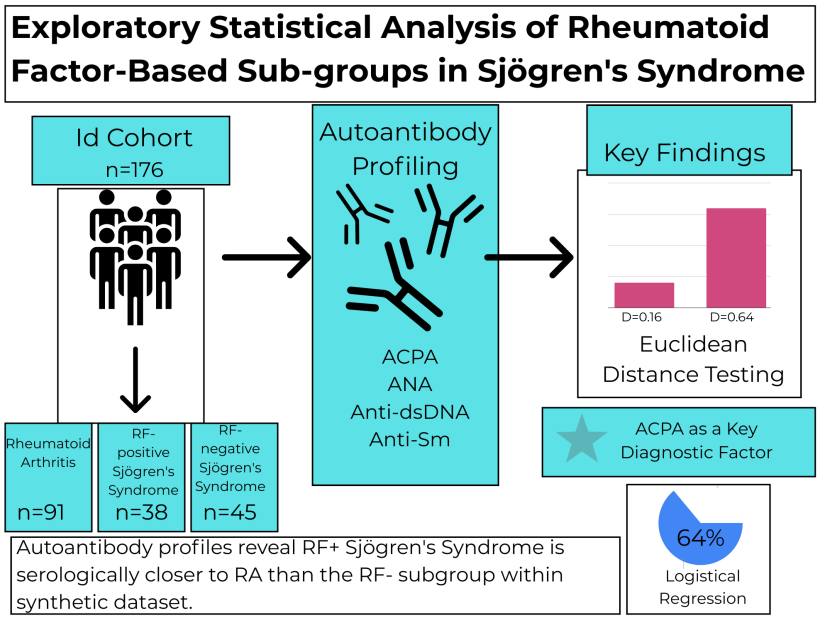
Graphical Abstract

Exploratory Statistical Analysis of Rheumatoid Factor–Based Subgroups in Sjögren's Syndrome

Nathan Zhuang, Jacqueline Howells



Highlights

Exploratory Statistical Analysis of Rheumatoid Factor–Based Subgroups in Sjögren's Syndrome

Nathan Zhuang, Jacqueline Howells

- RF-positive SS shows immunological similarity to Rheumatoid Arthritis in Welch's T-tests.
- RF-negative SS exhibits distinct lupus-like autoantibody profile in Tukey's post hoc.
- This study supports hypothesis generation of personalized treatment approaches in SS.

Exploratory Statistical Analysis of Rheumatoid Factor–Based Subgroups in Sjögren's Syndrome

Nathan Zhuang[a,b], Jacqueline Howells[a,c]

[a] *Polygence, Menlo Park, California, United States of America*
[b] *Western Canada High School, Calgary, Alberta, Canada*
[c] *Brown University, Providence, Rhode Island, United States of America*

Abstract

Rheumatoid factor (RF) is a core diagnostic marker for rheumatoid arthritis (RA) but is also common in Sjögren's syndrome (SS). RF-positive SS patients often exhibit RA-like features, suggesting a potential overlapping subtype. This study investigated whether RF status in SS is associated with distinct autoantibody profiles and greater immunological similarity to RA. We analyzed a synthetic dataset (n=176), including 91 RA patients and 85 SS patients (38 RF-positive, 45 RF-negative), across four autoantibodies: anti-citrullinated protein antibodies (ACPA), antinuclear antibodies (ANA), anti-double-stranded DNA (anti-dsDNA), and anti-Smith (anti-Sm). Statistical analyses included t-tests, Chi squared, Effect sizes, ANOVA, MANOVA, Euclidean distance measures, and logistic regression. RF-positive SS showed similar ACPA levels to RA in Welch's T-tests, whereas RF-negative SS differed significantly in ANA and anti-dsDNA, consistent with a lupus-like profile within ANOVA. MANOVA and distance analyses indicated that RF-positive SS is immunologically closer to RA, while RF-negative SS forms a distinct cluster. Logistic regression supported moderate-low classification performance. Overall, these findings may support the hypothesis of an RF-associated SS subtype overlapping with RA and highlight the potential of RF testing to improve disease stratification and guide therapeutic decisions.

All data and code are publicly available (Zhuang, 2025).

Keywords: Rheumatoid arthritis, Sjögren's syndrome, Rheumatoid factor, Autoantibodies, Synthetic Dataset.

1. Introduction

Rheumatoid Arthritis (RA) is a chronic autoimmune disease characterized by persistent joint inflammation, systemic involvement, and progressive tissue damage, affecting approximately 0.5–1% of the global population (Aletaha et al., 2010). RA remains a leading cause of autoimmune-related disability. A key diagnostic and mechanistic hallmark of RA is the presence of Rheumatoid Factor (RF), an autoantibody directed against the Fc portion of immunoglobulins G, A, and M. These immune complexes trigger inflammation and contribute to joint degradation (Bournazos et al., 2020; Feldman and Ginsburg, 2021; Ingegnoli et al., 2013). Because RF is highly prevalent in RA and included in the ACR/EULAR diagnostic criteria, it remains a critical benchmark for autoantibody research (Aletaha et al., 2010).

RF is not exclusive to RA. Elevated RF levels are also observed in other autoimmune conditions, notably Sjögren's Syndrome (SS), in which 75–95% of patients test positive (Elagib et al., 1999). SS is a systemic autoimmune disorder primarily targeting exocrine glands, leading to symptoms such as dry eyes and mouth (Brito-Zerón et al., 2016). In RF-positive SS patients, overlapping features with RA, such as joint pain and stiffness, are also often observed, complicating diagnosis and treatment (Kim et al., 2020). RA and SS are among the most interconnected autoimmune diseases in terms of pathogenesis and late-stage symptoms, and up to 8.7% of RA patients are also diagnosed with secondary SS (Kim et al., 2020).

Given these overlapping serological markers and clinical symptoms between SS and RA, we hypothesize that RF positive SS patients would exhibit greater immunological similarity to RA patients than RF-negative SS patients, within a synthetic dataset modeled on real-world autoimmune distributions. Specifically, we will be examining shared autoantibody profiles (e.g., anti-citrullinated protein antibodies [ACPAs], antinuclear antibodies [ANAs]) that may be more characteristic of RA than of RF-negative SS (Bournazos et al., 2020; Nashi and Shmerling, 2021).

To explore this, we analyzed the publicly available: Comprehensive Autoimmune Disorder Dataset (Ragheb, 2024), a synthetic dataset generated from population distributions and clinical guidelines to reflect real-world autoimmune patterns. Using this dataset, we focused on a targeted set of key variables, including RF, ACPA, Anti-dsDNA, Anti-Sm and ANA, to assess whether RF status stratifies SS into two potential subgroups (Damoiseaux and Van Beers, 2023; Kurowska et al., 2017; Liu et al., 2024; Nashi and

Shmerling, 2021; Van Beers and Schreurs, 2022).

This study aims to: Determine whether RF-positive SS patients show greater immunological similarity to RA patients than RF-negative SS patients through comparison of autoantibody pressence profiles.

The completion of this work will be able to clarify whether RF-positive SS represents a true overlap with RA, a misclassification, or a distinct clinical subtype. While healthy controls are not included, their absence is acknowledged as a limitation and a potential focus for future studies. By refining interpretation of RF across autoimmune diagnoses, this study aims to inform strategies for improving diagnostic precision and understanding immunological patterns in complex autoimmune conditions.

2. Methods

2.1. Dataset Description and Deduplication

Data for this study were obtained from the publicly available Comprehensive Autoimmune Disorder Dataset (Ragheb, 2024), hosted on Kaggle (DOI: 10.34740/KAGGLE/DSV/9807734) under an MIT license. This dataset is synthetically generated, designed to reflect real-world distributions of autoimmune diseases while maintaining patient privacy. The dataset was derived from human patients and de-identified, creating what is referenced as a synthetic data set that can be used for future metadata analysis relevant to translational biomedical research. It includes patient-level data across 114 autoimmune disorders, including Sjögren's Syndrome and Rheumatoid Arthritis, as well as antibody profiles, hematologic markers, and symptom variables. The dataset was modeled on peer-reviewed clinical studies, following the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR). Additionally, existing clinical databases, including publicly accessible datasets and research from health institutions focused on autoimmune disease characteristics (Ragheb, 2024).

For the present analysis, we focused exclusively on SS and RA patients. The initial dataset contained multiple entries per patient, therefore a deduplication step was performed by matching patient IDs and antibody profiles to ensure that each individual contributed only a single record. After deduplication and filtering for complete antibody data, the final dataset (of RA and SS) included 176 unique patients: 85 with SS and 91 with RA. Within the SS group, patients were further stratified by rheumatoid factor (RF) sta-

tus, described from now on as RF-positive and RF-negative, with a total of 38 RF-positive SS patients and 45 RF-negative SS patients.

From the aggregate data, patient age range was not significant over z-scores in the dataset. Age was therefore considered uniform, and sex distribution was roughly balanced across groups, again using z-scores. Because these demographic factors showed little variation, they were not included as covariates in subsequent analyses.

This approach allows for hypothesis generation regarding immunological similarities and differences between SS subgroups and RA, while avoiding privacy concerns associated with real painted data. All preprocessing and analysis scripts are available at `https://github.com/NathanZhuangCalgary/SS_RA_Antibody_Analysis` to ensure reproducibility.

2.2. Variable Selection

A targeted set of four autoantibodies was selected for analysis based on their known clinical and diagnostic relevance:

Table 1: Key autoantibodies analyzed in this study

Antibody	Disease Association	Target	Diagnostic Role
ACPA	RA[a,c]	Citrullinated proteins	High specificity for RA[a,c]
ANA	SS, RA, SLE[d,e]	Nuclear components	Broad autoimmune screening
Anti-dsDNA	SLE[b]	Double-stranded DNA	Detect overlap syndromes
Anti-Sm	SLE[f]	snRNPs	Specific for lupus

[a](Aletaha et al., 2010) [b] (Damoiseaux and Van Beers, 2023) [c] (Kurowska et al., 2017) [d] (Liu et al., 2024) [e] (Nashi and Shmerling, 2021) [f] (Van Beers and Schreurs, 2022)

This set includes RA-specific markers (ACPA), broad autoimmune markers (ANA), and markers relevant to potential overlap syndromes (Anti-dsDNA, Anti-Sm). Each antibody was coded as a binary variable: 1 = positive/present, 0 = negative/absent. These antibodies capture RA-specific markers (ACPA), broad autoimmune autoimmunity (ANA), and potential overlap syndromes (Anti-dsDNA, Anti-Sm) (Damoiseaux and Van Beers, 2023; Kurowska et al.,

2017; Liu et al., 2024; Nashi and Shmerling, 2021; Van Beers and Schreurs, 2022).

2.3. Statistical Analysis

All statistical analyses were conducted in Python 3.9, using NumPy v2.2.5 for numerical operations, pandas v2.2.3 for data manipulation, SciPy v1.16.2 for hypothesis testing, statsmodels v0.14.5 for additional statistical analyses, and seaborn v0.13.2 and matplotlib v3.10.1 for visualization. Auxiliary packages included plotly v6.3.0 for interactive plots, pillow v11.2.1 for image processing, orjson v3.11.3 for fast JSON handling, kaleido v1.1.0 for static image export, and packaging v25.0 for version management. Development and testing tools included pytest v8.4.2 and pytest-timeout v2.4.0.

2.3.1. Group Comparisons

To evaluate whether antibody presence differed between RA, RF-positive SS, and RF-negative SS patients, Chi-squared tests of independence were conducted for each antibody. This non-parametric test is suitable because it assesses whether the categorical antibody positivity (present/absent) is associated with diagnostic groups, accommodating the binary nature of the data (McHugh, 2013).

For comparisons of antibody levels, antibodies were treated as continuous variables, Welch's t-tests were applied. Welch's t-test is appropriate as it accounts for unequal variances and sample sizes between groups, ensuring robust comparisons of antibody levels across RA and SS subgroups (Delacre et al., 2017).

2.3.2. Effect Size and Cohen's d

To complement significance testing and quantify the magnitude of observed differences, effect sizes were calculated (Lakens, 2013). For continuous antibody measures (e.g., normalized levels), Cohen's d was used because it standardizes the difference between group means, providing a clear measure of practical significance for antibody level differences. For categorical variables (e.g., antibody presence/absence), effect sizes were calculated to quantify the strength of associations, suitable for the binary antibody data (Gaddis, 1998).

2.3.3. Multiple Comparisons

To test differences across all three diagnostic groups simultaneously, one-way ANOVA was applied to each continuous antibody measure. ANOVA is ideal as it compares means across multiple groups (RA, RF-positive SS, RF-negative SS) while controlling the overall error rate, detecting any group differences in antibody levels (Chatzi and Doody, 2023). Post hoc Tukey's tests were used to identify specific group pairs driving any significant differences. Tukey's HSD is appropriate because it controls for multiple comparisons, pinpointing which group pairs differ significantly (Kim, 2015).

To predict diagnostic group membership (RA vs. SS) based on antibody profiles, logistic regression was employed. This method fits for teh variables, as it models the probability of group membership using binary predictors, aligning with the study's aim to assess diagnostic utility of autoantibodies (Sperandei, 2014).

2.3.4. Multivariate Analysis

To assess the overall immunological profile beyond single antibodies, Euclidean distances were calculated between group centroids in the multidimensional space of Z-normalized antibody profiles (Glazko and Mushegian, 2010). This approach is effective as it quantifies serological similarity between groups in a multidimensional space, capturing overall profile differences. Additionally, Multivariate Analysis of Variance (MANOVA) was conducted to test whether the combined antibody profiles differ across diagnostic groups (Gaddis, 1998). MANOVA is well-suited because it evaluates multiple correlated antibody measures simultaneously, reducing Type I error and detecting multivariate patterns across RA and SS subgroups.

2.3.5. Significance Thresholds

All tests were two-sided with a significance of $\alpha = 0.05$. Post hoc adjustments, including Holm and FDR corrections, were applied for multiple comparisons to control false positives, ensuring robust inference across the multiple tests conducted (Aickin and Gensler, 1996; Korthauer et al., 2019).

2.4. Reproducibility

To ensure transparency and reproducibility, all preprocessing, statistical analyses, and figure generation scripts are publicly available at `https://github.com/NathanZhuangCalgary/SS_RA_Antibody_Analysis`. The synthetic dataset is publicly accessible via KaggleDOI: `https://doi.org/10.`

34740/KAGGLE/DSV/9807734.. All analyses were performed on the deduplicated dataset, ensuring each id contributed only a single record. Figures were generated programmatically using Seaborn and Matplotlib to maintain reproducibility of visualizations.

2.5. Ethical Considerations

This study exclusively used a synthetic, de-identified dataset, which contains no information that could identify real individuals. No human participants were directly involved, and therefore institutional review board (IRB) approval was not required. All analyses comply with open-data usage guidelines and respect the original dataset's MIT license. The synthetic nature of the dataset ensures that patient privacy is fully preserved while allowing investigation of autoimmune disease patterns.

3. Results

3.1. Overview of Synthetic Dataset

After filtering for complete records and removing duplicates where necessary, the final dataset included 176 patients, consisting of 85 Sjögren's syndrome (SS) patients and 91 rheumatoid arthritis (RA) patients. The SS group is being further subdivided by rheumatoid factor (RF) status, there were 38 RF-positive SS patients and 45 RF-negative SS patients. RA patients were 63.1% RF-positive, while RF-negative SS patients, by definition, had 0% RF positivity.

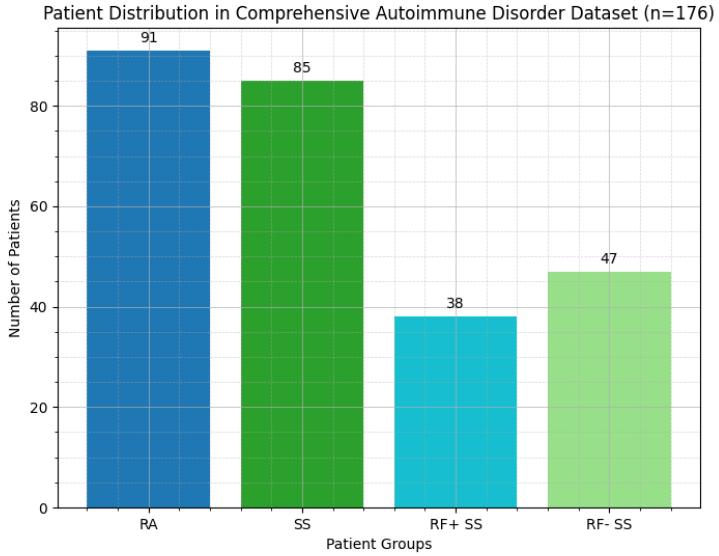


Figure 1: **Patient Distribution in Comprehensive Autoimmune Disorder Dataset (n=176).** Bar plot showing the numbers of id's of each subgroup analyzed, extrapolated from the dataset.

3.2. Autoantibody Comparison

Autoantibody levels–anti-citrullinated protein antibodies (ACPA), antinuclear antibodies (ANA), anti-double-stranded DNA (anti-dsDNA), and anti-Smith (anti-Sm) were compared between SS subgroups and RA patients using independent Welch's t-tests.

Table 2: Results of independent Welch's t-tests comparing autoantibody levels between RA and SS subgroups (RF-positive and RF-negative)

Antibody	RF-positive SS vs RA (t)	RF-positive SS vs RA (p)	RF-negative SS vs RA (t)	RF-negative SS vs RA (p)
ACPA	-0.194	0.847	0.514	0.608
ANA	0.691	0.492	2.424	0.017
Anti-dsDNA	-0.372	0.711	2.842	0.005
Anti-Sm	-0.194	0.847	-0.673	0.503

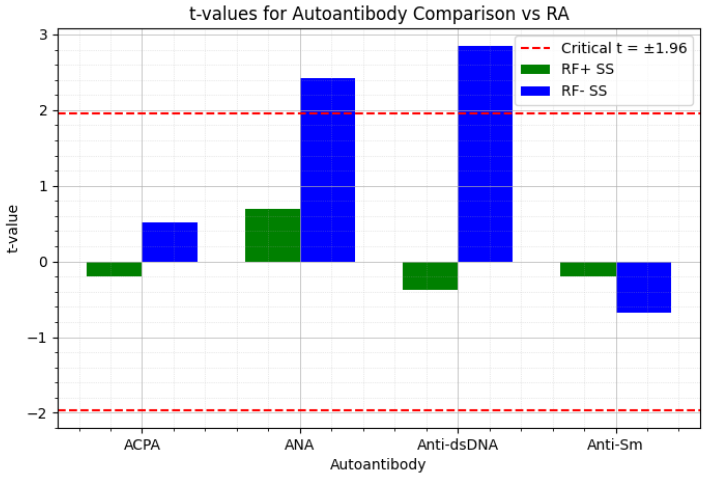


Figure 2: **T-test statistics for autoantibody levels in RF-positive and RF-negative Sjögren's syndrome compared with rheumatoid arthritis.** Bar plot showing two-group t-values for ACPA, ANA, anti-dsDNA, and anti-Sm. RF-positive SS values are shown in green; RF-negative SS values in blue. Dashed horizontal lines indicate the critical t-value threshold ($\alpha = 0.05$) for statistical significance. Positive t-values indicate higher mean levels in Sjögren's groups, while negative values indicate higher levels in RA.

3.3. Chi-squared Test of Independence

Chi-squared tests were performed to examine whether categorical autoantibody positivity (present vs absent) was associated with diagnosis (RA, RF-positive SS, RF-negative SS).

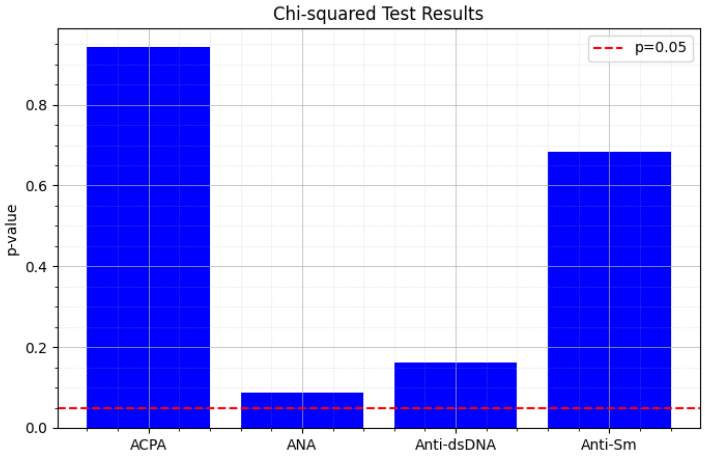


Figure 3: **Bar Plot of Chi-Squared Test Results for Autoantibody Frequency Distributions.** This bar plot illustrates chi-squared values for categorical comparisons of ACPA, ANA, Anti-dsDNA, and Anti-Sm frequencies across RA and SS groups. The relative bar heights highlight the magnitude of association for each autoantibody, providing a visual summary of distributional differences between diagnostic categories.

3.4. Effect Size Analysis

We calculated Cohen's d to quantify the effect sizes of antibody level differences between RA and the two SS subgroups (RF-positive and RF-negative).

Table 3: Cohen's d Effect Sizes Comparing Autoantibody Levels Between RA and RF-positive / RF-negative Sjögren's Syndrome Subgroups

Antibody	RA vs RF-positive SS	RA vs RF-negative SS
ACPA	0.038	-0.092
ANA	-0.131	-0.404
Anti-dsDNA	0.072	-0.494
Anti-Sm	0.038	0.122

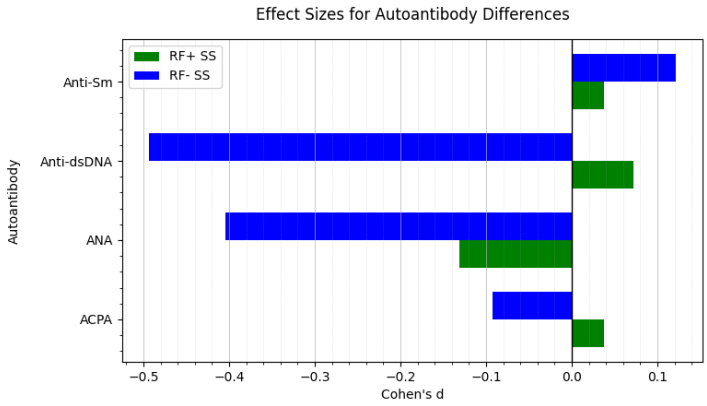


Figure 4: **Cohen's d effect sizes for autoantibody differences between Sjögren's subgroups and rheumatoid arthritis.** Horizontal bar chart of effect sizes (d) for ACPA, ANA, anti-dsDNA, and anti-Sm in RF-positive and RF-negative Sjögren's syndrome vs. rheumatoid arthritis. Positive values indicate higher antibody levels in the Sjögren's group, negative values indicate higher levels in RA. Vertical gridlines mark small (0.2), medium (0.5), and large (0.8) effect size thresholds.

3.5. One-way ANOVA

A one-way Analysis of Variance (ANOVA) was conducted to determine if there were statistically significant differences in mean autoantibody levels across the three diagnostic groups: Rheumatoid Arthritis (RA), Rheumatoid Factor-positive Sjögren's disease (RF-positive SS), and Rheumatoid Factor-negative Sjögren's disease (RF-negative SS). For antibodies where the overall ANOVA was significant (anti-dsDNA), Tukey's HSD test was used for post-hoc pairwise comparisons to identify which specific groups differed. To ensure minimal errors, the significance of the ANOVA results was also evaluated using both Holm and False Discovery Rate (FDR) corrections for multiple comparisons.

Table 4: One-way ANOVA test results.

Antibody	F	p-value	Significant
ACPA	0.201	0.818	No
ANA	2.507	0.085	No
Anti-dsDNA	4.577	0.012	Yes
Anti-Sm	0.228	0.796	No

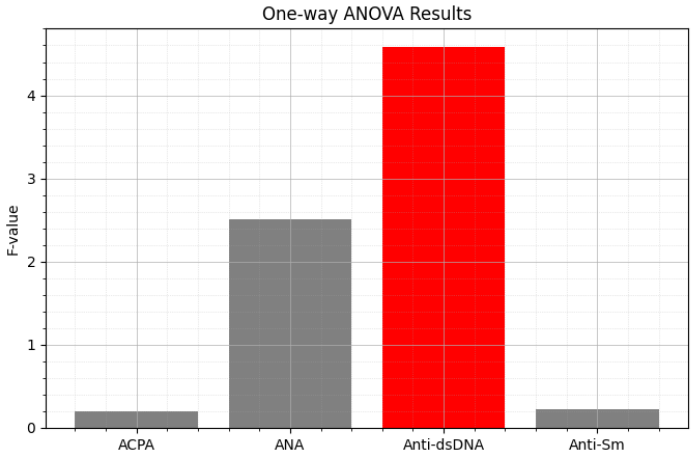


Figure 5: **One-way ANOVA results.** This bar plot displays F-statistics from one-way ANOVA tests for ACPA, ANA, Anti-dsDNA, and Anti-Sm.

Tukey's HSD was performed as a post hoc test for the one-way ANOVA on anti-dsDNA levels. No significant difference was found between RA and RF-positive SS (mean difference = -0.0361, $p = 0.923$). RA and RF- SS differed significantly (mean difference = 0.2399, $p = 0.019$), as did RF-positive SS compared to RF- SS (mean difference = 0.276, $p = 0.029$).

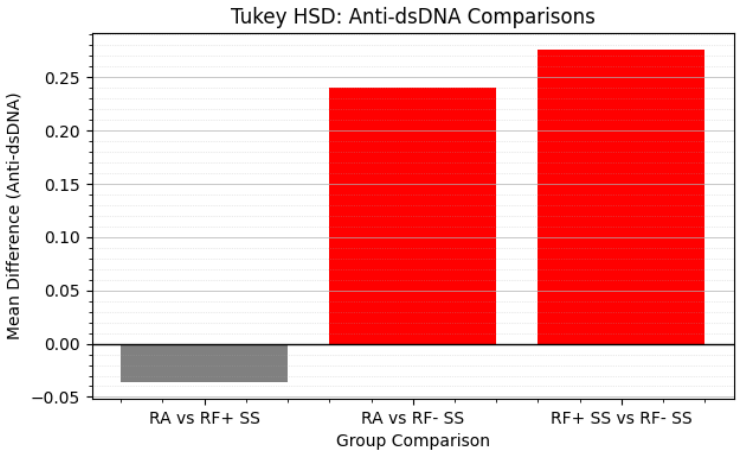


Figure 6: **Tukey's HSD post-hoc comparisons for anti-dsDNA levels.** Bar plot showing mean differences in anti-dsDNA levels for RA vs. RF-positive SS, RA vs. RF-negative SS, and RF-positive SS vs. RF-negative SS. Bars are color-coded by significance (red = significant, gray = non-significant). Positive differences indicate higher mean values in the first group of the comparison; negative values indicate higher levels in the second group.

After running ANOVA testing, we also implemented FDR and Holm-Bonferroni corrections for multivariate testing. The implementation of multivariate testing minimized false positives. Testing was done to confirm the results obtained from both the raw one-way ANOVA reaults and Tukey's post hoc.

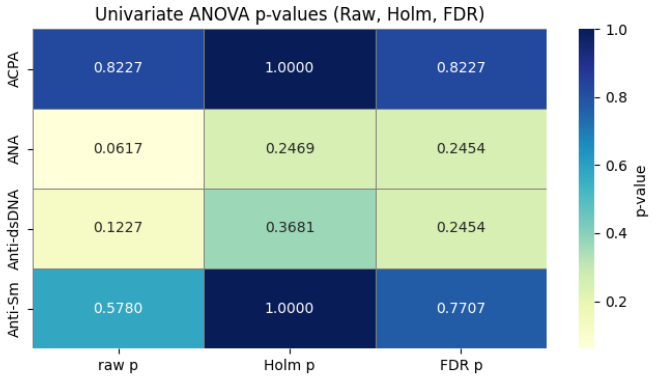


Figure 7: **Univariate ANOVA p-values (Raw, Holm, FDR)** Heatmap displaying all values for ANOVA corrections, color gradient represents p-value, where lighter is closer to zero and darker is closer to 1.

3.6. MANOVA

Multivariate analysis of variance (MANOVA) tested whether the combined autoantibody profiles (ACPA, ANA, anti-dsDNA, anti-Sm) differed across the diagnostic groups (RA, RF-positive SS, RF-negative SS).

Table 5: Multivariate Analysis of Variance (MANOVA) Results for Autoantibody Profiles Across RA and SS Subgroups

Model Component	Test	Value	Num DF	Den DF	F Value	Pr > F
Intercept	Wilks' lambda	0.2707	4	171	115.1665	0.0000
Intercept	Pillai's trace	0.7293	4	171	115.1665	0.0000
Intercept	Hotelling-Lawley trace	2.6940	4	171	115.1665	0.0000
Intercept	Roy's greatest root	2.6940	4	171	115.1665	0.0000
Diagnosis_simple	Wilks' lambda	0.9689	4	171	1.3729	0.2453
Diagnosis_simple	Pillai's trace	0.0311	4	171	1.3729	0.2453
Diagnosis_simple	Hotelling-Lawley trace	0.0321	4	171	1.3729	0.2453
Diagnosis_simple	Roy's greatest root	0.0321	4	171	1.3729	0.2453

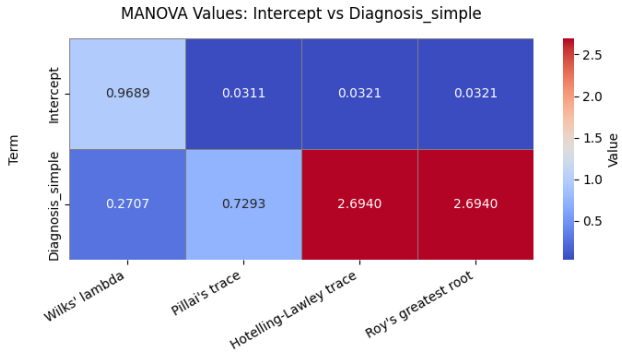


Figure 8: **Heatmap of MANOVA values.** Heatmap showing MANOVA values, lower values (cooler colors) indicate greater similarity to RA, while higher values (warmer colors) indicate more distinct antibody profiles.

3.7. Euclidean Distance Analysis

We calculated Euclidean distances between Z-normalized autoantibody centroids for RA vs RF-positive SS and RA vs RF-negative SS. This provided a direct measure of overall similarity or dissimilarity in antibody profiles between groups.

Table 6: Euclidean Distances Between Z-Normalized Autoantibody Centroids of RA and SS Subgroups

Comparison	Distance
RF-positive SS vs RA	0.1641
RF-negative SS vs RA	0.6443

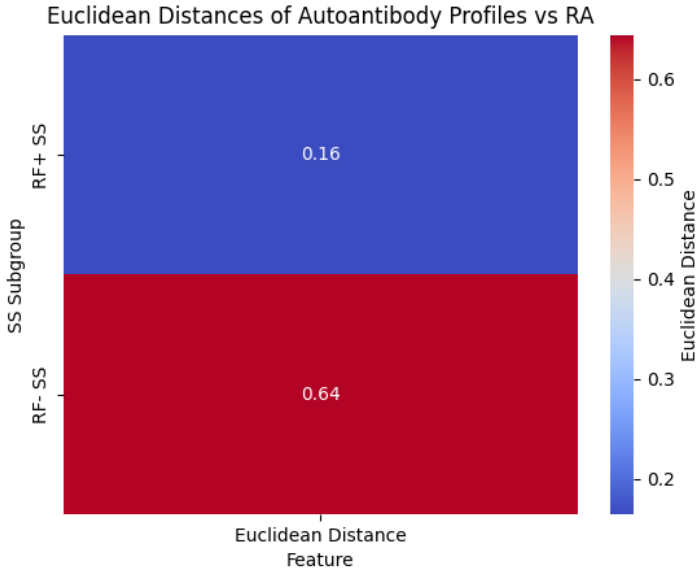


Figure 9: **Euclidean distances between rheumatoid arthritis and Sjögren's syndrome subgroups based on autoantibody profiles.** Heatmap showing Euclidean distances between RA and RF-positive SS, and RA and RF-negative SS after Z-normalization of autoantibody levels. Lower values (cooler colors) indicate greater similarity to RA, while higher values (warmer colors) indicate more distinct antibody profiles.

3.8. Logistic Regression

We applied logistic regression to classify RA and SS cases based on their autoantibody profiles. This test tested the diagnosis based off of the RF factor identity.

Table 7: Logistic Regression Classification Performance for Distinguishing RA and SS Subgroups Based on Autoantibody Profiles

Class	Precision	Recall	F1-Score	Support
RA	0.645	0.659	0.652	91
SS	0.627	0.612	0.619	85
Accuracy			0.636	176
Macro Avg	0.636	0.636	0.636	176
Weighted Avg	0.636	0.636	0.636	176

Figure 10: **Bar Plot of Logistic Regression Classification Metrics for Distinguishing RA and SS** This bar plot presents model performance across precision (green), recall (blue), and F1-score (yellow) for RA and SS subgroups. The dotted horizontal line indicates overall model accuracy.

4. Discussion

4.1. Summary of Findings

Across multiple analyses (t-tests, χ^2 tests, ANOVA with Tukey's HSD, MANOVA, Holm/FDR corrections, logistic regression, and Euclidean distances), RF-positive SS displayed moderate serological similarity to RA than RF-negative SS in this synthetic dataset. RF-negative SS showed higher ANA and anti-dsDNA relative to RA, a pattern consistent with (but not proof of) SLE-like serology (Damoiseaux and Van Beers, 2023; Liu et al., 2024). Chi-squared tests indicated that antibody positivity alone was not sufficient to distinguish RA from SS, while Holm/FDR corrections removed the apparent ANOVA significance for anti-dsDNA, suggesting that these effects may be exploratory rather than confirmatory. Logistic regression showed only modest classification performance, further supporting the view that antibodies contribute to group differentiation but cannot serve as a standalone classifier. Together, these findings are hypothesis-generating and suggest that RF status may stratify SS into subgroups with distinct immunological profiles. However, none of the results obtained from the dataset should be used as concrete evidence of association.

4.2. Interpretation of Findings

T-tests and the subsequent Cohen's d effect sizes suggested that RF-positive SS patients shared similarities to RA serological profiles. Specifically, the similar prevalence levels of ACPA ($t = -0.194$, Cohen's d = 0.038), an antibody with high specificity for RA, supported this potential overlap (p = 0.846). Other antibodies showed smaller or negligible differences between RF-positive SS and RA (ANA: t = 0.691, d = -0.131; anti-dsDNA: t = -0.372, d = 0.072; anti-Sm: t = -0.194, d = 0.038). In contrast, RF-negative SS diverged more strongly from RA in several antibody measures, particularly ANA ($t = 2.424$, d = -0.404) and anti-dsDNA ($t = 2.842$, d = -0.494), both showing moderate-to-large effect sizes.

The chi-squared tests for independence further showed that the distribution of antibody positivity did not differ significantly between diagnostic groups (ACPA: $\chi^2 = 0.005$, p = 0.942; ANA: $\chi^2 = 2.910$, p = 0.088; anti-dsDNA: $\chi^2 = 1.954$, p = 0.162; anti-Sm: $\chi^2 = 0.167$, p = 0.683). This suggests that diagnosis (RA vs SS) cannot be fully explained by the presence or absence of specific antibodies alone.

Univariate one-way ANOVA supported this interpretation. Most antibodies were not significantly different between groups, except for anti-dsDNA (F = 4.577, p = 0.012), which initially appeared significant. Tukey's post hoc test clarified this further: RF-negative SS patients differed significantly from both RA and RF-positive SS, while RA and RF-positive SS remained indistinguishable. However, after correction for multiple comparisons (Holm and FDR), the anti-dsDNA effect was no longer statistically significant, suggesting that the result may represent a false positive in the context of multiple hypothesis testing. This highlights the exploratory nature of the findings, where certain pairwise contrasts appear promising but require validation in clinical conditions.

At the multivariate level, MANOVA did not reveal significant differences in overall antibody profiles between diagnostic groups (Wilks' $\lambda = 0.969$, F(4,171) = 1.373, p = 0.245). This indicates that when all antibodies are considered simultaneously, the distinction between RA and SS does not reach statistical significance.

Predictive modeling provided an additional perspective. Logistic regression using antibody data yielded a modest overall accuracy of 63.6%, with balanced performance across RA (precision = 0.645, recall = 0.659) and SS (precision = 0.627, recall = 0.612). This reinforces the idea that while

antibodies contribute to differentiation, they cannot serve as a standalone diagnostic classifier with high reliability.

Euclidean distance analysis further quantified these relationships. In the autoantibody profile space, RF-positive SS was closer to RA (distance = 0.1641) than RF-negative SS was to RA (distance = 0.6443). This suggests that the serological overlap between RF-positive SS and RA extends into broader clinical characteristics, whereas RF-negative SS may represent a more distinct subgroup.

4.3. Diagnostic Implications

The findings from the synthetic dataset suggest that RF status may define clinical subtypes within Sjögren's syndrome (SS). The one-way ANOVA and t-tests showed that RF-negative SS patients had higher ANA and anti-dsDNA levels, aligning with a "classic" SS-dominant subtype characterized by stronger ANA and anti-dsDNA presence, and potentially aligning more closely with systemic lupus erythematosus (SLE). However, the ANOVA value is only confirmed in Tukey's post hoc, and dissproved in further corrections. This pattern is supported by existing literature, though the loss of significance after Holm/FDR corrections in the univariate ANOVA indicates these findings are exploratory and require further validation. This raises the possibility that RF-negative SS shares an immunological link with SLE, though further targeted comparisons with SLE cohorts would be needed to confirm this relationship.

In contrast, RF-positive SS demonstrates features intermediate between SS and rheumatoid arthritis (RA), consistent with a potential RA-overlap subtype. The t-tests showed serological similarity between RF-positive SS and RA, particularly for ACPA, while the one-way ANOVA and MANOVA found no significant differences in antibody profiles, reinforcing this overlap. The Euclidean distance analysis further confirmed that RF-positive SS is serologically closer to RA than RF-negative SS. Recognizing these immunological subgroups could enable more precise, tailored therapeutic strategies, particularly for RF-positive SS patients, who may benefit from early intervention with RA-targeted treatments. The logistic regression analysis showed modest classification performance, indicating that while antibodies contribute to differentiating RA and SS, they are not sufficient alone, supporting the inclusion of RF and ACPA in a broader diagnostic panel. This could also reduce the risk of misdiagnosis, minimize unnecessary treatment side effects, and help anticipate disease exacerbations.

From a diagnostic standpoint, these results highlight the need to revisit current classification frameworks. Existing SS criteria (e.g., ACR or EULAR) place emphasis on ANA and SSA/SSB antibodies, alongside hallmark symptoms, but RF is often considered only a supportive marker. The t-tests and euclidean distance testing suggest that RF positivity may play a more critical role in identifying RA-like SS subtypes, particularly when combined with other serological markers. Our findings suggest RF positivity in SS may potentially have the need to be weighted more heavily, however keeping in mind these results stemmed from a synthetic dataset generated from peer reviewed datasets.

Additionally, incorporating ACPA testing into the SS diagnostic process may further aid in identifying patients with RA-like pathology. The t-tests and chi-squared tests showed that ACPA levels are similar between RF-positive SS and RA, highlighting its potential as a marker for RA-overlap in SS. As autoimmune disease research advances through biomarker profiling, updated classification systems could benefit from integrating serological subtyping, particularly RF and ACPA status, for patients presenting with ambiguous or early-stage symptoms. The exploratory nature of these findings, driven by the loss of significance after multiple testing corrections and modest predictive performance in logistic regression, underscores the need for prospective studies to validate the diagnostic utility of RF and ACPA. Such refinements may improve diagnostic accuracy, inform prognosis, and guide more personalized management approaches.

These observations are hypothesis-generating and motivate prospective studies to test whether RF (with ACPA) improves subtype classification or prediction of RA overlap in real-world cohorts.

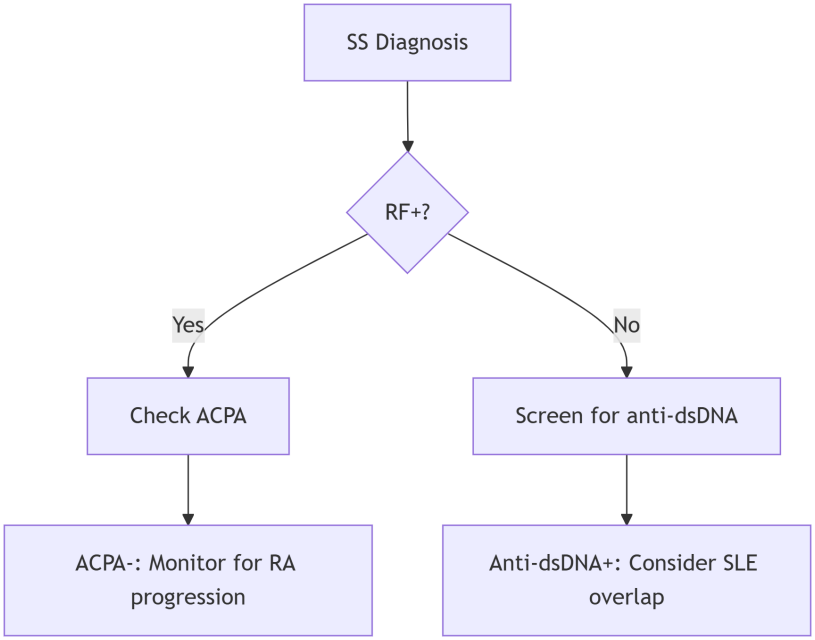


Figure 11: **Conceptual decision pathway (hypothesis-generating)** Based on findings from the synthetic dataset, RF positivity in SS may indicate a subgroup with greater serological similarity to RA, as supported by t-tests showing overlap in ACPA levels, and one-way ANOVA and MANOVA confirming minimal differences in antibody profiles, with Euclidean distance analysis indicating closer proximity to RA. This suggests consideration of RA-targeted evaluation or management. RF-negative SS may reflect a "classic" SS phenotype with possible systemic lupus erythematosus (SLE)-like serology, as indicated by one-way ANOVA and t-tests showing higher ANA and anti-dsDNA levels, though the loss of significance after Holm/FDR corrections highlights the exploratory nature of this finding. The moderate-low performance of logistic regression further suggests that RF and ACPA, combined with other markers, may aid in identifying these subgroups. This pathway is intended for research purposes, specifically hypothysis generating only, and should be validated in real patient cohorts before any clinical application.

4.4. Limitations

This study relies on a single, literature-modeled synthetic dataset, which approximates, but does not replicate real patient heterogeneity; results are hypothesis-generating. Antibody and symptom variables were binary, precluding analysis of titers or severity. The synthetic dataset lacks clinical variables such as disease duration or treatment history, which may influence serological profiles and limit the applicability of findings to real-world diagnostic scenarios. The cohort size (n = 176) may limit generalizability.

SSA/SSB and RNP were unavailable, constraining SS-specific inference.

Because RF was used to stratify SS subgroups, some observed associations with RA-linked markers may reflect expected correlation structures embedded in the synthetic data (potential circularity). The discrepancy in Anti-dsDNA significance between one-way ANOVA and univariate ANOVA with corrections suggests potential inconsistencies in statistical modeling or data handling, which may affect the reliability of this biomarker's role in distinguishing groups. Additionally, the non-significant MANOVA result may reflect limited statistical power due to the modest sample size or correlations among biomarkers, potentially masking multivariate differences between RA and Sjögren's syndrome subgroups. Finally, while ANOVA post hoc comparisons used Tukey's HSD, multiple pairwise tests may inflate Type I error; future work should consider false discovery rate control or multivariable modeling (e.g., logistic regression). Additionally, there was only 1 positive control set, being the RA patients. Whereas, bringing in patients from SLE could help pinpoint the disease subtype even more, potentially stratifying RF-negative towards SLE. Furthermore, there were not normal (healthy) patient controls within this study, however the justification for this is that the study only aimed to compare SS patients to RA patients and the potential of RF stratifying SS into 2 subgroups. Therefore, the implications that unhealthy patients needed to be considered doesn't hold much prevalence within the aims of this study.

However, given the constraints of the dataset, the values obtained present hypothesis generating ideas. These ideas serve to progress understanding of autoimmune diseases, rather than provide concrete evidence of assocation. While the results may fall short in real world diagnostic situations, they provide insight into potential links, that can be explored in clinical settings.

4.5. Future directions

There are a multitude of future directions that can be explored after the results of this exploratory study. Firstly, the expansion of this exploratory design towards a set of real patients would provide more concrete evidence towards the potential of a SS subgroup. Prospective studies in real patient cohorts should test whether RF-based stratification predicts RA overlap, clinical outcomes, or therapeutic response in SS. Inclusion of SLE and healthy controls, comprehensive autoantibody panels (SSA/SSB, RNP), and quantitative titers will enable finer discrimination.

Incorporating hematologic and clinical severity scales and conducting longitudinal assessments can relate serology to disease courses.

Integrative multi-omics (proteomics/transcriptomics) and multivariable modeling (e.g., penalized logistic regression allowing for reverse diagnosis based on hypothesis generation, a more robust form of analysis completed withint this study) could identify biomarker signatures distinguishing RF-positive from RF-negative SS. External validation across datasets will be critical to establish robustness and translational utility.

5. Conclusion

In this synthetic dataset, RF-positive SS exhibits greater serological similarity to RA than RF-negative SS, which shows higher ANA/anti-dsDNA (Ragheb, 2024). These findings are hypothesis-generating and support the rationale for RF-informed stratification in future SS research. If validated in clinical cohorts, RF (with ACPA) could help identify patients at risk for RA-like disease trajectories and guide serology-informed, personalized approaches to diagnosis and monitoring.

Ethics Statement

This study used only synthetic, de-identified data; no ethical approval was required (Ragheb, 2024).

CRediT authorship contribution statement

Nathan Zhuang: Writing - review & editing, Writing - original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Jacqueline Howells: Writing – review & editing, Supervision, Project administration.

Competing interests

This work was developed as part of the Polygence mentorship program. Polygence provided guidance on research design but had no role in data collection/analysis or manuscript.

Acknowledgments

The author's thank Polygence for providing a platform and support to conduct this research, as well as guidance throughout the project.

References

Aickin, M., Gensler, H., 1996. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. American Journal of Public Health 86, 726–728. URL: https://pubmed.ncbi.nlm.nih.gov/8629727/, doi:10.2105/ajph.86.5.726.

Aletaha, D., Neogi, T., Silman, A.J., Funovits, J., Felson, D.T., Bingham, C.O., Birnbaum, N.S., Burmester, G.R., Bykerk, V.P., Cohen, M.D., Combe, B., Costenbader, K.H., Dougados, M., Emery, P., Ferraccioli, G., Hazes, J.M.W., Hobbs, K., Huizinga, T.W.J., Kavanaugh, A., Kay, J., Kvien, T.K., Laing, T., Mease, P., Ménard, H.A., Moreland, L.W., Naden, R.L., Pincus, T., Smolen, J.S., Stanislawska-Biernat, E., Symmons, D., Tak, P.P., Upchurch, K.S., Vencovský, J., Wolfe, F., Hawker, G., 2010. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis & Rheumatism 62, 2569–2581. URL: https://doi.org/10.1002/art.27584, doi:10.1002/art.27584.

Bournazos, S., Gupta, A., Ravetch, J.V., 2020. The role of IgG Fc receptors in antibody-dependent enhancement. Nature reviews. Immunology 20, 633–643. URL: https://www.nature.com/articles/s41577-020-00410-0?, doi:10.1038/s41577-020-00410-0.

Brito-Zerón, P., Baldini, C., Bootsma, H., Bowman, S.J., Jonsson, R., Mariette, X., Sivils, K., Theander, E., Tzioufas, A., Ramos-Casals, M., 2016. Sjögren syndrome. Nature Reviews Disease Primers 2. URL: https://pubmed.ncbi.nlm.nih.gov/27383445/, doi:10.1038/nrdp.2016.47.

Chatzi, A., Doody, O., 2023. The one-way ANOVA test explained. Nurse Researcher 31, 8–14. URL: https://pubmed.ncbi.nlm.nih.gov/37317616/, doi:10.7748/nr.2023.e1885.

Damoiseaux, J., Van Beers, J., 2023. Autoantibodies to dsDNA in the diagnosis, classification and follow-up of patients with systemic lupus erythematosus. Journal of Translational Autoimmunity 6, 100191. URL:

https://pmc.ncbi.nlm.nih.gov/articles/PMC9883238/, doi:10.1016/j.jtauto.2023.100191.

Delacre, M., Lakens, D., Leys, C., 2017. Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. International review of social psychology 30, 92–101. URL: https://doi.org/10.5334/irsp.82, doi:10.5334/irsp.82.

Elagib, K.E.E., Børretzen, M., Jonsson, R., Haga, H.J., Thoen, J., Thompson, K.M., Natvig, J.B., 1999. Rheumatoid factors in primary Sjögren's syndrome (pSS) use diverse VH region genes, the majority of which show no evidence of somatic hypermutation. Clinical & Experimental Immunology 117, 388–394. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1905326, doi:10.1046/j.1365-2249.1999.00963.x.

Feldman, M., Ginsburg, I., 2021. A novel hypothetical approach to explain the mechanisms of pathogenicity of rheumatic arthritis. Mediterranean Journal of Rheumatology 32, 112. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8369279/, doi:10.31138/mjr.32.2.112.

Gaddis, M.L., 1998. Statistical Methodology: IV. Analysis of Variance, Analysis of Co variance, and Multivariate Analysis of Variance. Academic Emergency Medicine 5, 258–265. URL: https://pubmed.ncbi.nlm.nih.gov/9523936/, doi:10.1111/j.1553-2712.1998.tb02624.x.

Glazko, G., Mushegian, A., 2010. Measuring gene expression divergence: the distance to keep. Biology Direct 5, 51. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC2928186/, doi:10.1186/1745-6150-5-51.

Ingegnoli, F., Castelli, R., Gualtierotti, R., 2013. Rheumatoid factors: clinical applications. Disease Markers 35, 727–734. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3845430/, doi:10.1155/2013/726598.

Kim, H., Cho, S.K., Kim, H.W., Han, J., Kim, Y., Hwang, K.G., Sung, Y.K., 2020. The prevalence of Sjögren's syndrome in rheumatoid arthritis patients and their clinical features. Journal of Korean Medical Science 35. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7683240/, doi:10.3346/jkms.2020.35.e369.

Kim, H.Y., 2015. Statistical notes for clinical researchers:post-hocmultiple comparisons. Restorative Dentistry & Endodontics 40, 172. URL:

https://pmc.ncbi.nlm.nih.gov/articles/PMC4432262/, doi:10.5395/rde.2015.40.2.172.

Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E.J., Hicks, S.C., 2019. A practical guide to methods controlling false discoveries in computational biology. Genome biology 20. URL: https://pubmed.ncbi.nlm.nih.gov/31164141/, doi:10.1186/s13059-019-1716-1.

Kurowska, W., Kuca-Warnawin, E.H., Radzikowska, A., Maśliński, W., 2017. The role of anti-citrullinated protein antibodies (ACPA) in the pathogenesis of rheumatoid arthritis. Central European Journal of Immunology 42, 390–398. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC5820977/, doi:10.5114/ceji.2017.72807.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology 4. URL: https://doi.org/10.3389/fpsyg.2013.00863, doi:10.3389/fpsyg.2013.00863.

Liu, F., Wang, X.Q., Zou, J.W., Li, M., Pan, C.C., Si, Y.Q., 2024. Association between serum antinuclear antibody and rheumatoid arthritis. Frontiers in Immunology 15. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11070521/, doi:10.3389/fimmu.2024.1358114.

McHugh, M.L., 2013. The Chi-square test of independence. Biochemia Medica , 143–149URL: https://pubmed.ncbi.nlm.nih.gov/23894860/, doi:10.11613/bm.2013.018.

Nashi, R.A., Shmerling, R.H., 2021. Antinuclear antibody testing for the diagnosis of systemic lupus erythematosus. Medical Clinics of North America 105, 387–396. URL: https://pubmed.ncbi.nlm.nih.gov/33589110/, doi:10.1016/j.mcna.2020.10.003.

Ragheb, A., 2024. Comprehensive Autoimmune Disorder Dataset. URL: https://www.kaggle.com/datasets/abdullahragheb/all-autoimmune-disorder-10k/data.

Sperandei, S., 2014. Understanding logistic regression analysis. Biochemia Medica , 12–18URL: https://pubmed.ncbi.nlm.nih.gov/24627710/, doi:10.11613/bm.2014.003.

Van Beers, J.J., Schreurs, M.W., 2022. Anti-Sm antibodies in the classification criteria of systemic lupus erythematosus. Journal of Translational Autoimmunity 5, 100155. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9026971/, doi:10.1016/j.jtauto.2022.100155.

Zhuang, N., 2025. GitHub - NathanZhuangCalgary/SSRAAntibodyAnalysis: Code and data processing scripts for the analysis of autoantibody profiles in Sjögren's syndrome and rheumatoid arthritis (synthetic dataset, N=176). URL: https://github.com/NathanZhuangCalgary/SSRAAntibodyAnalysis.

Figures

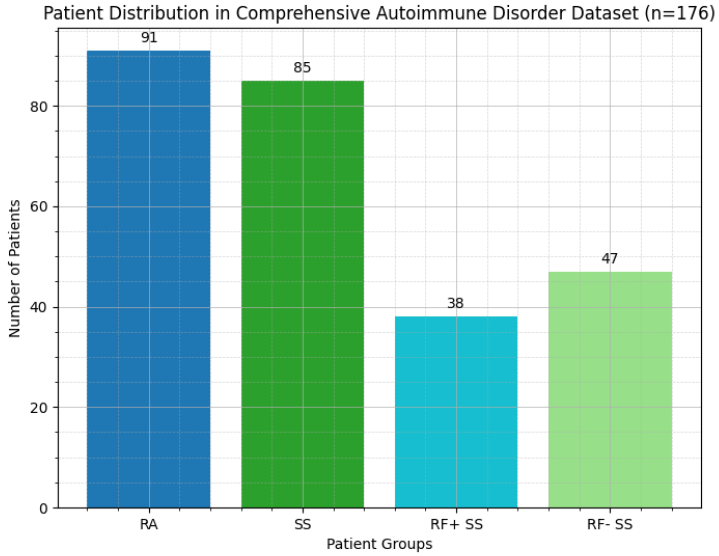


Figure 1: **Patient Distribution in Comprehensive Autoimmune Disorder Dataset (n=176).** Bar plot showing the numbers of id's of each subgroup analyzed, extrapolated from the dataset.

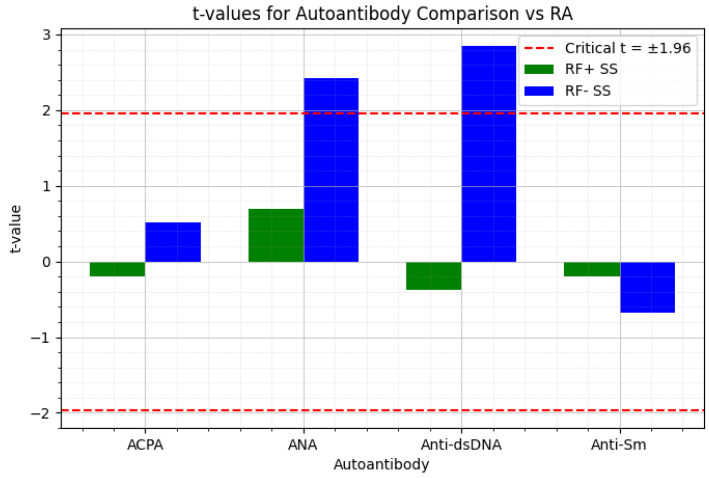


Figure 2: **T-test statistics for autoantibody levels in RF-positive and RF-negative Sjögren's syndrome compared with rheumatoid arthritis.** Bar plot showing two-group t-values for ACPA, ANA, anti-dsDNA, and anti-Sm. RF-positive SS values are shown in green; RF-negative SS values in blue. Dashed horizontal lines indicate the critical t-value threshold ($\alpha = 0.05$) for statistical significance. Positive t-values indicate higher mean levels in Sjögren's groups, while negative values indicate higher levels in RA.

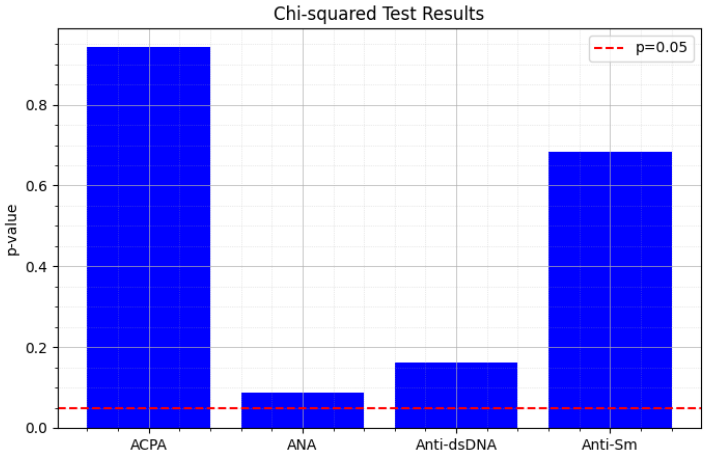


Figure 3: **Bar Plot of Chi-Squared Test Results for Autoantibody Frequency Distributions.** This bar plot illustrates chi-squared values for categorical comparisons of ACPA, ANA, Anti-dsDNA, and Anti-Sm frequencies across RA and SS groups. The relative bar heights highlight the magnitude of association for each autoantibody, providing a visual summary of distributional differences between diagnostic categories.

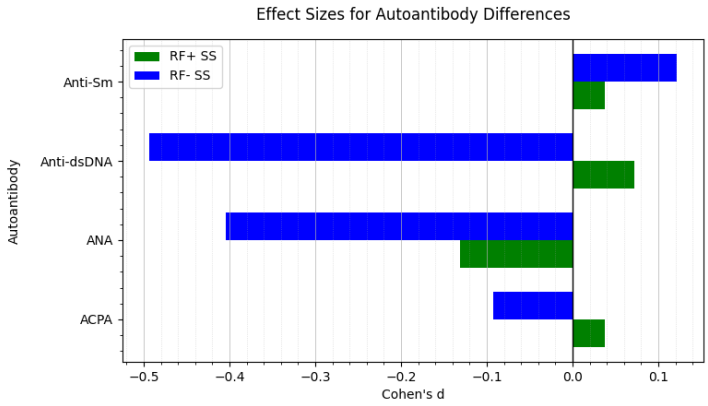


Figure 4: **Cohen's d effect sizes for autoantibody differences between Sjögren's subgroups and rheumatoid arthritis.** Horizontal bar chart of effect sizes (d) for ACPA, ANA, anti-dsDNA, and anti-Sm in RF-positive and RF-negative Sjögren's syndrome vs. rheumatoid arthritis. Positive values indicate higher antibody levels in the Sjögren's group, negative values indicate higher levels in RA. Vertical gridlines mark small (0.2), medium (0.5), and large (0.8) effect size thresholds.

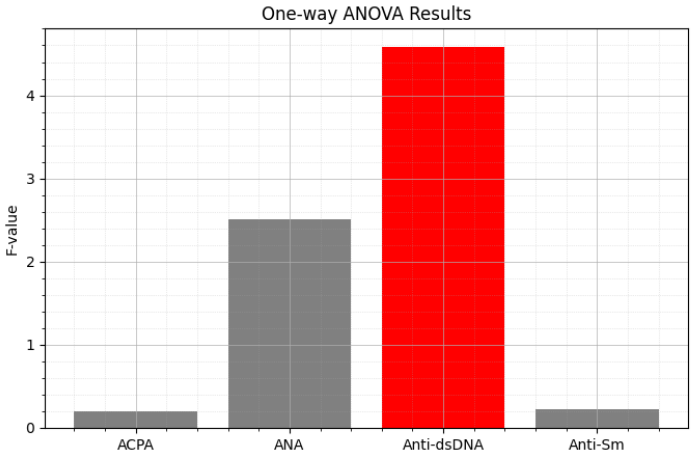


Figure 5: **One-way ANOVA results.** This bar plot displays F-statistics from one-way ANOVA tests for ACPA, ANA, Anti-dsDNA, and Anti-Sm.

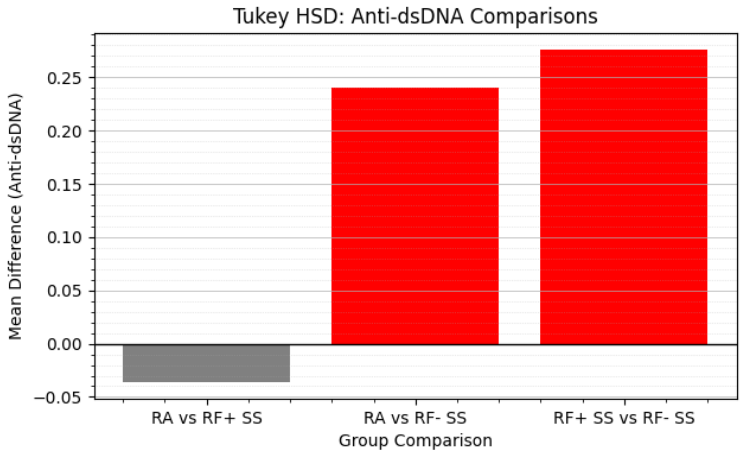


Figure 6: **Tukey's HSD post-hoc comparisons for anti-dsDNA levels.** Bar plot showing mean differences in anti-dsDNA levels for RA vs. RF-positive SS, RA vs. RF-negative SS, and RF-positive SS vs. RF-negative SS. Bars are color-coded by significance (red = significant, gray = non-significant). Positive differences indicate higher mean values in the first group of the comparison; negative values indicate higher levels in the second group.

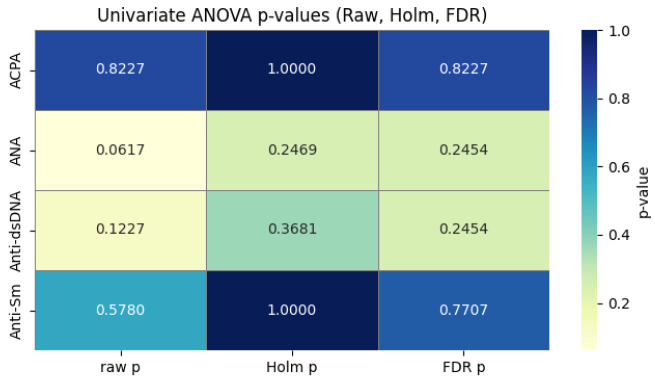


Figure 7: **Univariate ANOVA p-values (Raw, Holm, FDR)** Heatmap displaying all values for ANOVA corrections, color gradient represents p-value, where lighter is closer to zero and darker is closer to 1.

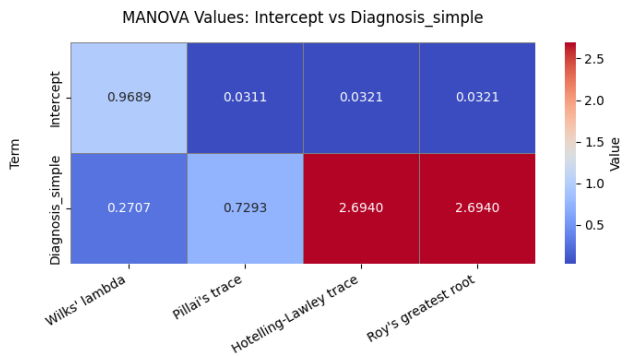


Figure 8: **Heatmap of MANOVA values.** Heatmap showing MANOVA values, lower values (cooler colors) indicate greater similarity to RA, while higher values (warmer colors) indicate more distinct antibody profiles.

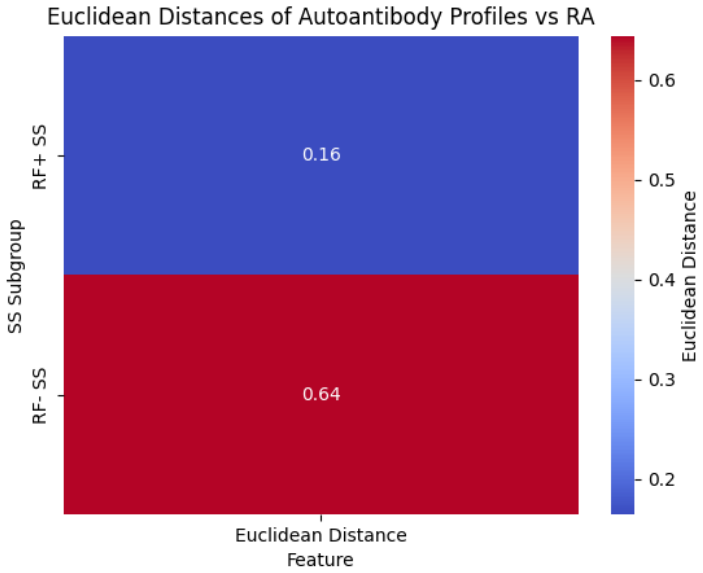


Figure 9: **Euclidean distances between rheumatoid arthritis and Sjögren's syndrome subgroups based on autoantibody profiles.** Heatmap showing Euclidean distances between RA and RF-positive SS, and RA and RF-negative SS after Z-normalization of autoantibody levels. Lower values (cooler colors) indicate greater similarity to RA, while higher values (warmer colors) indicate more distinct antibody profiles.

Figure 10: **Bar Plot of Logistic Regression Classification Metrics for Distinguishing RA and SS** This bar plot presents model performance across precision (green), recall (blue), and F1-score (yellow) for RA and SS subgroups. The dotted horizontal line indicates overall model accuracy.

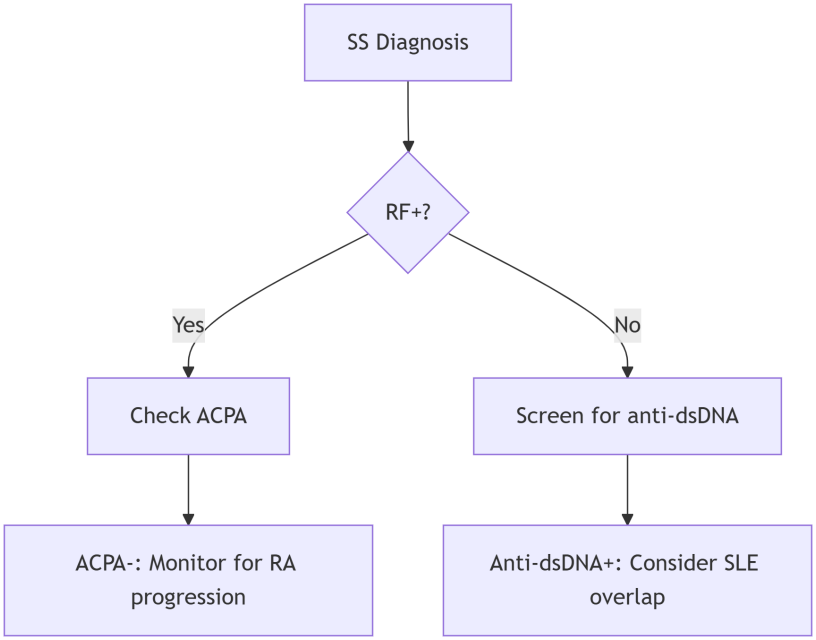


Figure 11: **Conceptual decision pathway (hypothesis-generating)** Based on findings from the synthetic dataset, RF positivity in SS may indicate a subgroup with greater serological similarity to RA, as supported by t-tests showing overlap in ACPA levels, and one-way ANOVA and MANOVA confirming minimal differences in antibody profiles, with Euclidean distance analysis indicating closer proximity to RA. This suggests consideration of RA-targeted evaluation or management. RF-negative SS may reflect a "classic" SS phenotype with possible systemic lupus erythematosus (SLE)-like serology, as indicated by one-way ANOVA and t-tests showing higher ANA and anti-dsDNA levels, though the loss of significance after Holm/FDR corrections highlights the exploratory nature of this finding. The moderate-low performance of logistic regression further suggests that RF and ACPA, combined with other markers, may aid in identifying these subgroups. This pathway is intended for research purposes, specifically hypothysis generating only, and should be validated in real patient cohorts before any clinical application.