

# Poster: Classifying Primary Outcomes in Rheumatoid Arthritis: Knowledge Discovery from Clinical Trial Metadata

Yuanyuan Feng<sup>1</sup>, Vandana P. Janeja<sup>1</sup>, Yelena Yesha<sup>1</sup>, Naphthali Rische<sup>2</sup>, Michael A. Grasso<sup>3</sup>, Amanda Niskar<sup>3</sup>

<sup>1</sup>University of Maryland Baltimore County, Baltimore, USA,

<sup>2</sup>Florida International University, Florida, USA

<sup>3</sup>University of Maryland School of Medicine, Baltimore, USA,

<sup>4</sup>Arthritis Foundation, USA

**Abstract**— Early prediction of treatment outcomes in RA clinical trials is critical for both patient safety and trial success. We hypothesize that an approach employing metadata of clinical trials could provide accurate classification of primary outcomes before trial implementation. We retrieved RA clinical trials metadata from ClinicalTrials.gov. Four quantitative outcome measures that are frequently used in RA trials, i.e., ACR20, DAS28, and AE/SAE, were the classification targets in the model. Classification rules were applied to make the prediction and were evaluated. The results confirmed our hypothesis. We concluded that the metadata in clinical trials could be used to make early prediction of the study outcomes with acceptable accuracy.

**Keywords**—data mining; clinical trials metadata; rheumatoid arthritis; outcome prediction

## I. INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease, which results in irreversible joint damage and disability characterized by reduction in both physical function and quality of life. Early prediction of the outcomes in RA clinical trials is critical for both patient safety and the success of drug development. The most common measures for the outcomes of RA trials include ACR20 and DAS28 [1].

Clinical trials outcomes are often predicted by simulation models using pharmacological information and clinical data [2]. For example, Anderson et al. used baseline data on disease activity to predict ACR20, DDAS, nACR and O'Brien's test [3]. This model requires that the recruitment has completed and information about the participants is gathered in advance of the prediction. Thus the model would at best work as the evaluation of a cohort, instead of decision support for the design.

In this study, we develop classification models based on the metadata of RA trials from ClinicalTrial.gov to study the potential of predicting the primary outcomes of a clinical study in a supervised manner. We hypothesize that an approach employing metadata of clinical trials can provide accurate classification of primary outcomes before trial implementation, and separating clinical trials based on the study design can improve the classification performance.

The main contributions of our work are threefold: (1) we integrate previous knowledge from clinical trials for classifying primary outcomes of the trials; (2) we propose to leverage the differences between treatment arms and placebo arms in the classification models; and (3) we evaluate the results across the training sets, as well as among different classification algorithms.

## II. METHODS

We retrieved study description, eligibility criteria and baseline data from ClinicalTrials.gov. Attributes were selected based on their frequencies in the retrieved cases to minimize the impact of missing values. Numeric attributes were discretized based on the quartiles in their distributions.

Clinical trial data are characterized by heterogeneous sources and high-dimensional outcome measures. Altiparmak et al. applied clustering algorithms and association rules in finding frequent lab outcome sets, in order to identify a group of factors that indicate the health state [4]. They identified homogeneous subsets of data by clustering the lab outcomes of each patient, over which they further identified common patterns. We extend their idea by dealing with heterogeneous data in our study. To disease activity measures, i.e., ACR20 and DAS28, the trial cohorts are often designed to include a placebo ground and a treatment. Thus we divided the dataset based on the major differences of treatments – placebo vs. treatment, instead of automatically clustering the data. In treatment safety evaluation, the trial cohorts are recruited focused on patients under the treatment. Thus, we didn't divide datasets for adverse events (AE) and serious adverse events (SAE) classification.

We applied different classification rules to the datasets and evaluated the classifier's performance by Friedman test.

## III. RESULTS

We focused on four primary outcome measures in RA trials – ACR20, DAS28, AE and SAE. We designed three different approaches to train the classification algorithms for ACR20 and DAS28, including a treatment set, a placebo set and a combined set. We used all cohorts to classify AE and SAE.

a) *ACR20*: The models for placebo set achieved higher performance, with the highest average scores for accuracy (82.70%), AUC (0.95), Kappa statistic (0.74), precision (0.84) and recall (0.83). However, we failed to find any significant difference of the performance between treatment set and combined set in post hoc Friedman test with bonferroni correction.

b) *DAS28*: The placebo set has the best performance with high average scores for accuracy (74.75%), AUC (0.863), Kappa statistic (0.62), precision (0.76) and recall (0.75); the performance for treatment set is also acceptable with 66% accuracy and 0.79 AUC on average. Our test results show that models for treatment and placebo outcomes independently perform better than the combined outcomes ( $p < 0.01$ ). However, we didn't detect any significant difference between predictive performances of the treatment set and placebo set, except for the precision measure, in which placebo set outperformed the treatment set ( $p = 0.015$ ).

c) *Adverse events and serious adverse events*: The adverse events set achieves higher performance than serious adverse events, with average scores of 68.92% accuracy, 0.83 AUC, 0.53 Kappa statistic, 0.72 precision and 0.69 recall.

In the classification models for ACR20, the performance of classifiers varied significantly in response to accuracy ( $p = 0.015$ ), Kappa statistic ( $p = 0.013$ ) and recall ( $p = 0.015$ ). We had limited power to identify any significant difference for AUC and precision. In the classification models for DAS28, we identified significant difference in the rank of classifiers for accuracy ( $p = 0.036$ ), AUC ( $p = 0.029$ ), precision ( $p = 0.036$ ) and recall ( $p = 0.036$ ). We used the mean rankings of performance values in further identification of best classifiers. The Friedman tests for the rankings in both ACR20 and DAS28 indicated significant differences among the classifiers. Random Forrest ranks as the best classification algorithm in classifying the outcome of ACR20 ( $p = 0.02$ ). For DAS28, we found that Random Forrest, J48 and J48 graft performed better than the other classifiers in our analysis ( $p = 0.03$ ).

#### IV. DISCUSSION

In this study, we developed classification models using metadata from clinical trials to classify primary outcomes. This lays the foundation for early prediction using metadata from clinical trials. The early prediction of trial outcomes is especially important for patient safety in studies with irreversible diseases. Rheumatoid arthritis is one such disease. We used both study description data and criteria text data to classify the most common outcomes for RA trials. We included all the studies with the condition of RA that can be downloaded from ClinicalTrials.gov. Our cases span from drug treatment to device intervention, with adult and juvenile patients under varied disease activities. With such diverse cases, we categorized our attributes and applied decision tree algorithms in our classification models. Further, the design of clinical trials – different arms for the comparison – informed our methodology development. We noticed the inherent differences between treatment arms and placebo arm, and created multiple training sets in order to improve the classification accuracy.

We found that classification models for the treatment set and placebo set perform better than for the combined set, which confirmed our hypothesis. Among the two, placebo set has a better performance, which indicates that the treatment arm may have greater variations that may not be so evident from metadata of the trials, whereas the placebo arm with clear outcomes has common characteristics in majority of cases.

The poor performance when using combined set is due to two main reasons. First, the combined set merged the placebo and treatment arms together and lost the identifiable features for the outcomes. As we observed in our dataset, the number of enrollment, age and gender ratio often varied among arms in a single study. These factors are closely related with participants' disease progression and study evaluation. When substituted with general descriptions, the attributes lost their variations among different outcome categories. Secondly, the outcomes in the combined set were presented by the difference between treatment arms and placebo arms. The calculated values could not fully represent the original values. For example, the outcome of DAS28 in each arm can be the change of value from the baseline. In the combined set, the result was calculated by the difference of the changes between placebo arm and treatment arm. It overlooked the difference between two arms in the baseline data and oversimplified the outcome measure.

Without dividing into any subgroups, the AE set achieved a model with high accuracy. The reason for the high accuracy is mainly because the adverse events include both drug related events and naturally developed events. Since the same condition has similar manifestations, the adverse events can be easily described by the patterns in the metadata. The models for SAE performed poorly. We only retrieved 47 cases that had serious adverse events as primary outcomes. SAE are often rare conditions, which are difficult to classify in small dataset.

#### V. CONCLUSION

This study provides different methods in classifying primary outcomes for rheumatoid arthritis clinical trials using the trial metadata. The good performance of our classification models confirms our hypothesis and lays the foundation for early prediction using metadata from clinical trials. Our results also suggest that classifying outcomes based on the study design improves the model performance.

#### REFERENCES

- [1] J. Anderson, L. Caplan, J. Yazdany, M. Robbins, T. Neogi, K. Michaud, K. Saag, J. O'dell and S. Kazi, "Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice", *Arthritis Care & Research*, vol. 64, no. 5, pp. 640-647, 2012.
- [2] N. Holford, H. Kimko, J. Monteleone and C. Peck, "Simulation of clinical trials." *Annual review of pharmacology and toxicology*, 40, no. 1: 209-234, 2000.
- [3] J. Anderson, J. Bolognese and D. Felson, "Comparison of rheumatoid arthritis clinical trial outcome measures: A simulation study", *Arthritis & Rheumatism*, vol. 48, no. 11, pp. 3031-3038, 2003.
- [4] F. Altıparmak, H. Ferhatosmanoglu, S. Erdal and D. Trost, "Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases", *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 2, pp. 254-263, 2006.