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

Nathan Z. Zhuang*[1,2] and Jacqueline M. Howells, PhD[1,3]

[1]Polygence, Menlo Park, California, United States of America
[2]Western Canada High School, Calgary, Alberta, Canada
[3]Brown University, Providence, Rhode Island, United States of America

September 11, 2025

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 possible overlap 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, modeled on real-world clinical trends and comprised of de-identified patient data (Comprehensive Autoimmune Disorder Dataset; synthetic; n=176), including 85 SS patients (38 RF-positive, 47 RF-negative) and 91 RA patients. Four autoantibodies: anti-citrullinated protein antibodies (ACPA), antinuclear antibodies (ANA), anti-double-stranded DNA (anti-dsDNA), and anti-Smith (anti-Sm). These variables were compared across groups using Welch's t-tests including Cohen's d Effect Zoning, chi-squared tests, one-way ANOVA with Tukey's post hoc, MANOVA, and Euclidean distance analysis of Z-normalized antibody profiles. RF-positive SS showed no significant differences from RA in ACPA prevalence (t=-0.194, p=0.847), whereas RF-negative SS differed significantly from RA in ANA (t=2.424, p=0.017) and anti-dsDNA (t=2.842, p=0.005). ANOVA and Tukey's post hoc tests revealed higher anti-dsDNA in RF-negative SS versus RA (p=0.019) and RF-positive SS (p=0.029). MANOVA indicated minimal overall profile differences between RF positive SS and RA (Wilks' $\lambda = 0.969$, p=0.245), but distinct clustering for RF-negative SS. Euclidean distances supported these findings, with RF-positive SS closer to RA (0.1641) than RF-negative SS (0.6443). RF-positive SS demonstrates greater immunological similarity to RA than RF-negative SS, supporting the concept of an RF-associated SS subtype with RA overlap features. RF-negative SS showed stronger ANA and anti-dsDNA associations, consistent with a lupus-like profile. RF testing in SS may improve disease stratification, inform prognosis, and guide tailored therapeutic approaches.

*nathan.zhuangzh@gmail.com
*ORCID iD: Nathan Zhuang: 0009-0009-5930-2875
†ORCID iD: Jacqueline Howells: 0000-0003-2664-5694

All relevant data and code are available at GitHub repository.

Keywords: Rheumatoid arthritis (RA), Sjögren's syndrome (SS), Rheumatoid factor (RF), positive rheumatoid factor (RF-positive), negative rheumatoid factor (RF-negative).

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 dry 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. 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 RF is associated with systemic inflammation, shared autoantibody profiles (e.g., anti-citrullinated protein antibodies [ACPAs], antinuclear antibodies [ANAs]) 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: 1) Determine whether RF-positive SS patients show greater immunological similarity to RA patients than RF-negative SS patients; 2) Compare the prevalence of key autoantibodies across RA, RF-positive SS, and RF-negative SS patients.

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 (Ragheb, 2024). It includes patient-level data across 114 autoimmune disorders, including Sjögren's syndrome (SS) and rheumatoid arthritis (RA), 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; 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 included 176 unique patients: 85 with SS and 91 with RA. Within the SS group, patients were further stratified by rheumatoid factor (RF) status, hereafter described as RF-positive and RF-negative, with a total of 38 RF-positive SS patients and 47 RF-negative SS patients.

From the aggregate data, patient age range is within 45–50 years in the dataset. Age was therefore considered uniform, and sex distribution was roughly balanced across groups. 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 GitHub repository to ensure reproducibility (Zhuang, 2025).

2.2 Variable Selection

2.2.1 Autoantibody Profiles

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

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

Table 1: Key autoantibodies analyzed in this study

Binary coding: 1 = positive/present, 0 = negative/absent. This set includes RA-specific markers (ACPA), broad autoimmune markers (ANA), and markers relevant to 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).

2.2.2 Optional Clinical and Hematologic Variables

Although the original dataset includes clinical symptoms (e.g., dry eyes, dry mouth, joint pain, fatigue) and hematologic markers (e.g., hemoglobin, mean corpuscular hemoglobin, reticulocyte count), the present analysis focuses on antibody profiles (Ragheb, 2024). Other variables may be included in future studies examining clinical correlation.

2.3 Statistical Analysis

All statistical analyses were conducted in Python 3.9, using pandas for data manipulation, SciPy for hypothesis testing, NumPy for numerical operations, statsmodels for additional analyses, and seaborn for visualization. All analyses were conducted in Python 3.9 using NumPy v2.2.4, SciPy v1.15.2, Statsmodels v0.14.4, Seaborn v0.13.2, matplotlib v3.10.5, and Pandas v2.2.3.

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, non-parametric test assesses whether observed frequencies deviate from expected frequencies under the null hypothesis of independence (McHugh, 2013).

For comparisons of antibody levels treated as continuous variables, Welch's t-tests were applied. Welch's t-test is robust to unequal variances and small, unequal sample sizes. This makes Welch's t-tests more appropriate than a standard Student's t-test when group variances may differ (Delacre et al., 2017).

2.3.2 Effect Size

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 provides a standardized measure of the difference between

two group means relative to pooled variability, allowing interpretation of practical significance independent of sample size. For categorical variables (e.g., antibody presence/absence) (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 justified as it evaluates whether at least one group mean differs while controlling the family-wise error rate (Chatzi and Doody, 2023). Post hoc Tukey's tests were used to identify specific group pairs driving any significant differences (Kim, 2015).

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. Smaller distances indicate greater similarity. Additionally, Multivariate Analysis of Variance (MANOVA) was conducted to test whether the combined antibody profiles differ across diagnostic groups. MANOVA is appropriate here because it evaluates multiple correlated dependent variables simultaneously, reducing Type I error and capturing multidimensional patterns (Gaddis, 1998).

2.3.5 Significance Thresholds

All tests were two-sided with a significance of $\alpha = 0.05$. Post hoc adjustments were applied where multiple comparisons were performed.

2.4 Reproducibility

To ensure transparency and reproducibility, all preprocessing, statistical analyses, and figure generation scripts are publicly available at GitHub repository to ensure reproducibility (Zhuang, 2025). The synthetic dataset is publicly accessible via Kaggle DOI: 10.34740/KAGGLE/DSV/9807734 . All analyses were performed on the deduplicated dataset, ensuring each patient contributed only a single record. Where applicable, random seeds were set in Python to allow identical replication of results. 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 (Ragheb, 2024).

3 Results

3.1 Overview of Dataset

After filtering for complete records and removing duplicates where necessary, the final dataset included 176 patients, consisting of 91 rheumatoid arthritis (RA) patients and 85 Sjögren's syndrome (SS) patients and. The SS group is being further subdivided by rheumatoid factor (RF) status, there were 38 RF-positive SS patients and 47 RF-negative SS patients. RA patients were 63.1% RF-positive, while RF-negative SS patients, by definition, had 0% RF positivity.

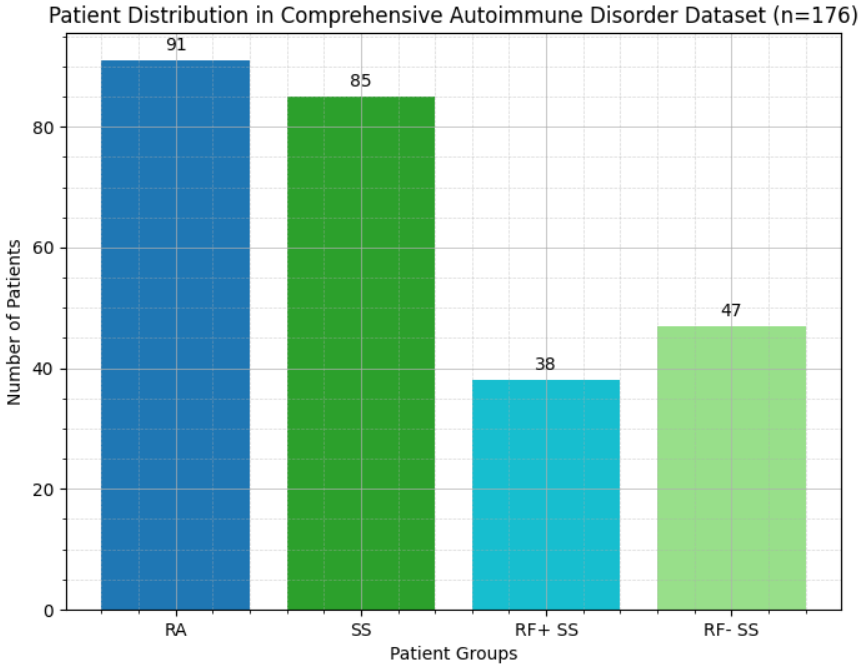

Figure 1: **Patient Distribution in the Comprehensive Autoimmune Disorder Dataset.** The final analytic cohort consisted of 176 patients, including rheumatoid arthritis (RA, n=91) and Sjogren's syndrome (SS, n=85) patients. The SS group was stratified into RF-positive (RF+ SS, n=38) and RF-negative (RF- SS, n=47) subgroups.

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.

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

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

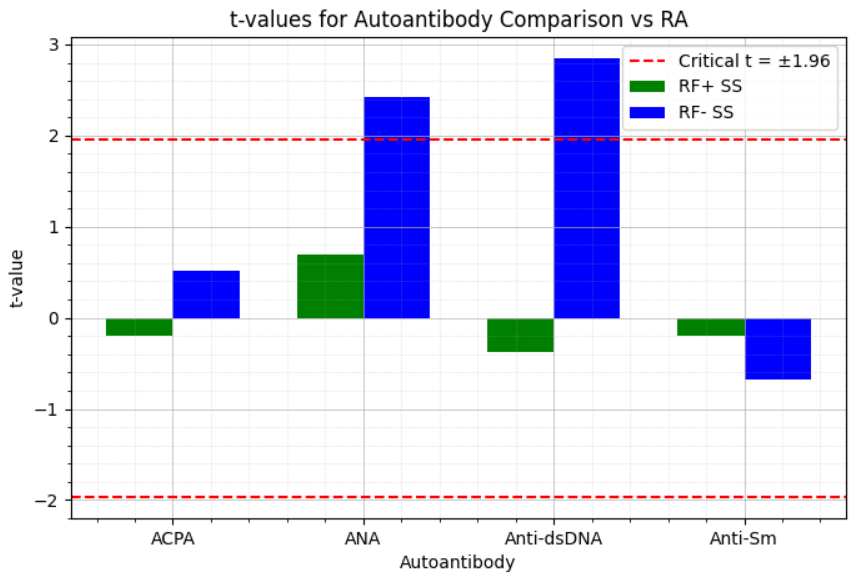

Figure 2: T-test statistics for autoantibody levels in RF-positive and RF– Sjögren's syndrome compared with rheumatoid arthritis. Bar plot showing two-group t-values for ACPA, ANA, anti-dsDNA, and anti-Sm. RF-positive SS values are shown in green; RF– SS values in blue. Dashed horizontal lines indicate 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). This test was appropriate because both variables were categorical.

Antibody	χ^2	p-value	Significant
ACPA	0.005	0.942	No
ANA	2.910	0.088	No
Anti-dsDNA	1.954	0.162	No
Anti-Sm	0.167	0.683	No

Table 3: Results of Chi-squared test for independence

Cohen's d for autoantibodies (RA vs SS subgroups):

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). This was used to complement significance testing with a standardized measure of group separation.

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

Table 4: Cohen's d Effect Sizes Comparing Autoantibody Levels

3.4 One-way ANOVA

One-way ANOVA tested whether mean autoantibody levels differed across the three diagnostic groups (RA, RF-positive SS, RF-negative SS). This was appropriate for continuous antibody measures compared across more than 2 groups. Tukey's HSD was then used for pairwise comparisons to identify which groups specifically differed.

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

Table 5: One-way ANOVA test results

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, $p = 0.029$).

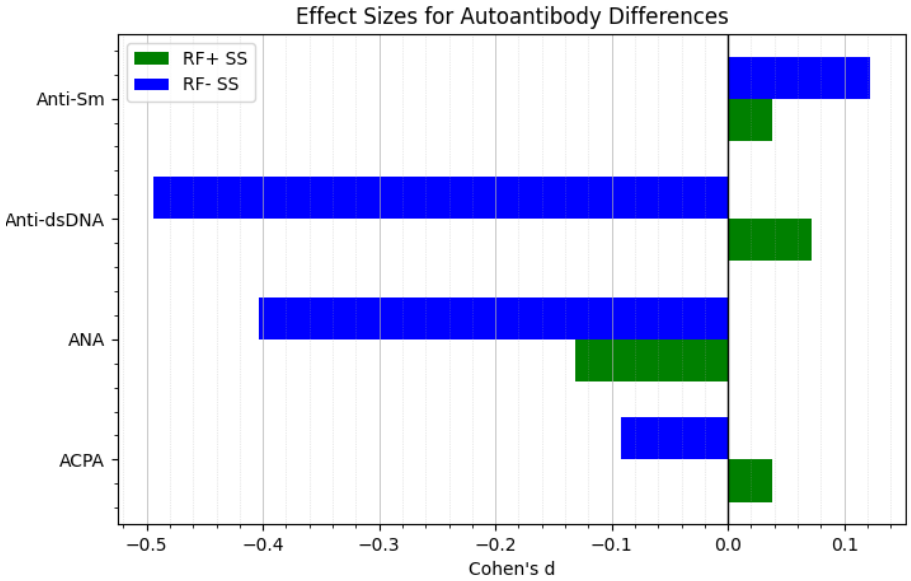

Figure 3: Cohen's d effect sizes for autoantibody differences between Sjögren's subgroups and rheumatoid arthritis. Horizontal bar chart of effect sizes (d) for ACPA, ANA, anti-dsDNA, and anti-Sm in RF-positive and RF– Sjögren's syndrome vs. rheumatoid arthritis. Positive values indicate higher antibody levels in the Sjögren's group, negative values indicate higher levels in RA.

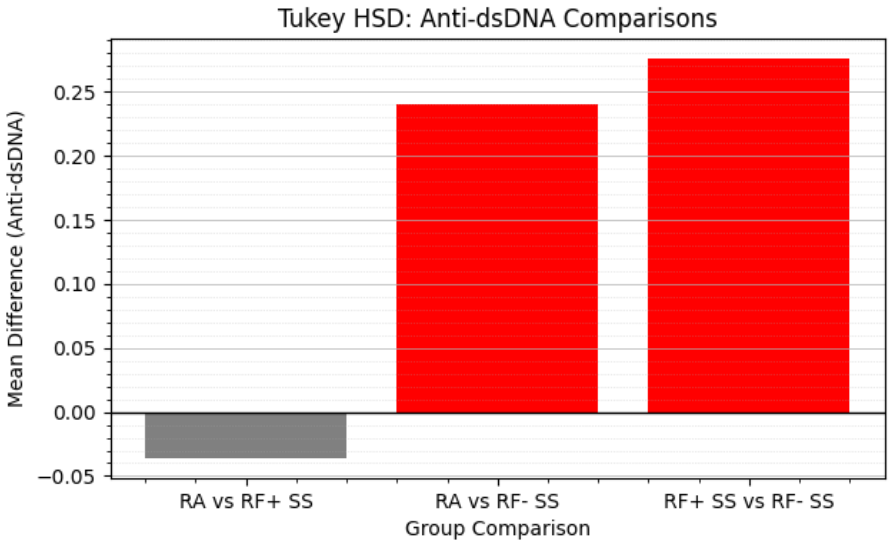

Figure 4: Tukey's HSD post-hoc comparisons for anti-dsDNA levels. Bar plot showing mean differences in anti-dsDNA levels for RA vs. RF-positive SS, RA vs. RF– SS, and RF-positive SS vs. RF– SS. Bars are color-coded by significance (red = significant, gray = non-significant). Positive differences indicate higher mean values in the first group of the comparison; negative values indicate higher levels in the second group.

3.5 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). This allowed us to assess the multivariate structure of antibody expression rather than evaluating each antibody in isolation.

Effect	**Wilks' λ**	**Pillai's trace**	**F Value**	**p-value**
Intercept	0.271	0.729	115.17	<0.001
Diagnosis	0.969	0.031	1.373	0.245

Table 6: MANOVA Results for Autoantibody Profiles

3.6 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, complementing inferential statistics with a distance-based exploratory approach.

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

Table 7: Euclidean Distances Between Autoantibody Centroids

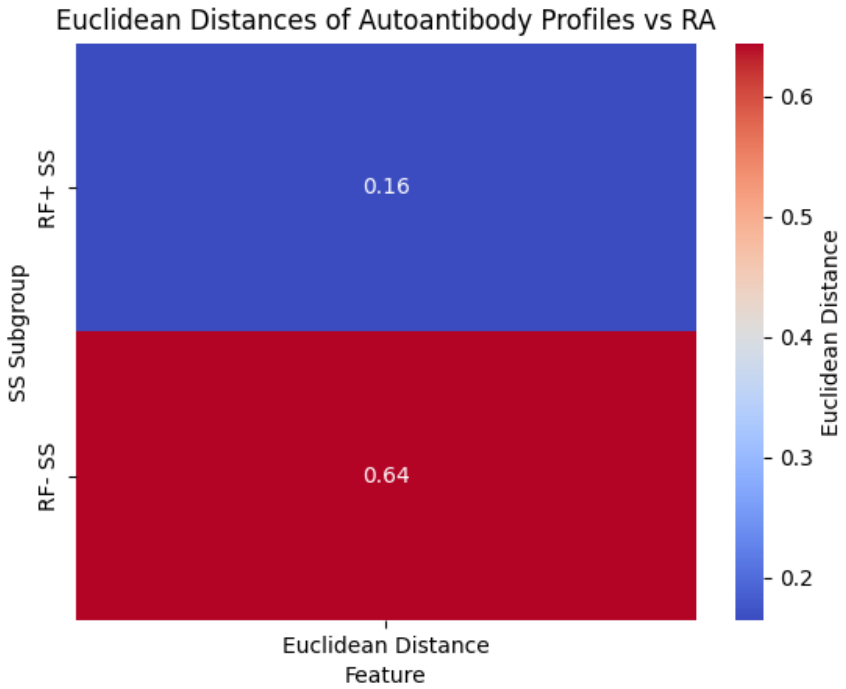

Figure 5: Euclidean distances between rheumatoid arthritis and Sjögren's syndrome subgroups based on autoantibody profiles. Heatmap showing Euclidean distances between RA and RF-positive SS, and RA and RF– SS after Z-normalization of autoantibody levels. Lower values (cooler colors) indicate greater similarity to RA, while higher values (warmer colors) indicate more distinct antibody profiles.

4 Discussion

4.1 Summary of Findings

Across multiple analyses (t-tests, χ^2 tests, ANOVA with Tukey's HSD, MANOVA, and Euclidean distances), RF-positive SS displayed greater 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). Together, these findings are hypothesis-generating and suggest that RF status may stratify SS into subgroups with distinct immunological profiles.

4.2 Interpretation of Findings

T-tests (Table 2) and the subsequent Cohen's d effect sizes (Table 4) suggested that RF-positive SS patients shared stark 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 (Table 3) revealed that the presence or absence of none of the four autoantibodies (ACPA, ANA, anti-dsDNA, anti-Sm) was significantly associated with the diagnosis of RA versus SS (RF-positive and RF-negative combined) in this cohort. All p-values were well above the significance threshold of α=0.05 (ACPA: χ^2=0.005, p=0.942; ANA: χ^2=2.910, p=0.088; anti-dsDNA: χ^2=1.954, p=0.162; anti-Sm: χ^2=0.167, p=0.683). This indicates that while serological profiles provide valuable clues, the mere positivity for any single one of these antibodies is not, on its own, a definitive discriminator between a diagnosis of RA and SS in this synthetic dataset.

ANOVA (Table 5) and post hoc Tukey's HSD revealed that for anti-dsDNA levels, RF-negative SS patients differed significantly from both RF-positive SS and RA (mean difference from RA = 0.2399, $p < 0.05$; mean difference from RF-positive SS = 0.2760, $p < 0.05$), while RF-positive SS and RA showed no significant difference (mean difference = -0.0361, $p > 0.05$). This supports the concept that RF-negative SS represents a distinct immunological profile from RA, whereas RF-positive SS clusters more closely with RA. In multivariate testing, MANOVA (Table 6) was not significant (Wilks' $\lambda = 0.969$, p = 0.245), indicating that the combined antibody profile differences did not reach multivariate significance in this dataset.

Euclidean distance analysis (Table 7) 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.

4.3 Diagnostic Implications

The findings from the synthetic dataset suggest that RF status may define clinical subtypes within Sjögren's syndrome (SS). RF-negative SS appears to reflect the "classic" SS-dominant subtype, characterized by stronger ANA and anti-dsDNA presence, and potentially aligning more closely with systemic lupus erythematosus (SLE). 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. 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. 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. 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.

Additionally, incorporating ACPA testing into the SS diagnostic process may further aid in identifying patients with RA-like pathology. 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. Such refinements may improve diagnostic accuracy, inform prognosis, and guide more personalized management approaches.

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

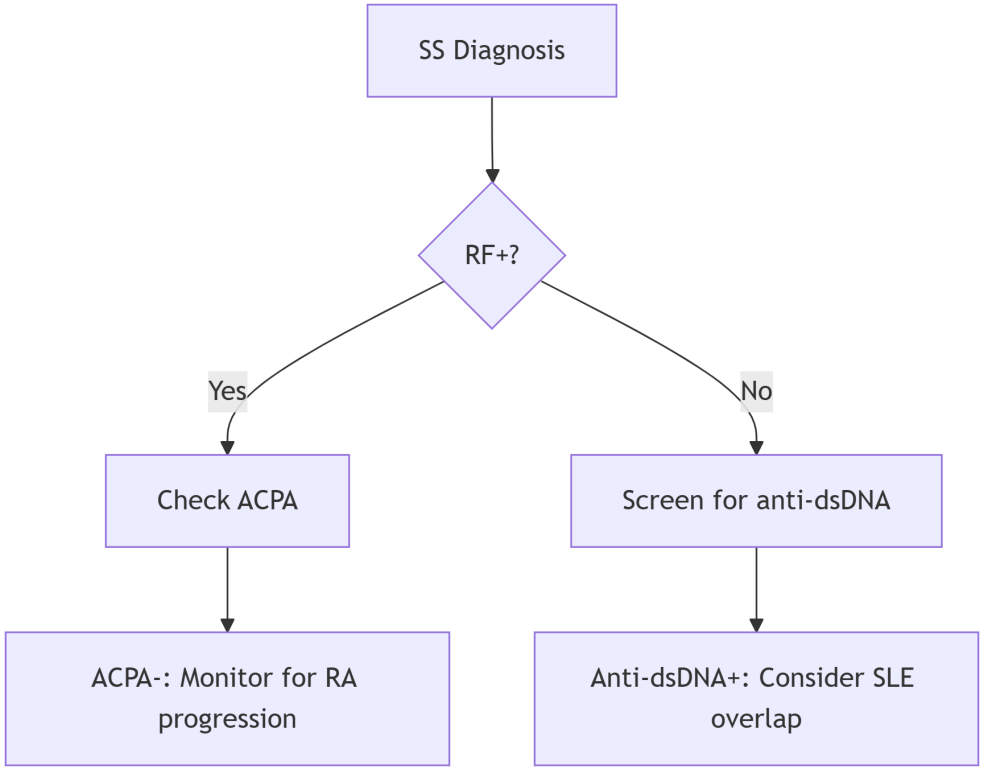

Figure 6: Conceptual decision pathway (hypothesis-generating). Based on findings from the synthetic dataset, RF positivity in SS may indicate a subgroup with greater serological similarity to rheumatoid arthritis (RA), suggesting consideration of RA-targeted evaluation or management. RF-negative SS may reflect a "classic" SS phenotype with possible systemic lupus erythematosus (SLE)-like serology. This pathway is intended for research purposes only and should be validated in real patient cohorts before 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 therefore strictly hypothesis-generating. Antibody and symptom variables were binary, precluding analysis of titers or severity. The cohort size (n = 176) and narrow age range (45–50 years) may limit generalizability. SSA/SSB and RNP were unavailable, constraining SS-specific inference.

A primary methodological consideration is that RF was used to stratify SS subgroups, meaning some observed associations with RA-linked markers could reflect expected correlation structures embedded in the synthetic data (potential circularity). However, the use of multiple independent antibodies and multivariate analyses (MANOVA, Euclidean distance) provides converging evidence that the serological similarity between RF-positive SS and RA is a robust pattern, not solely an artifact of the stratification. 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 one positive control set (RA patients). Including an SLE cohort could have helped pinpoint the RF-negative disease subtype more clearly. Furthermore, healthy controls were not included, as the study aim was specifically to compare SS and RA patients and to test RF-based stratification within SS, not to establish absolute diagnostic thresholds against health. Therefore, the absence of healthy controls is a limitation but does not directly contradict the study's internal comparisons.

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) could identify biomarker signatures distinguishing RF-positive from RF-negative SS. External validation across datasets will be critical to establish robustness and translational utility.

4.6 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

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., and Hawker, G. (2010). 2010 rheumatoid arthritis classification criteria: An american college of rheumatology/european league against rheumatism collaborative initiative. *Arthritis & Rheumatism*, 62(9):2569–2581.

Bournazos, S., Gupta, A., and Ravetch, J. V. (2020). The role of igg fc receptors in antibody-dependent enhancement. *Nature Reviews Immunology*, 20(10):633–643.

Brito-Zerón, P., Baldini, C., Bootsma, H., Bowman, S. J., Jonsson, R., Mariette, X., Sivils, K., Theander, E., Tzioufas, A., and Ramos-Casals, M. (2016). Sjögren syndrome. *Nature Reviews Disease Primers*, 2(1).

Chatzi, A. and Doody, O. (2023). The one-way anova test explained. *Nurse Researcher*, 31(3):8–14.

Damoiseaux, J. and 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.

Delacre, M., Lakens, D., and 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(1):92–101.

Elagib, K. E. E., Børretzen, M., Jonsson, R., Haga, H. J., Thoen, J., Thompson, K. M., and 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(2):388–394.

Feldman, M. and Ginsburg, I. (2021). A novel hypothetical approach to explain the mechanisms of pathogenicity of rheumatic arthritis. *Mediterranean Journal of Rheumatology*, 32(2):112.

Gaddis, M. L. (1998). Statistical methodology: Iv. analysis of variance, analysis of covariance, and multivariate analysis of variance. *Academic Emergency Medicine*, 5(3):258–265.

Ingegnoli, F., Castelli, R., and Gualtierotti, R. (2013). Rheumatoid factors: clinical applications. *Disease Markers*, 35:727–734.

Kim, H., Cho, S.-K., Kim, H. W., Han, J., Kim, Y., Hwang, K.-G., and Sung, Y.-K. (2020). The prevalence of sjögren's syndrome in rheumatoid arthritis patients and their clinical features. *Journal of Korean Medical Science*, 35(45).

Kim, H.-Y. (2015). Statistical notes for clinical researchers: post-hoc multiple comparisons. *Restorative Dentistry & Endodontics*, 40(2):172.

Kurowska, W., Kuca-Warnawin, E. H., Radzikowska, A., and Maśliński, W. (2017). The role of anti-citrullinated protein antibodies (acpa) in the pathogenesis of rheumatoid arthritis. *Central European Journal of Immunology*, 42(4):390–398.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4.

Liu, F., Wang, X.-Q., Zou, J.-W., Li, M., Pan, C.-C., and Si, Y.-Q. (2024). Association between serum antinuclear antibody and rheumatoid arthritis. *Frontiers in Immunology*, 15.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, pages 143–149.

Nashi, R. A. and Shmerling, R. H. (2021). Antinuclear antibody testing for the diagnosis of systemic lupus erythematosus. *Medical Clinics of North America*, 105(2):387–396.

Ragheb, A. (2024). Comprehensive autoimmune disorder dataset.

Van Beers, J. J. and Schreurs, M. W. (2022). Anti-sm antibodies in the classification criteria of systemic lupus erythematosus. *Journal of Translational Autoimmunity*, 5:100155.

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).