritis of ankle (disorder)
	C0409636	Rheumatoid arthritis of tibiofibular joint (disorder)
	C0409637	rheumatoid arthritis of knee (diagnosis)
	C0409638	Rheumatoid arthritis of sacroiliac joint (disorder)
	C0409639	Rheumatoid arthritis of hip
	C0409640	Rheumatoid arthritis of distal interphalangeal joint of finger (disorder)
	C0409641	Rheumatoid arthritis of proximal interphalangeal joint of finger (disorder)
	C0409642	Rheumatoid arthritis of metacarpophalangeal joint (disorder)
	C0409643	rheumatoid arthritis of wrist (diagnosis)
	C0409644	Rheumatoid arthritis of distal radioulnar joint (disorder)
	C0409645	Rheumatoid arthritis of elbow
	C0409646	Rheumatoid arthritis of acromioclavicular joint (disorder)
	C0409647	Rheumatoid arthritis of sternoclavicular joint (disorder)
	C0409648	Rheumatoid arthritis of shoulder
	C0409649	Oth rheumatoid arthritis-spine
	C0409650	Rheumatoid arthritis of cervical spine (disorder)
	C0409651	Seropositive rheumatoid arthritis, unspecified
	C0409652	ARTHRITIS RHEUMATOID SERONEGATIVE
	C0409653	Rheumatoid arthritis with organ / system involvement (disorder)
	C0409657	Rheumatoid arthritis with multisystem involvement (disorder)
	C0477542	Other specified rheumatoid arthritis
	C0564784	Rheumatoid arthritis of multiple joints (disorder)
	C0564785	Rheumatoid arthritis of hand joint (disorder)
	C0564786	Rheumatoid arthritis of ankle and/or foot (disorder)
	C0564787	Rheumatoid arthr - other joint
	C0581345	Flare of rheumatoid arthritis (disorder)
	C0837507	Felty's syndrome, multiple sites
	C0837508	Felty's syndrome, shoulder region
	C0837509	Felty's syndrome, upper arm
	C0837510	Felty's syndrome, forearm
	C0837511	Felty's syndrome, hand
	C0837512	Felty's syndrome, pelvic region and thigh
	C0837513	Felty's syndrome, lower leg
	C0837514	Felty's syndrome, ankle and foot
	C0837515	Felty's syndrome, other site
	C0837516	Felty's syndrome, site unspecified
	C0837597	Rheumatoid nodule, multiple sites
	C0837598	Rheumatoid nodule, shoulder region
	C0837599	Rheumatoid nodule, upper arm
	C0837600	Rheumatoid nodule, forearm
	C0837601	Rheumatoid nodule, hand
	C0837602	Rheumatoid nodule, pelvic region and thigh
	C0837603	Rheumatoid nodule, lower leg
	C0837604	Rheumatoid nodule, ankle and foot
	C0837605	Rheumatoid nodule, other site
	C0837606	Rheumatoid nodule, site unspecified
	C0837627	Rheumatoid arthritis, unspecified, multiple sites
	C0837628	Rheumatoid arthritis, unspecified, shoulder region
	C0837629	Rheumatoid arthritis, unspecified, upper arm
	C0837630	Rheumatoid arthritis, unspecified, forearm
	C0837631	Rheumatoid arthritis, unspecified, hand
	C0837632	Rheumatoid arthritis, unspecified, pelvic region and thigh
	C0837633	Rheumatoid arthritis, unspecified, lower leg
	C0837634	Rheumatoid arthritis, unspecified, ankle and foot
	C0837635	Rheumatoid arthritis, unspecified, other site
	C1304214	Rheumatoid nodulosis (disorder)

SLE	C0024137	Cutaneous lupus erythematosus (disorder)
	C0024138	LUPUS ERYTHEMATOSUS DISCOID
	C0024140	Subacute cutaneous lupus erythematosus (diagnosis)
	C0024141	LUPUS ERYTHEMATOSUS SYSTEMIC
	C0024143	Systemic lupus erythematosus glomerulonephritis syndrome (disorder)
	C0024145	Sarcoidosis, lupus pernio type (disorder)
	C0030327	Lupus erythematosus profundus (disorder)
	C0155180	Discoid lupus erythematosus of eyelid (disorder)
	C0242380	Nonbacterial verrucal endocarditisis (disorder)
	C0263591	Drug-induced lupus erythematosus (disorder)
	C0263592	Drug-induced lupus erythematosus due to procainamide (disorder)
	C0263593	Drug-induced lupus erythematosus due to hydralazine (disorder)
	C0263594	Drug-induced lupus erythematosus due to diphenylhydantoin (disorder)
	C0264514	Lupus disease of the lung (disorder)
	C0267807	Lupus hepatitis (disorder)
	C0268754	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class I (disorder)
	C0268755	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class II (disorder)
	C0268756	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class III (disorder)
	C0268757	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class IV (disorder)
	C0268758	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class V (disorder)
	C0268759	Systemic lupus erythematosus glomerulonephritis syndrome, World Health Organization (WHO) class VI (disorder)
	C0339908	Lung disease with systemic lupus erythematosus (disorder)
	C0393968	Systemic lupus erythematosus encephalitis (disorder)
	C0406633	Lupus erythematosus chronicus (disorder)
	C0406634	Lupus erythematosus migrans (disorder)
	C0406635	Lupus erythematosus nodularis (disorder)
	C0406636	Lupus erythematosus tumidus (disorder)
	C0406637	Erythema multiforme-like lupus erythematosus (disorder)
	C0406638	Lupus erythematosus unguium mutilans (disorder)
	C0409974	Lupus erythematosus
	C0409975	Limited lupus erythematosus (disorder)
	C0409976	Systemic lupus erythematosus with organ/system involvement
	C0409977	Bullous systemic lupus erythematosus (disorder)
	C0409978	Systemic lupus erythematosus with multisystem involvement (disorder)
	C0409979	Neonatal lupus erythematosus (disorder)
	C0477525	Other local lupus erythematosus
	C0477587	Other forms of systemic lupus erythematosus
	C0521471	Systemic lupus erythematosus rash
	C0521513	Systemic lupus erythematosus arthritis (disorder)
	C0542297	LE SYNDROME AGGRAVATED
	C0543635	localized discoid lupus erythematosus (diagnosis)
	C0558705	Lupus erythematosus associated with renal failure
	C0587239	Systemic lupus erythematosus with pericarditis
	C0740415	Lupus encephalitis
	C0752332	Lupus Vasculitis, Central Nervous System [Disease/Finding]
	C1274831	Acute systemic lupus erythematosus (disorder)
	C1274833	Fulminating systemic lupus erythematosus (disorder)
	C1274834	Systemic lupus erythematosus of childhood (disorder)
	C1274836	Subacute cutaneous lupus erythematosus, papulosquamous type (disorder)
	C1274838	Hypertrophic type discoid lupus erythematosus (disorder)
	C1274839	Rosaceous type discoid lupus erythematosus (disorder)
	C1274840	Discoid lupus erythematosus of mucous membranes (disorder)
	C1274841	Discoid lupus erythematosus of oral mucosa (disorder)
	C1274842	Discoid lupus erythematosus of genital mucous membranes (disorder)
	C1274843	Discoid lupus erythematosus of scalp (disorder)
	C1274844	Discoid lupus erythematosus of face (disorder)
	C1274845	Discoid lupus erythematosus of lip (disorder)
	C1274846	Discoid lupus erythematosus of hands (disorder)
	C1274847	Discoid lupus erythematosus of foot (disorder)
	C1274848	Disseminated discoid lupus erythematosus (disorder)
	C1274858	Lupus erythematosus-associated nail dystrophy (disorder)
	C1384022	Systemic lupus erythematosus-related syndrome (disorder)

JRA	C0157916	acute polyarticular juvenile rheumatoid arthritis (diagnosis)
	C0157917	Pauciarticular juvenile rheumatoid arthritis (disorder)
	C0157918	Monarticular juvenile rheumatoid arthritis (disorder)
	C0263739	Chronic polyarticular juvenile rheumatoid arthritis (disorder)
	C0811221	polyarticular juvenile rheumatoid arthritis (diagnosis)
	C0409625	Juvenile reactive arthritis
	C0409667	Juvenile Chronic Polyarthritis
	C0409671	Juvenile rheumatoid arthr.unsp
	C0409678	Juvenile arthritis of inflammatory bowel disease (disorder)
	C0477545	Other juvenile arthritis
	C0553662	Chronic Childhood Arthritis
	C0837691	Juvenile rheumatoid arthritis, multiple sites
	C0837692	Juvenile rheumatoid arthritis, shoulder region
	C0837693	Juvenile rheumatoid arthritis, upper arm
	C0837694	Juvenile rheumatoid arthritis, forearm
	C0837695	Juvenile rheumatoid arthritis, hand
	C0837696	Juvenile rheumatoid arthritis, pelvic region and thigh
	C0837697	Juvenile rheumatoid arthritis, lower leg
	C0837698	Juvenile rheumatoid arthritis, ankle and foot
	C0837699	Juvenile rheumatoid arthritis, other site
	C0837700	Juvenile rheumatoid arthritis, site unspecified
	C0837741	Juvenile arthritis, unspecified, multiple sites
	C0837742	Juvenile arthritis, unspecified, shoulder region
	C0837743	Juvenile arthritis, unspecified, upper arm
	C0837744	Juvenile arthritis, unspecified, forearm
	C0837745	Juvenile arthritis, unspecified, hand
	C0837746	Juvenile arthritis, unspecified, pelvic region and thigh
	C0837747	Juvenile arthritis, unspecified, lower leg
	C0837748	Juvenile arthritis, unspecified, ankle and foot
	C0837749	Juvenile arthritis, unspecified, other site
	C0837750	Juvenile arthritis, unspecified, site unspecified
	C1384600	Systemic onset juvenile chronic arthritis (disorder)
	C1444840	Juvenile seronegative polyarthritis (disorder)
	C1444841	Juvenile idiopathic arthritis, oligoarthritis (disorder)
	C1444844	Juvenile idiopathic arthritis, enthesitis related arthritis (disorder)
	C1444845	Juvenile idiopathic arthritis, undifferentiated arthritis (disorder)
PsA	C0003872	Arthritis,psoriatic
	C0343176	Psoriatic arthritis with spine involvement (disorder)
	C0409672	Juvenile psoriatic arthritis (disorder)
	C0409682	Psoriatic arthritis with distal interphalangeal joint involvement
	C0409883	Psoriatic dactylitis (disorder)
	C0477543	Other psoriatic arthropathies
	C1444809	Iritis in psoriatic arthritis (disorder)
RF	C0035448	Rheumatoid factor (substance)
	C0201860	RF
	C0201861	Qualitative Rheumatoid Factor Test
	C0201862	Rheumatoid factor, quantitative (procedure)
	C0584621	Serum rheumatoid antigen measurement (procedure)
	C1254833	RF Titer Test
	C1256157	Quant Rheumatoid Factor Test
	C1272655	Fluid rheumatoid factor measurement (procedure)
	C1279454	Rheumatoid factor screening test
	C1295062	IgM rheumatoid factor assay (procedure)
	C1532421	Rheumatoid factor IgG measurement (procedure)
	C1532542	IgA rheumatoid factor measurement (procedure)
	C1972676	Rheumatoid factor	bid-ser-plas
	C1972677	Rheumatoid factor	cerebral spinal fluid
	C1972678	Rheumatoid factor	body fluid
	C1972679	Rheumatoid factor	synovial fluid
	C1972680	Rheumatoid factor IgA	bid-ser-plas
	C1972681	Rheumatoid factor IgG	bid-ser-plas
	C1972682	Rheumatoid factor IgM	bid-ser-plas
	C1591264	Rheumatoid factor	Pleural fluid
+R	C0409651	Seropositive rheumatoid arthritis, unspecified
	C0477541	Other seropositive rheumatoid arthritis
	C0585962	Seropositive erosive rheumatoid arthritis
	C0837548	Other seropositive rheumatoid arthritis, shoulder region
	C0837549	Other seropositive rheumatoid arthritis, upper arm
	C0837550	Other seropositive rheumatoid arthritis, forearm
	C0837551	Other seropositive rheumatoid arthritis, hand
	C0837552	Other seropositive rheumatoid arthritis, pelvic region and thigh
	C0837553	Other seropositive rheumatoid arthritis, lower leg
	C0837554	Other seropositive rheumatoid arthritis, ankle and foot
	C0837555	Other seropositive rheumatoid arthritis, other site
	C0837556	Other seropositive rheumatoid arthritis, site unspecified
	C0837557	Seropositive rheumatoid arthritis, unspecified, multiple sites
	C0837558	Seropositive rheumatoid arthritis, unspecified, shoulder region
	C0837559	Seropositive rheumatoid arthritis, unspecified, upper arm
	C0837560	Seropositive rheumatoid arthritis, unspecified, forearm
	C0837561	Seropositive rheumatoid arthritis, unspecified, hand
	C0837562	Seropositive rheumatoid arthritis, unspecified, pelvic region and thigh
	C0837563	Seropositive rheumatoid arthritis, unspecified, lower leg
	C0837564	Seropositive rheumatoid arthritis, unspecified, ankle and foot
	C0837565	Seropositive rheumatoid arthritis, unspecified, other site
	C0837566	Seropositive rheumatoid arthritis, unspecified, site unspecified
Erosions *	C0333307	Superficial ulcer
	C0587240	Erosion of bone
CCP**	C1824602	Anti-Antibodies

Limitations

- Still only one chronic disease well represented in ICD-9s
- May not represent portability of other machine learning methods