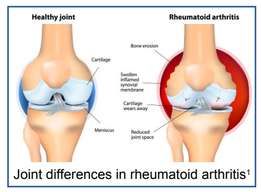
Predicting Methotrexate Response in Treatment-Naïve Early Arthritis Patients Through the Blood Proteome

Cara Chang[1,2], Benjamin Hur[3], John M. Davis III[4], and Jaeyun Sung[4,5]

[1]Department of Statistics and Data Science, Northwestern University, [2]Summer Internship Grant Program (SIGP), Northwestern University, [3]Microbiomics Program, Center for Individualized Medicine, Mayo Clinic, [4]Division of Rheumatology, Department of Internal Medicine, Mayo Clinic, [5]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic

BACKGROUND

- **Arthritis** is characterized by inflammation in joint tissue, causing chronic pain
- In 2022, the age-adjusted prevalence of diagnosed arthritis in U.S. adults was 18.9%[2]



Joint differences in rheumatoid arthritis[1]

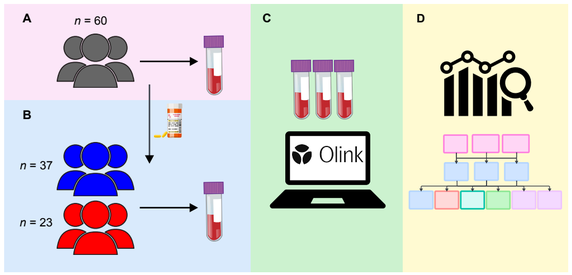
- Females and older adults have a higher risk of developing arthritis
- **Methotrexate (MTX)** is a disease-modifying anti-rheumatic drug (DMARD) that is recommended for initial treatment of arthritis[3]
- However, **MTX is ineffective in up to 50% of cases**[4], and there is currently no way to tell which patients will respond positively

OBJECTIVES

Develop a machine learning approach that uses **plasma proteomic profiles to predict patient response** (i.e., responders and non-responders to MTX) in patients with early arthritis. Specifically, we will:

- Identify a subgroup of proteins that best predicts response
- Uncover features relevant to MTX treatment

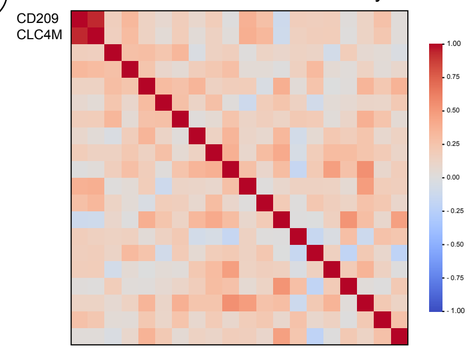
STUDY DESIGN



A. Blood samples of ***n* = 60 treatment-naïve patients** with early arthritis were collected at baseline (i.e., before treatment)
B. Samples were collected again after 3–4 months following MTX treatment
   - 37 responders & 23 non-responders
C. Olink® Proximity Extension Assay technology measured the **relative abundance of 2,904 proteins** from plasma
D. Machine learning was performed with data transformation, cross validation, feature selection
   - A pipeline was created to automated all combinations of techniques to systematically identify models with highest performance

EXPLORATORY DATA ANALYSIS

1 Heatmap of 20 Most Correlated Features Reveals Little Multi-colinearity



2 No Initial Differences Between Responders and Non-Responders

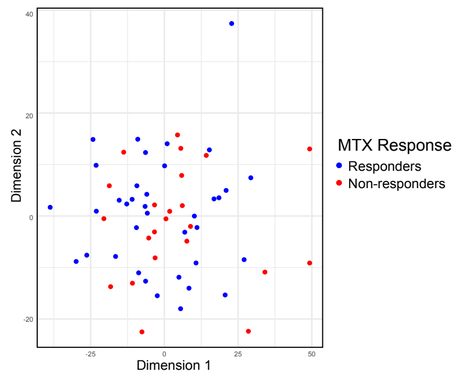


Figure 1. Heatmap of 20 Most Correlated Features Reveals Little Multi-colinearity: Only two proteins were found to be highly correlated of the 2,904, meaning that multi-colinearity was not an issue.
Figure 2. No Initial Significant Differences Between Response Groups: Multidimensional scaling plot does not contain clusters that would indicate higher similarity within groups.

MACHINE LEARNING PIPELINE DESIGN

3

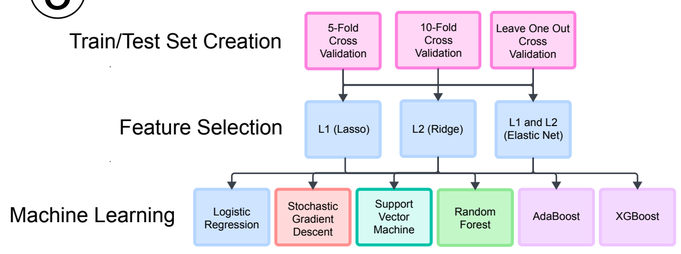


Figure 3. Machine Learning Pipeline: Using only baseline proteomics data, predictive models were built in 3 steps:

1. Train/test set creation
2. Feature selection
3. Machine learning

The pipeline created models by combining these steps in all possible ways

MACHINE LEARNING RESULTS

Cross-Validation	Feature Selection	Model Type	ROC-AUC	Accuracy	Precision
10-fold	L1 Penalization (Lasso)	Logistic Regression	0.7486	68.33%	77.17%
10-fold	L1 Penalization (Lasso)	Support Vector Machine	0.7389	68.33%	77.17%
10-fold	None	Support Vector Machine	0.7389	68.33%	72.83%
5-fold	L1 Penalization (Lasso)	Support Vector Machine	0.6086	66.67%	72.11%

CONCLUSIONS

- **Simple models (logistic regression, support vector machines) tend to perform the best**
  - Other models may be prone to overfitting because of small sample size
  - Simplicity is better for interpretability and computation
  - Some ensemble models still show promise
- Approximately **100 proteins** were consistently selected throughout feature selection
  - The relationship with MTX response is currently unknown
- Multiple models outperform a blind guess of response or no response (ROC-AUC > 0.5)
- Plasma proteomics may have clinical utility in understanding the effects of MTX treatment
  - Especially important in the early course of disease
- A holistic approach using multi-omics may further improve predictive power (work ongoing)

REFERENCES

1. https://www.advancedrheumatology.net/service/rheumatoid-arthritis-treatment/
2. “Arthritis in Adults Age 18 and Older: United States, 2022” Elgaddal, *et al*., CDC/National Center for Health Statistics, 2022.
3. “Methotrexate (Rheumatrex, Trexall, Otrexup, Rasuvo).” American College of Rheumatology 2024.
4. “Effectiveness profiles and dose dependent retention of traditional disease modifying antirheumatic drugs for rheumatoid arthritis. An observational study.” Aletaha and Smolen 2002.

ACKNOWLEDGMENTS

- This project was funded by the Mayo Clinic Center for Individualized Medicine and Division of Rheumatology
- We would like to thank the Sung Lab for their scientific and editorial support and feedback, especially Thomas J. Pelowitz
- We would also like to thank Rene Mohammadi and Jasmine Lunia for their support throughout this project

Sung Lab at Mayo Clinic computational systems medicine

Northwestern University