Apply Machine Learning to Identify Unique Patient Clusters and Associated Key Biomarkers in Rheumatoid Arthritis

Developing a Point of Care Test with a Multi Biomarker Panel for Patient Classification and Disease Progression

Gabriella Lui
Fair ID: USWA03
Project ID: TMED032
Category: Translational Medical Science / Disease Detection and Diagnosis
What is Rheumatoid Arthritis?

- **Rheumatoid Arthritis (RA)** is an autoimmune disease, causing inflammation in joints and eventually bone erosion and joint deformity.
- As of now, there is **no cure**– the current goal of RA treatment is **remission**
- Super complex disease – necessitates a trial and error treatment approach

The Important Role of Biomarkers in the Pathogenesis Of RA

**Biomarkers can predict the onset of RA and forecast disease activity in established RA patients**

- Cytokines are biomarkers of systemic inflammation that may appear prior to symptoms
- While RA **has no cure**, pre-emptive treatment using disease-modifying antirheumatic drugs (DMARDs) in the ‘Window of Treatment (WoT)’, may prevent the onset of RA
- Biomarkers also actively appear in later stages of RA and forecast disease progression or remission
- Advantageous for disease and treatment monitoring
The Need for a Point of Care Test for RA Disease Progression

• Increase Health equity and access to quality patient care
• Due to Covid-19, patients have been unable to travel to the clinic
• Resorting to Telemedicine, rheumatologists have not had access to quantitative measures to assess their RA patients’ conditions due to travel restrictions
• This will aid research in RA and precision medicine, allowing physicians to pinpoint an optimal treatment plan for a specific phenotype of RA patients sooner

What’s a Point of Care (POC) Test?

Proof of concept LFT that I designed last summer using IL-6 as the assay analyte
Apply machine learning modeling to a 15 year longitudinal RA dataset to determine the optimal panel of a few biomarkers (out of 40) that might accurately predict RA disease progression and treatment response.

Create a Point of Care (POC) assay that tests for the identified panel of highly correlated biomarkers that can reveal a patient’s RA disease progression.
University of California San Francisco RA Cohort is ethnically and racially diverse. Clinical observations spanned over 15 years.

The biomarkers that were obtained at each visit for the 2 cohorts:

**VECTRA panel** (EGF, IL6, LEPTIN, VEGF, CRP, SAA, VCAM1, MMP1, MMP3, RESISTIN, TNFRI, YKL40, HsCRP)

**Extended Cytokine Panel** (IL1b, IL1ra, IL2, IL4, IL5, IL6, IL7, IL8, IL9, IL10, IL12, IL13, IL15, IL17, Eotaxin, FGF, GCSF, GMCSF, IFNg, IP10, MCP1, MIP1a, MIP1b, PDGF, RANTES, TNFa, VEGF)
Identifying Key Biomarkers for RA Prognosis of a General Cohort

Data Preparation, Cleansing & Transformation

Maximize number of observations without introducing additional error

- Any observations with missing or out-of-range Vectra-DA & Extended Cytokine Panel values were eliminated
- Input raw longitudinal observations without imputation
- Numeric observations were “scaled” by subtracting the values by the center value and then divide by the standard deviation

More on LASSO
- LASSO (Least Absolute Shrinkage and Selection Operator)
  \[
  \sum_{i=1}^{n} (y_i - \sum_{j} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|
  \]
  - As \( \lambda \) increases, more coefficients are set to zero. Best \( \lambda \) was calculated and then fed to the algorithm
- \( y_i \) is DAS28 (RA Disease Activity)
- \( x_{ij} \) are extended cytokines panel & Vectra components
- Some \( \beta \)s are shrunk to exactly zero => variables are being eliminated

Dimensionality Reduction & Feature Selection

Parse out only a few highly correlated key biomarkers

- Principal component analysis does not retain original features. Instead, it provides vectors that are linear combinations of the original features.
- LASSO retains original features by performing feature selection by setting the coefficient of nondominant variables to zero.

- Apply LASSO machine learning algorithm to select the key biomarkers / Optimal Panel that are highly correlated to Disease Activity
- Apply Multi-Variant Regression to the optimal biomarker panel to “Predict” Disease Activity
Optimal Biomarker Panel for the Prognostication of a General Patient Cohort

- **EGF** - expressed on synovium (lining of cell) and promotes fibroblast proliferation in RA and production of pro-inflammatory cytokines.
- **IL6** - promotes production of autoantibodies, induces endothelial cells to produce IL8 and stimulate osteoclasts (erosions). Enhance production of MMPs ➔ damage cartilage. Amongst many other things.
- **Leptin** - pro-inflammatory, secreted by fat cells. Stimulates the immune system in many ways.
- **SAA** - Similar to CRP, made by liver, many downstream proinflammatory effects.
- **MMP3** - marker of tissue destruction released by chondrocytes/synovial cells leading to cartilage destruction.
- **HSCRP** - CRP released by liver, similar to SAA, many downstream proinflammatory effects.

**Interleukin 9 (IL9)**
- Secreted by T Helper type 9 cells (subset of CD4+ T Cells)
- **Implicates the pathogenesis of autoimmune diseases**
- Increases MMP production - a protein that contributes to tissue damage
- Facilitates Th17 differentiation – pro inflammatory T cells

**IP10 / CXCL10 (Gamma Interferon Induced Protein 10)**
- Overexpressed in synovial tissue + serum in pre-and established RA
- Correlates with disease activity DAS28ESR
Identifying Biomarker Panels for Unique phenotypes of established RA patients

Data Preparation, Cleansing & Transformation

- Any observations with out-of-range values were eliminated
- Missing values were imputed using MICE (Multivariate Imputation via Chained Equations)
- Cross-sectional observations were created by Z-score normalization to collapse data per patient
- Numeric observations were “scaled” by subtracting the values by the center value and then divided by the standard deviation

Associate Observations with Individual Patients

Unsupervised Machine Learning Clustering

- Unsupervised Machine Learning K-prototype Mixed variable types (numerical and categorical) clustering
- K-prototype Dunn Index to compute optimal number of Clusters ($k$)
- Calculate the best $\lambda$ (weightage for categorical values)
- Set $K$ and $\lambda$ in K-prototype modeling

<table>
<thead>
<tr>
<th>Model based</th>
<th>K mean</th>
<th>Hierarchical</th>
<th>K proto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbiased groupings- without user dictated # of clusters</td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Mixed parameters- categorical and continuous</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Optimized to handle large amounts of data</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Identifying Biomarker Panels for Unique phenotypes of established RA patients

Dimensionality Reduction & Feature Selection

- Parse out only a few highly correlated key biomarkers within each patient phenotype
- Un-collapse the aggregated patient data with associated cluster identifications
- Apply LASSO machine learning algorithm onto the longitudinal data over time to select the key biomarkers / Optimal Panel that are highly correlated to Disease Activity in each cluster
- Apply Multi-Variant Regression to the optimal biomarker panel to “Predict” Disease Activity

Longitudinal Data Analysis

- Perform Summary Statistics on each cluster’s clinical baseline data and longitudinal data over time
- Observe unique characteristics in each of the clusters
### Summary Statistics for 4 Distinct Patient Clusters

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>54.8 (13.6)</td>
<td>63.6 (9.7)</td>
<td>50.8 (14.9)</td>
<td>58.2 (15.8)</td>
<td>50.3 (12.1)</td>
</tr>
<tr>
<td><strong>Female Sex</strong></td>
<td>318 (85.3%)</td>
<td>101 (87.1%)</td>
<td>57 (81.4%)</td>
<td>11 (78.6%)</td>
<td>149 (86.1%)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>181 (48.5%)</td>
<td>22 (19.0%)</td>
<td>47 (67.1%)</td>
<td>5 (35.7%)</td>
<td>107 (61.8%)</td>
</tr>
<tr>
<td>Asian</td>
<td>123 (33.0%)</td>
<td>73 (62.9%)</td>
<td>8 (11.4%)</td>
<td>6 (42.9%)</td>
<td>36 (20.8%)</td>
</tr>
<tr>
<td>Black</td>
<td>35 (9.4%)</td>
<td>12 (10.3%)</td>
<td>8 (11.4%)</td>
<td>2 (14.3%)</td>
<td>13 (7.5%)</td>
</tr>
<tr>
<td>White + Other</td>
<td>34 (9.1%)</td>
<td>9 (7.7%)</td>
<td>7 (10.0%)</td>
<td>1 (7.1%)</td>
<td>17 (9.8%)</td>
</tr>
<tr>
<td><strong>RF</strong></td>
<td>315 (84.5%)</td>
<td>104 (89.7%)</td>
<td>56 (80.0%)</td>
<td>13 (92.9%)</td>
<td>142 (82.1%)</td>
</tr>
<tr>
<td><strong>ACPA</strong></td>
<td>297 (79.6%)</td>
<td>98 (84.5%)</td>
<td>54 (77.1%)</td>
<td>12 (85.7%)</td>
<td>133 (76.9%)</td>
</tr>
<tr>
<td><strong>Disease Duration</strong></td>
<td>7.8 (7.6)</td>
<td>13.7 (9.7)</td>
<td>5.4 (4.6)</td>
<td>6.7 (5.7)</td>
<td>4.9 (3.8)</td>
</tr>
<tr>
<td>csDMARD</td>
<td>344 (92.2%)</td>
<td>108 (93.1%)</td>
<td>63 (90.0%)</td>
<td>13 (92.9%)</td>
<td>160 (92.5%)</td>
</tr>
<tr>
<td>bDMARD</td>
<td>185 (49.6%)</td>
<td>45 (38.8%)</td>
<td>30 (42.9%)</td>
<td>4 (28.6%)</td>
<td>106 (61.3%)</td>
</tr>
<tr>
<td>Prednisone</td>
<td>6.70 (3.81)</td>
<td>6.02 (3.96)</td>
<td>8.62 (4.92)</td>
<td>5.77 (1.37)</td>
<td>6.33 (2.81)</td>
</tr>
<tr>
<td><strong>BMI</strong></td>
<td>28.2 (4.5)</td>
<td>26.6 (3.8)</td>
<td>28.7 (3.8)</td>
<td>28.0 (6.2)</td>
<td>29.1 (4.7)</td>
</tr>
<tr>
<td><strong>DAS28 ESR</strong></td>
<td>4.2 (1.1)</td>
<td>4.2 (1.0)</td>
<td>5.5 (0.8)</td>
<td>3.9 (0.8)</td>
<td>3.7 (0.9)</td>
</tr>
</tbody>
</table>

### Lasso Results

| **EGF**                                  | -0.16*         | -0.41***        | --              | --              | -0.20**          |
| **Leptin**                               | 0.15**         | --              | 0.21*           | --              | 0.21**           |
| **CRP**                                  | 0.34**         | 0.54***         | --              | --              | --               |
| **VCAM1**                                | --             | --              | --              | -0.73           | --               |
| **YLK40**                                | --             | 0.26*           | --              | --              | --               |

*p < 0.05*, <0.01**, <0.001***
4 Distinct Patient Clusters Identified Biomarker Panels for Each Cluster
Findings submitted to the EULAR (European League Against Rheumatism) 2021 Conference

Cluster 1 (n=116) → CRP, EGF, YKL40

**Patient Profile** – long standing RA patients with high chronic inflammation.
Older (63.6±9.7) with long disease duration (13.7±9.7) notable these patients are on prednisone.

**CRP** generalized downstream proinflammatory protein, making it a good summary marker for many other proinflammatory proteins.

**YKL-40** is produced by chondrocytes and mature macrophages, has roles in cell proliferation. More inflammation = more YKL-40?

Cluster 2 (n=70) → Leptin

**Patient Profile** – This is an obesity related arthritis (not true RA) cohort
Higher BMI (28.7±3.8) - fat associated leptin
Highest Prednisone Dosage (8.6 ± 4.9) – possibly contributes to weight gain.
Notable lowest CCP and RF positivity (80%)

Cluster 3 (n=14) → VCAM1

**Patient Profile** – Non-aggressive RA cohort possibly in remission
Nuanced since it has the lowest n
Relatively low DAS28ESR (3.9 ± 0.8)

Cluster 4 (n=173) → EGF, Leptin

**Patient Profile** – early-stage RA cohort with high biologic usage and weight
Largest cluster with shortest disease duration (4.94.9±3.8 years)
Highest biologic bDMARD use (61.3%)
Higher BMI (29.1 ± 4.7) - fat associated leptin

**Leptin** is pro-inflammatory and secreted by fat cells. Stimulates the immune system in many ways.

**EGF** expressed on synovium and promotes fibroblast proliferation in RA. Produces pro-inflammatory cytokine (downstream effects) making it another good summary marker

**VCAM1** is a cellular adhesion molecule meaning it helps other cells get to sites of inflammation. Possibly indicates remission.
Testing the Lateral Flow Assay Sensitivities – Sequential vs. Premixed

Premixed
- Low cost
- Home use
- Sensitive
- Non-invasive (only requires a few drops of blood)

Sequential lateral flow test #3

<table>
<thead>
<tr>
<th>Concentration (ng/mL)</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Positive</td>
</tr>
<tr>
<td>2.5</td>
<td>Positive</td>
</tr>
<tr>
<td>1</td>
<td>Positive</td>
</tr>
<tr>
<td>0.5</td>
<td>Positive</td>
</tr>
<tr>
<td>0.25</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative control</td>
<td>Negative</td>
</tr>
</tbody>
</table>

- Visual LOD for sequential LFT is 0.5 ng/mL