Using LASSO Regression to Predict Rheumatoid Arthritis Treatment Efficacy

David J. Odgers MS, Natalie Tellis, Heather Hall, Michel Dumontier PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

ABSTRACT

Rheumatoid arthritis (RA) accounts for one-fifth of the deaths due to arthritis, the leading cause of disability in the United States. Finding effective treatments for managing arthritis symptoms are a major challenge, since the mechanisms of autoimmune disorders are not fully understood and disease presentation differs for each patient. The American College of Rheumatology clinical guidelines for treatment consider the severity of the disease when deciding treatment, but do not include any prediction of drug efficacy.

Using Electronic Health Records and Biomedical Linked Open Data (LOD), we demonstrate a method to classify patient outcomes using LASSO penalized regression. We show how Linked Data improves prediction and provides insight into how drug treatment regimes have different treatment outcome. Applying classifiers like this to decision support in clinical applications could decrease time to successful disease management, lessening a physical and financial burden on patients individually and the healthcare system as a whole.

INTRODUCTION

Rheumatoid arthritis: a poorly understood challenge for biologists and clinicians.

Arthritis is a broad term for more than a hundred conditions that affect the joints and connective tissues. Although disease presentation varies between conditions, all forms of arthritis are characterized by pain and stiffness around the joints. The health-related life quality in arthritis patients is 2-3 times worse than those without arthritis for physical and mental health indicators [1].

RA is an autoimmune form of the disease that presents in both the young and old, and affects 1% of the population [2,3]. For those with RA, arthritis is not only disabling, but also shortens life expectancy by approximately 10 years [2,4,5] and accounts for one-fifth of the deaths due to arthritis and rheumatic conditions [2,3]. Comorbidities such as infections, cardiac disease, and mental health issues also confound diagnosis, making treatment more challenging. These confounders along with variation in disease presentation and treatment-response preclude understanding of disease mechanisms and hinder progress that would allow effective treatment of patients.

Rapid diagnosis and management is important for reducing joint damage and the risk of premature death in patients with RA. Current guidelines require patients to try multiple medications based on their disease severity and progression, instead of considering patient-specific factors that predict their response to treatments [6]. With the available guidelines and treatment options, medications for rheumatoid arthritis patients are only 50% effective [7]. These inadequate guidelines reduce the quality of care for these patients and increase their risk for premature death. While efforts have been made to improve clinical guidelines, these efforts are preliminary and have only the parameters necessary for effective clinical guidelines [8]. We hypothesize that by using a combination of linked open data to describe underlying biological patterns of clinical variables found in the patient EHR data, we can 1) create a baseline classifier that shows improvement over non data driven RA outcome prediction 2) show classifier improvement over the baseline by adding linked data patterns as extended variables to LASSO regression models.

Coupling electronic health record data with public biomedical data for broad computational characterization.

Patient stratification methods [10,11,12,13] can help understand patient treatment profiles using a combination of semantic web enabled feature extraction and engineering techniques along with machine learning classification [14,15,16]. EHR data can be connected to semantic web resources such as Bio2RDF [17,18], an open-source, semantic web repository of life science data on the internet, containing ~11 billion triples across 35 datasets. STRIDE is a clinical data warehouse that contains EHR data [9,18,19] from the Lucile Packard Children's Hospital and Stanford Hospital and

Clinics. In this study, we examine the utility of incorporating clinical features from the EHR with public biomedical data to undertake a large scale data-mining approach for understanding patient outcomes on different drug regimens. This work provides an initial approach for patient classification and personalized clinical guidelines, which will have immediately translatable effects on personalized health trajectories for rheumatoid arthritis patients. It will be possible to extend this clinical outcomes modeling [21] to numerous other diseases that currently have a limited understanding of the underlying disease features.

METHODS

Disease Features: We developed a list of relevant International Classification of Disease, Ninth Revision (ICD9) codes related to "rheumatoid arthritis" to identify an initial RA cohort (Table 1).

Table 1: ICD9 codes selected to build RA cohort with a description of the code.

ICD9	Description
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
714.3	Juvenile chronic polyarthritis
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
714.4	Chronic postrheumatic arthropathy
714.8	RA Related
714.81	Rheumatoid lung
714.89	Other specified inflammatory polyarthropathies
714.9	Unspecified inflammatory polyarthropathy

Drug Features: We considered two different drug classes - Corticosteroids and Disease Modifying Antirheumatic drugs (DMARDs) - in this study. We chose them for their widely accepted therapeutic use for RA patients as well as the prevalence of use within our treatment cohorts. Table 2 provides a full list of all the medications used with their drug classification, compiled from literature, and verified using Healthline [22].

Table 2: Drug classes and drug names for the treatment cohorts within the RA cohort.

Class	Drug
DMARD	Sulfasalazine, Penicillamine, Minocycline, Methotrexate, Leflunomide, Hydroxychloroquine, Cyclosporine, Cyclophosphamide, Azathioprine, Auranofin
Corticosteroids	Prednisone, Prednisolone, Methylprednisolone, Hydrocortisone, Dexamethasone , Cortisone

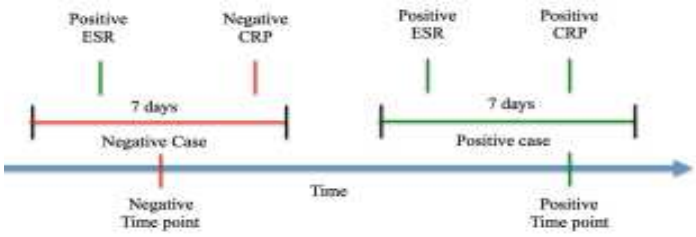
Lab Features: We chose Erythrocyte sedimentation rate lab (ESR) and C-reactive protein lab (CRP) as inflammation indicators for a clinical endpoint based on a recommendations from rheumatologists at Stanford Clinics. These indicators are usually concurrently determined to assess the level of inflammation experienced by the patient. Since each test interrogates the state of different biochemical pathways, a more accurate view of disease status is garnered when these indicators are concordant. If the values of each indicator are within normal range, the patient is assumed to be treated while if either of the test results are abnormal, the patient is considered to be in an inflamed state. Other indicators such as Rheumatoid Factor were available in the patient record but had too few occurrences to be considered for a standard measure for all patients in the RA cohort.

Building the rheumatoid arthritis patient cohort.

We constructed a RA patient cohort in which patients must have at least one of the ICD9 codes in Table 1 plus results for at least one ESR or CRP measurement. The RA cohort was further pruned to those who have been prescribed at least one corticosteroid or DMARD listed in Table 2. We used INTREPID, a Semantic Web database that combines an RDF version of STRIDE with RDF data from Bio2RDF [23], to obtain features for each patient in the RA cohort. We then used a sliding window to extract patient features (comorbidities and prescriptions).

Labs were used to establish ground truth with regards to the inflammation that the patient was experiencing at a given point in time. We used these time points to identify observation time points for the study. CRP and ESR lab values were assessed to be normal or high based on the accepted standard cutoff measurements: 10 mg/dL for CRP tests and 25 mm/hr for the ESR test. We only considered cases that had at least one ESR and one CRP test administered within one week of one another. If both values were below the threshold, then we labeled the observation time point a positive test case at the time point in which the second time point showed a positive result, clinically this indicates that the patient is exhibiting a properly treated state. If either one of the values were higher than the threshold, we labeled the observation time point a negative control and the time point was averaged between the two tests as demonstrated in Figure 1, this indicates that the patient is in an improperly treated state.

Figure 1: Windowing procedure for lab ESR and CRP values to establish the observation period.



Comorbidities were extracted 365 days back from the observation time point for both positive and negative cases. The comorbidities that fell within the observation window were comorbidity features.

Prescriptions were considered that fell within 365 days back from the observation time point, similar to the comorbidities. Once the prescriptions were evaluated to be temporally connected to the observation time point, the drugs of class Corticosteroid and DMARDs were identified within the observation window and the observation was labeled either a Corticosteroid, DMARDs or Corticosteroids/DMARDS observation, based on which drug classes were present in the observation window. We considered route of administration information to limit the drugs to oral, topical, inhalation, rectal, transdermal, nebulization, sublingual, nasal, buccal, swish & swallow, vaginal, inhaled (neb), inhaled (oral), inhaled (oral) w/spacer, transmucosal, mucous membrane, mouth/throat, joint, oral/nasogastric, both nostrils, translingual, intravenous, subcutaneous, unmapped, intramuscular and injection. We selected these routes to represent either the route of administration for the drugs in our treatment cohorts or drug routes that could be administered at home. We used the start and stop codes in Table 3 to verify that the patient was currently taking the medication over the particular interval. We used an additional parameter for time to efficacy (set at 15 days) and time to complete elimination of substance from body (set to 15 days) in the event that the patient had either started taking the drug less than 15 days prior to the observation period or had stopped taking the drug more than 15 days before the observation period ended. We aggregated the drugs to the RxNORM unique identifier for the active ingredient to normalize the drug names. We then used co-prescriptions to populate a matrix for co-prescription features.

Table 3: Start and Stop codes applied to prescriptions based on status.

Code	Prescription Status
START	New order, Sent, Verified, Order code status change, Dispensed, Extra Doses, Fill List Generation, Change order request
STOP	Discontinue order request, Discontinued, Canceled, Completed, Suspend, Deleted Order/Retracted Order

Linking out to external data sources.

We used the Online Mendelian Inheritance in Man knowledge base (OMIM) [24], Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] and DrugBank (Drugbank) that were integrated into INTREPID from Bio2RDF sources [26] to link comorbidities and co-prescriptions to inherited genetic disease phenotypes, drug activity and biochemical pathways through synonymous concepts found in the Unified Medical Language System (UMLS) as shown in Table 4. For example, we know from OMIM that a disease phenotype can have several inherited disease genes. Therefore, if a patient has a comorbidity that is inherited, we can infer that the genes that have been identified for that particular disease are also genes that are part of an aberrant pathway within the patient through KEGG. We now consider these genes and pathways as features, making further inference about the biochemical pathways that the gene affects. We used KEGG to find interaction information that describes molecular actions that take place when the medication is administered, metabolism information to describe physiologic systems that are affected when medication is metabolized, drug mechanism of action and pathway perturbation information. We used Drugbank to identify specific genes and enzymes that are targeted by medication.

Table 4: Features extracted from LOD resources in Bio2RDF.

Resource	Features in Resource	Connection to STRIDE	# of Extended Features
Bio2RDF: OMIM	Diseased genes	ICD9 Comorbidity	4959
Bio2RDF: KEGG	Pathway perturbation	ICD9 Comorbidity	52
Bio2RDF: KEGG	Pathway activity	RxNORM Co-prescription	452
Bio2RDF: KEGG	Pathway Interaction	RxNORM Co-prescription	79
Bio2RDF: KEGG	Drug Metabolism	RxNORM Co-prescription	96
Bio2RDF: KEGG	Drug Mechanism of Action	RxNORM Co-prescription	392
Bio2RDF: KEGG	Pathway perturbation	RxNORM Co-prescription	84
Bio2RDF: DrugBank	Gene Target	RxNORM Co-prescription	565
Bio2RDF: DrugBank	Enzyme Target	RxNORM Co-prescription	116

Patient Stratification and LASSO regression.

The laboratory measurements outcomes indicating a positive or negative observation time point were combined with treatment cohort information to assign each observation into the correct observation classification. The observations were assigned to a treatment group if they were actively prescribed the class or classes of drug at the observation time point as shown in Table 5. We limited drug classes to those associated associated with at least 100 patients, this eliminated drug classes for Biologics and NSAIDS. Additionally, we imposed a constraint to verify that each patient had at least one observation.

Table 5: Class distribution for treatment cohorts.

Treatment Cohort	Patients	Observations	% Positive Cases	% Negative Cases
Corticosteroids	143	297	30	70
Corticosteroids, DMARDs	127	297	33	67
DMARDs	113	235	44	56

LASSO Regression models [27,28,29] for all treatment cohorts were constructed using R programming language with the glmnet package. We used a binomial distribution response variable in the LASSO classification due to the binary output variable in the processed data and used the minimized lambda value for each model to optimize the model for cross validation. We constructed feature matrices that contained only the clinical variables and those feature matrices containing clinical and features from external sources. The models were built to determine which patient subgroups responded best to which RA treatment based on their health record data. LASSO regression was chosen because of its ability to shrink coefficients of variables that do not contribute information to the model to zero, eliminating the need to do features selection on high dimensional data. This approach was seen as superior given the number of variables in the model.

RESULTS

Description of rheumatoid arthritis cohort.

The RA cohort showed no significant difference in race, but did show a significant enrichment for females (81% vs 29%, p-value < <0.05). This finding is validating given that rheumatoid arthritis has previously been found to be more prevalent in females [2]. Almost all patients were documented with more than one comorbidity and more than half of patients had over fifty comorbidities in their record. We also found that approximately 50% of patients had at least 2 prescriptions and 25% had at least 6 prescriptions.

LASSO Regression.

The models were tested with 10 fold cross validation to generate performance metrics. The cross validation calculates the number of true positives and negatives as well as false positives and negatives for each hold out set of 10% of the data while training on 90% of the data, ten times over, until all data has been classified. A between group comparison was performed between classifiers of the only clinical features and those containing externally derived features referred to as extended features as shown in Table 6 to demonstrate the utility of adding engineered features from INTREPID to STRIDE clinical features. The AUC values for all cases were between .73 and .79, with significant (p-value < .05) improvements in classifier performance demonstrated in bot Corticosteroid and DMARD groups.

Table 6: Performance statistics for LASSO Regression classifiers for each cohort. AUC is the area under a Receiver Operator Curve, Test Error was generated from 10 fold cross validation, McNemar's p-value is a measure of significant classifier improvement.

Treatment - Variable Set	AUC	Test Error	Sensitivity	Specificity	PPV	NPV	McNemar's p-value
Corticosteroids - Clinical Variables	0.79	0.24	0.38	0.94	0.72	0.78	0.041
Corticosteroids - Extended Variables	0.79	0.21	0.42	0.95	0.77	0.79	0.041
DMARDS - Clinical Variables	0.73	0.37	0.66	0.61	0.56	0.74	0.001
DMARDS - Extended Variables	0.76	0.32	0.67	0.67	0.62	0.77	0.001
Corticosteroids/DMARDs - Clinical Variables	0.79	0.27	0.54	0.83	0.60	0.79	0.074
Corticosteroids/DMARDs - Extended Variables	0.79	0.25	0.44	0.90	0.68	0.77	0.074

DISCUSSION

Value of INTREPID extended features.

Overall, the results significantly improved with the addition of extended features for two of the three treatment groups. The group that did not have significant improvement was near the significance threshold at .074. The boost in PPV performance shows that our extended feature models will predict a positive outcome more precisely than our model with only clinical features will. These methods can be integrated with clinical decision support to aid patients getting on the right treatment more quickly. The improvement in test error indicates that the extended features are adding additional information to the model and not just adding collinear variables. This means the extended features are adding new patterns of biomedical knowledge that underpin the clinical manifestation. It is reasonable to assume that the addition of new externally linked features may provide additional information not captured in the clinical data alone.

INTREPID is designed for knowledge discovery by interrogating Linked Data to find underlying patterns. It is possible that other LOD resources can be interrogated to find even more underlying patterns that are not immediately identifiable from within the patient record itself. The INTREPID Linked Data design can easily integrate new knowledge bases which will provide many more features to interrogate. The three resources, OMIM [24], KEGG [25] and DrugBank [26] that were interrogated in this study grant further access to a large part of the biomedical LOD cloud. The graph structure of the INTREPID graph database will be capable of immediately using any new LOD resources that are uploaded from the Bio2RDF datasets.

Clinical interpretation.

As we achieve relatively high PPV and NPV with all of our classifiers, we are confident that using Clinical Data already available for the patient could use these to rule out treatments in the clinical setting. Additionally, highly predictive features (White ethnicity, Abdominal pain, Abnormal weight gain, Acute bronchitis, Acute gastritis, Acute laryngitis without mention of obstruction, Acute laryngotracheitis without mention of obstruction, Aftercare following organ transplant, Aftercare for healing pathologic fracture of other bone, Aftercare for healing traumatic fracture of lower arm, Allergic rhinitis due to pollen, Aphakia, Arthropathy associated with gastrointestinal conditions other than infections, Benign neoplasm of colon, Benign neoplasm of stomach, Bilateral partial paralysis of vocal cords) can be further assessed clinically to uncover the underlying mechanisms of RA and lead to a more complete treatment profile. Our classifiers can predict outcomes for specific patients based on their health record information. The delay in treatment present in our patient set highlights the challenges of finding an effective treatment, especially with a several weeks to months of latency in drug efficacy for each new course of treatment. By classifying patients by most effective treatment plan for RA, as opposed to using trial and error treatment formulation, these classifiers could shorten time to effective treatment. By predicting drug treatment regime efficacy early on by the patients' background EHR variables through clinical decision support, the patient and physician will have critical information up front to make an informed decision about RA treatment options. This

will result in physical and fiscal burden on the patient, as well as the time cost to a physician of repeated failed disease management.

Classifier performance.

The classifiers demonstrated the performance gained from engineering biochemical features from clinical variables through INTREPID. LASSO is an ideal machine learning algorithm for this process because of how well it handles very high dimensional data. There is effectively no limit to the size of a feature matrix that is input into a LASSO regression, so building LASSO models on the fly without needing to control for collinear variables is ideal for derived features. A more aggressively fitting LASSO model can be used if the feature space is so large that the classifier has difficulty converging.

Previous work by Collier demonstrated the ability to use a non data driven approach to derive clinical features from RA patients EMR and use an activity calculator to predict disease activity at the time of each clinical encounter [31]. The method uses a disease activity score collected by the physicians during a routine physical exam. The physician qualitatively graded the patients' joints, looking for signs of inflammation or tenderness. This prospective design allows physicians to focus their effort on collecting high quality data at the point of care over 12 weeks. Physicians most accurately predicted remission and high disease activity with 88% and 79% accuracy respectively and least accurately predicted moderate and low disease activity at 48% and 62% accuracy respectively. The error rate of this method was between 12% and 52%. This indicates that the test error of our retrospective data driven method between 21% and 34% is similar in mean and has a smaller variance.

CONCLUSION

Expanding feature-sets.

Our results suggest that the addition of other feature sets may improve classifier performance. Other data such as gene function annotations, disease phenotypes, and disease variants may provide additional insight into the clinical pathology. Similarly, the Rheumatology Consultant System Ontology [32] may provide additional knowledge through its taxonomic structure that could prove useful in building more effective classifiers. Finally, additional clinical features extracted from the text mining of clinical notes could provide additional information that are not contained in the structured parts of the EHR [32,33,34].

Clinical applications.

Machine learning, with its intrinsic ability to integrate thousands of signals at the same time, could prove extremely useful in making predictions for otherwise complex clinical situations. For RA patients, who often experience significant financial, emotional, and physical challenges in managing this chronic condition, having more effective treatments is key to a good quality of life. With further research and development, we see the integration of machine learning methods, such as that described here, as a key part of future delivery of health care. Indeed, tools capable of learning from existing EHR data and personalized for the patient at point of care promise better outcomes for the treatment of rheumatoid arthritis patients.

ACKNOWLEDGEMENTS

We gratefully acknowledge the work of Nigam Shah for providing the STRIDE annotating semantic features in the STRIDE clinical notes and providing guidance with data mining STRIDE. Additionally, we would like to thank the Stanford rheumatologists were key to defining clinical phenotypes in our study.

REFERENCES

1. S.E. Furner, et al., *Arth. Care Res*. **63**, 788 (2011)
2. http://www.cdc.gov/arthritis/data_statistics/arthritis_related_stats.htm (2015)
3. K.G. Saag, G.G. Teng, N.M. Patkar, *Arth. and Rheum*. **59**, 762 (2008)
4. T. Pincus, T. Sokka, F. Wolfe, *Arth. and Rheum*. **44**, 1234 (2001)
5. C. Kelly, J. Hamilton., *Rheumatology* **46**, 183 (2007)
6. J.A. Singh, D.E. Furst, A. Bharat, et al., *Arth. Care Res*. **64**, 625 (2012)
7. B.B. Spear, M. Heath-Chiozzi, J. Huff, *Trends Mol. Med*. **7**, 201 (2001)
8. P. Sonali, J.Y. Desai, *Arth. and Rheum*. **63**, 3649 (2011)
9. A.K. Jha, et al., *Health Affairs* **29,** 1951 (2010)
10. F.S. Roque, P.B. Jensen, H. Schmock, et al., *PLoS Comput. Biol*. **7**, epub (2011)
11. J. Pathak, R.C. Keifer, S.J. Bielinski, C.G. Chute, *AMIA Annu Symp Proc.* (2012)
12. J. Pathak, R.C. Keifer, C.G. Chute, *Data Int. Life Sci.* **128** (2013)
13. J. Pathak, *AMIA Jt. Summits Transl. Sci. Proc.* (2013)
14. J. Wu,, J. Roy, W.F. Stewart, *Medical care* **48**, S106 (2010): S106-S113.
15. R.J. Carroll, A.E. Eyler, J.C. Denny. *AMIA Annu Symp Proc.* (2011).
16. T.S. Cole, et al. *Pediatr. Rheumatol.* **11**, 45 (2013)
17. M.A. Nolin, et al., *Semantic Web Challenge (ISWC 2008)*. (2008)
18. Callahan, J. Cruz-Toledo, P. Ansell, M. Dumontier, *ESWC 2013*. 200 (2008)
19. D. Blumenthal, M. Tavenner, *New England Journal of Medicine.* **363**, 501 (2010)
20. J. Frankovich, C.A. Longhurst, S.M. Sutherland, *N Engl J Med* 365.**19,** 1758 (2011)
21. J. Blake, *Nature biotechnology* **22**, 773 (2004)
22. http://www.healthline.com/health/rheumatoid-arthritis/medications-list (2015)
23. http://bio2rdf.org/ (2015)
24. http://www.ncbi.nlm.nih.gov/omim/ (2015)
25. http://www.genome.jp/kegg/ (2015)
26. http://www.drugbank.ca/ (2015)
27. M.K. Lodhi, et al. *Advances in Data Mining: Applications and Theoretical Aspects*. 56 (2015)
28. V.M. Castro, et al. *American Journal of Psychiatry* **172**, 363 (2014).
29. H.S. Huang, et al. *Journal of the American Medical Informatics Association* **21**, 1069 (2014)
30. http://bioportal.bioontology.org/ontologies/AI-RHEUM (2015)
31. D.S. Collier, et al., *Arthritis & Rheumatism.* **61**, 495 (2009)
32. N. Noy, et al. *Nucleic acids research* (2009)
33. D.L. Rubin, N.H. Shah, N.F. Noy. *Briefings in bioinformatics* **9**, 75 (2008).
34. N. H. Shah, et al., *BMC bioinformatics* **10**, S1 (2009)