

Article

Not peer-reviewed version

Identifying Multi Factor Risk Combinations for IVF Failure in PCOS Patients Using Association Rule Mining

Xuehong Zhu[†], [Lina Ge](#)[†], [Guanghui Dong](#)[†], [Zhong Lin](#), [Feng Han](#)^{*}

Posted Date: 4 June 2025

doi: 10.20944/preprints202506.0339.v1

Keywords: association rule mining; Apriori algorithm; in vitro fertilization; Polycystic Ovary Syndrome; medical data analyst; risk factor combinations



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Identifying Multi Factor Risk Combinations for IVF Failure in PCOS Patients Using Association Rule Mining

Xuehong Zhu ^{1,†}, Lina, Ge ^{2,3,†}, Guanghui Dong ^{4,†}, Zhong Lin ¹ and Feng Han ^{2,3,*}

¹ The Reproductive Hospital of Guangxi, No.3 Longyuan Road, Qingxiu District, Nanning, 530006 Guangxi, China

² School of Artificial Intelligence, Guangxi Minzu University, No.188 Daxue East Road, 530006 Nanning, China

³ Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, No.188 Daxue East Road, Nanning, 530006 Guangxi, China

⁴ School of Public Health, Sun Yat-Sen University, No.74 Zhongshan Second Road, Guangzhou, 510080 Guangdong, China

* Correspondence: fenghan@gxmzu.edu.cn

† Equal contribution.

Abstract: Background: Polycystic ovary syndrome (PCOS) is a major indication for in-vitro fertilisation (IVF). Previous studies typically evaluate candidate risk factors in isolation, obscuring the multi-factor interactions that often govern clinical pregnancy outcomes. **Methods:** This study retrospectively analyzed EMR data from PCOS patients undergoing IVF. Key clinical variables (age, body mass index, years of infertility, hormonal/metabolic disorders, tubal or uterine abnormalities, ovarian conditions including luteinized unruptured follicle syndrome [LUFs], and IVF treatment details) were one-hot-encoded. Apriori association rule mining was then used to identify patterns associated with clinical pregnancy failure, with thresholds of support ≥ 0.05 , confidence ≥ 0.60 , and lift > 1 to ensure robust rules. This novel approach enabled the detection of multifactorial risk associations that were not apparent in a traditional single-factor analysis. **Results:** The overall clinical pregnancy success rate in the cohort was ~40%. Association rule mining uncovered several clinically meaningful patterns; notably, maternal age > 35 years was a recurrent component of high-risk combinations, often alongside other factors (e.g., metabolic or anatomical abnormalities). For example, a combination of LUFs and tubal obstruction was strongly associated with failure, indicating a synergistic negative effect. Many of these multifactorial associations would have been missed by analyzing variables individually. **Conclusions:** Apriori rule mining effectively identified complex risk factor combinations for IVF failure in PCOS, informing individualized treatment strategies. Clinically, recognizing patients with advanced age coupled with specific reproductive or metabolic abnormalities can guide tailored interventions to improve IVF success. More broadly, this work demonstrates the potential of integrating association rule mining with EMR data for clinical decision support, enabling the discovery of hidden patterns to enhance personalized medicine.

Keywords: association rule mining; Apriori algorithm; in vitro fertilization; Polycystic Ovary Syndrome; medical data analyst; risk factor combinations

1. Introduction

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine and metabolic disorder among women of reproductive age. It is characterized by hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology [1] and is a leading cause of infertility, affecting 5% to 20% of women in this age group globally [2]. Despite the effectiveness of ovulation induction and lifestyle

modifications for many patients, a significant proportion of women with PCOS still require assisted reproductive technology (ART), particularly in vitro fertilization (IVF), to achieve pregnancy. However, PCOS patients undergoing IVF face several challenges, including an increased risk of ovarian hyperstimulation syndrome (OHSS), reduced embryo quality, and lower clinical pregnancy rates [3].

In recent years, advancements in medical technology and a deeper understanding of the pathophysiology of PCOS have led to improvements in IVF success rates. Nevertheless, a considerable number of PCOS patients still fail to achieve pregnancy after IVF treatment. Studies have shown that various factors may influence the success of IVF in PCOS patients, including age, body mass index (BMI), duration of infertility, hormonal imbalances, tubal factors, uterine anomalies, and more [4]. Moreover, it is increasingly recognized that the interplay between these factors can significantly impact IVF outcomes. For example, insulin resistance and hyperandrogenemia have been shown to synergistically impair oocyte competence and endometrial receptivity in PCOS patients undergoing IVF [5].

In clinical practice, accurately predicting the success of IVF treatment for PCOS patients is crucial for developing personalized treatment plans. However, due to the complex and multifactorial nature of the influences on IVF outcomes, traditional statistical methods often fall short of fully capturing these intricate relationships. Therefore, there is a need for more advanced data analysis techniques to uncover hidden patterns and associations within clinical data, thereby providing more robust support for clinical decision-making. Machine learning approaches such as random forest and gradient boosting have recently demonstrated promising performance in predicting IVF outcomes, particularly in heterogeneous populations such as those with PCOS [6].

Association rule mining (ARM) is a data mining technique that identifies hidden relationships in data by discovering “if-then” patterns. It has been successfully applied in the medical field to reveal dependencies between clinical factors and inform disease diagnosis and treatment [7,8]. In this study, we aimed to employ ARM to uncover the interactions between different factors associated with IVF failure in PCOS patients and determine their collective impact on IVF outcomes. By doing so, we hope to provide clinicians with more precise decision-making support to enhance the success rates of IVF treatments for PCOS patients.

2. Methods and Materials

2.1. Data Source

This study retrospectively analyzed de-identified electronic medical records (EMRs) from the Reproductive Hospital of Guangxi, covering all **in vitro fertilization** (IVF) treatment cycles between 2018 and 2023 for females diagnosed with **Polycystic Ovary Syndrome** (PCOS). Inclusion was restricted to PCOS patients, diagnosed according to standard clinical criteria, undergoing IVF with or without intracytoplasmic sperm injection (ICSI) during the study period. The dataset captured each patient’s baseline characteristics, relevant diagnoses, treatment details, and IVF outcomes. *Clinical pregnancy* (the outcome of interest) was defined as a positive fetal heartbeat on ultrasound, approximately 6–7 weeks post-embryo transfer, and this binary outcome—pregnant or not pregnant—was recorded for each IVF cycle. Only records with complete information on key variables and outcomes were retained for analysis. The study was approved by the hospital’s institutional ethics board, and due to its retrospective nature and the use of de-identified data, informed consent was waived.

2.2. Data Preprocessing

The data from raw EMR extracts were rigorously preprocessed to ensure high-quality data for further analysis. The following steps were implemented:

- (1) **Terminology standardization:** Clinical descriptions and diagnoses were normalized to a consistent vocabulary. In the raw EMRs, the same concept could be documented with varying

terms or abbreviations. For example, “*Polycystic Ovary Syndrome*” was recorded as “*Polycystic ovarian syndrome*”, “*PCOS*”, “*Poly-ovary syndrome*”, etc., depending on the doctor’s style. To address this, all symptom and diagnosis labels were mapped to a unified terminology, effectively harmonizing synonyms and abbreviations into one standard descriptor. This normalization ensured that each clinical concept, such as *hyperprolactinaemia*, was represented uniformly across all records, preventing duplicate features arising from naming variations.

- (2) **Record filtering and de-duplication:** Duplicate entries and records with substantial missing or incomplete data were removed. It is common for secondary-use healthcare data to contain duplicate records or omissions, which can bias the results. Any repeated patient entries, as well as cases lacking critical fields, such as outcome or key diagnoses, were identified and removed. This step improved data integrity and reliability by focusing the analysis on unique, complete cases.
- (3) **One-hot encoding:** One-hot encoding is a standard data transformation technique for converting categorical data into a numeric matrix for machine learning [9]. In practice, each distinct feature value, such as *primary infertility*, *secondary infertility*, or different stimulation protocols, was turned into its own column with a binary value, where the value “1” indicates the presence of that attribute in a given record and “0” indicates its absence. This encoding resulted in a Boolean feature matrix where each row corresponds to a patient’s IVF cycle, and each column corresponds to the presence of a specific condition or attribute. Continuous variables such as age, body mass index (BMI), and duration of infertility were discretized into clinically meaningful ranges before encoding, according to the standard of the World Health Organization [10,11].

All data cleaning and preprocessing steps were carried out in compliance with best practices for secondary analysis of clinical data. Overall, this process yielded a curated dataset of PCOS IVF cases that was consistent, free of major quality issues, and ready for pattern mining analyses.

2.3. Feature Selection

In this study, a comprehensive dataset of clinical and treatment features believed to be relevant to IVF outcomes in PCOS patients based on previous studies was extracted from the cleaned EMRs. These variables included demographic factors, comorbid conditions, anatomical factors, and treatment parameters, as detailed in Table 1.

All features were encoded as 0/1 variables after preprocessing. Continuous measurements (age, BMI, infertility years, oocyte count) were categorized into discrete bins to allow their inclusion as categorical items in the association rule analysis. The selection of these features was guided by clinical domain knowledge and the literature on PCOS and infertility, ensuring that the dataset captured most known factors that might interact to influence IVF success in PCOS patients.

Table 1. Features extracted from EMRs.

Domain	Extracted Variables	Reference
Demographics	Female age: discretized $\leq 35, >35$	[10,11]
Anthropometry	Female BMI (kg m^{-2}): discretized $\leq 2, >24$	
Reproductive history	Years of infertility: discretized $\leq 5, >5$; infertility type: primary/secondary	[12,13]
Hormonal and metabolic disorders	Hyperprolactinemia, sub-clinical hypothyroidism, insulin resistance	[14–17]
Tubo-uterine factors	Unilateral tubal obstruction, bilateral tubal obstruction, hydrosalpinx, pelvic adhesion, intra-uterine adhesion	[18,19]s
Ovarian and endocrine factors	Luteinized unruptured follicle syndrome (LUFS), ovarian cysts	[18,19]
Uterine malformations	Septate uterus, uterine fibroids (leiomyomas), adenomyosis	[20]
ART treatment data	Stimulation protocol (categorical), number of oocytes retrieved (binned), fertilization method (IVF vs. ICSI)	[21]
Outcome	Clinical pregnancy: 1 = success, 0 = failure	

2.4. Association Rule Mining

Association rule mining is a data mining technique that identifies hidden relationships in data by finding *if-then* patterns [22], and it has been shown to be useful in the medical domain for discovering dependencies between clinical factors [23,24]. In this analysis, the classical *Apriori* algorithm [25,26] was used to identify frequently co-occurring feature sets in the dataset and then derive association rules that predict the likelihood of clinical pregnancy, with a focus on rules where the consequent—the “then” part of the rule—was *clinical pregnancy failure* as the IVF outcome in PCOS patients.

The Apriori algorithm comprises two main steps: frequent itemset generation and association rule extraction.

- (1) **Frequent itemset generation:** The one-hot-encoded dataset was used to compute the frequent itemset. Apriori systematically explores combinations of items (features) and counts their occurrences in the one-hot-encoded dataset generated in the data preprocessing stage. A minimum *support* threshold of 0.05 was set, meaning that an itemset had to appear in at least 5% of patient records to be considered frequent, balancing the need to find non-trivial patterns with the desire to ignore extremely rare combinations and ensuring that any reported association involves a patient subset of meaningful size in the cohort.
- (2) **Association rule extraction:** Upon the identification of frequent itemsets, the association rules function commenced the calculation of rule metrics. A minimum confidence threshold of 0.60 was applied to filter the rules. Confidence measures the conditional probability of the consequent given the antecedent. It was utilized to assess the likelihood of pregnancy under specific conditions. A rule’s confidence was required to meet or exceed 60% for consideration in this study, ensuring that, among patients exhibiting the antecedent feature set, at least 60% achieved the expected result. Furthermore, a criterion of lift > 1.0 was imposed for all rules. Lift, defined as the ratio of the rule’s confidence to the baseline probability of the consequent, indicates the degree to which the antecedent and outcome co-occur more frequently than expected by chance. The lift > 1 requirement guaranteed that any reported rule represented a positive association, signifying an improvement over random chance in predicting pregnancy.

After generating the initial set of rules, we specifically filtered for those rules **ending in** \Rightarrow *clinical pregnancy failed*. This yielded rules of the form **antecedents to clinical pregnancy outcome**, highlighting combinations of patient features that were associated with failed conception in this PCOS IVF cohort. Rule results were then examined for clinical plausibility and ranked by their metrics. In the end, only filtered rules meeting all the threshold criteria and reaching statistical significance according to Fisher’s test were retained for interpretation.

2.5. Computational Environment

All analyses were conducted using Python version 3.12 in a Jupyter Notebook environment. Data handling and preprocessing were performed with the **pandas** library and **scikit-learn**. Association rule mining was carried out with the **mlxtend** library. The tools and models employed in this study are shown in Table 2.

Throughout the analysis, we adhered to an academically rigorous approach: cleaning the clinical dataset to a high standard, encoding features in a manner suitable for the discovery of patterns, and using validated data mining techniques with justified parameter choices. The flow of methods and materials is shown in Figure 1.

Table 2. Computing Environment for This Study.

Category	Specification	Purpose
Programming language	Python 3.12	Core scripting/analysis
Interactive IDE	Jupyter Notebook	Reproducible, stepwise workflow
Key libraries	pandas, scikit-learn, mlxtend	Data wrangling; preprocessing; Apriori and rule generation
Plotting	matplotlib, seaborn	Descriptive and network visualization
Hardware	2 × Intel Xeon (48 physical cores), 128 GB RAM	Parallel support-counting and rule filtering
Operating system	Ubuntu 22.04 LTS	Stable Linux environment for multi-threaded tasks

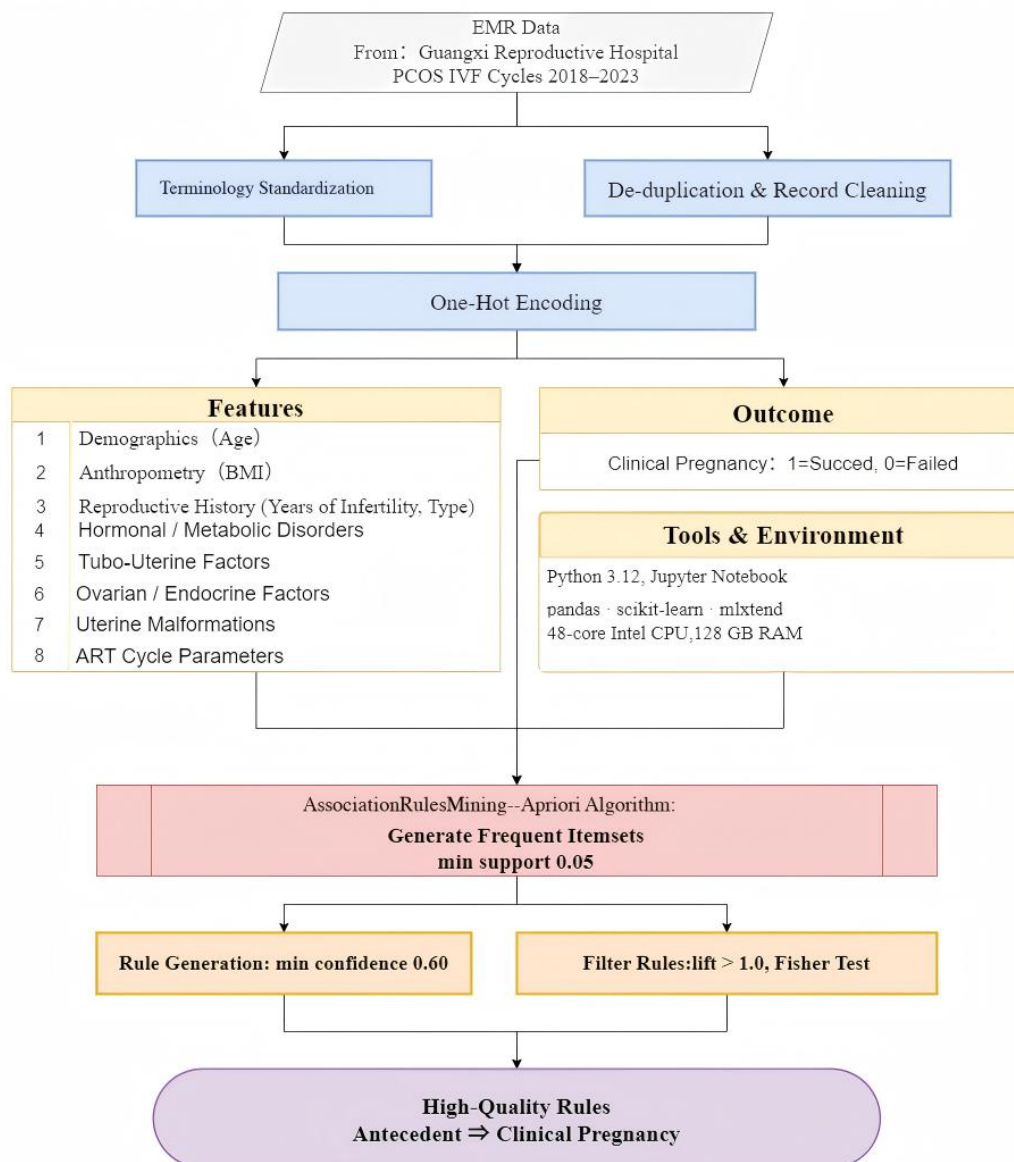


Figure 1. The flow of research methods and materials.

3. Results

3.1. Descriptive Statistics

3.1.1. Clinical Pregnancy Outcomes

The overall clinical pregnancy rate among patients diagnosed with Polycystic Ovary Syndrome (PCOS) undergoing their first IVF treatment was approximately **18.1%**, indicating that around one-fifth of IVF cycles successfully resulted in clinical pregnancy. This finding aligns with previously published clinical pregnancy rates observed among PCOS populations undergoing IVF [27].

3.1.2. Demographic and Clinical Characteristics

The demographic characteristics indicated a relatively young patient population, consistent with typical PCOS demographics. The mean age was **31.39 ± 4.45 years** (range: 20–46 years). The body mass index (BMI) averaged **23.53 ± 3.43 kg/m²**, with a range between 14.42 and 35.58 kg/m², suggesting that most patients were in the normal-weight to overweight category. The average duration of infertility was **4.45 ± 3.17 years**, ranging widely from 0.1 to 22 years, reflecting diverse patient fertility histories (Figure 2).

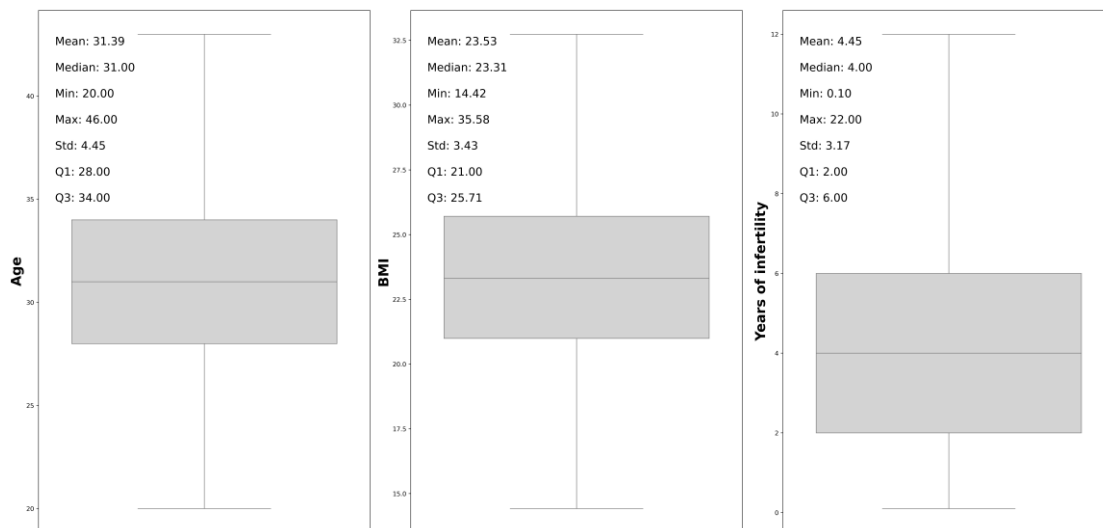


Figure 2. Boxplots of age, BMI, and infertility duration distributions.

Clinical conditions were analyzed according to their prevalence among successful and unsuccessful IVF cycles, presented in Figure 3. Conditions such as bilateral tubal obstruction and habitual abortion appeared predominantly in unsuccessful pregnancy cases, implying negative implications for IVF success. Pelvic adhesions and undiagnosed adnexal masses were relatively frequent diagnoses in both outcomes, indicating a high prevalence but limited specificity concerning pregnancy outcomes. In contrast, conditions such as hypertension, insulin resistance, and hyperprolactinemia were less prevalent overall, thus limiting their interpretative significance in isolation.

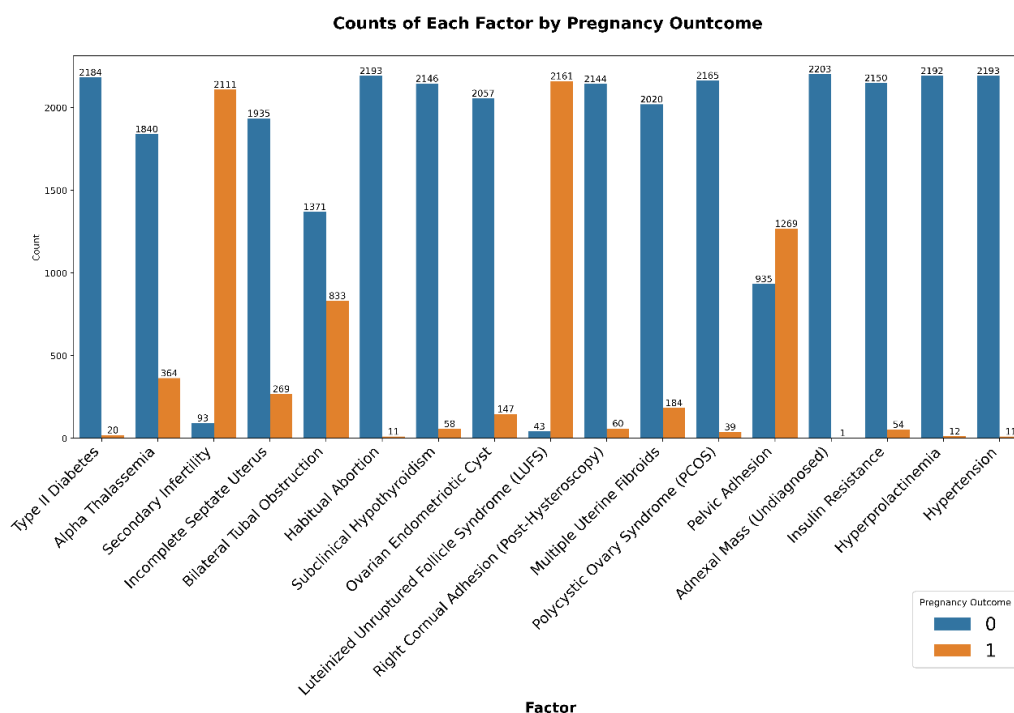


Figure 3. The prevalence of clinical features by pregnancy outcome.

3.1.3. Comparison of Pregnancy Outcomes by Demographic and Clinical Groups

The comparative prevalence of each clinical feature between pregnancy success and failure groups is illustrated in Figure 4.

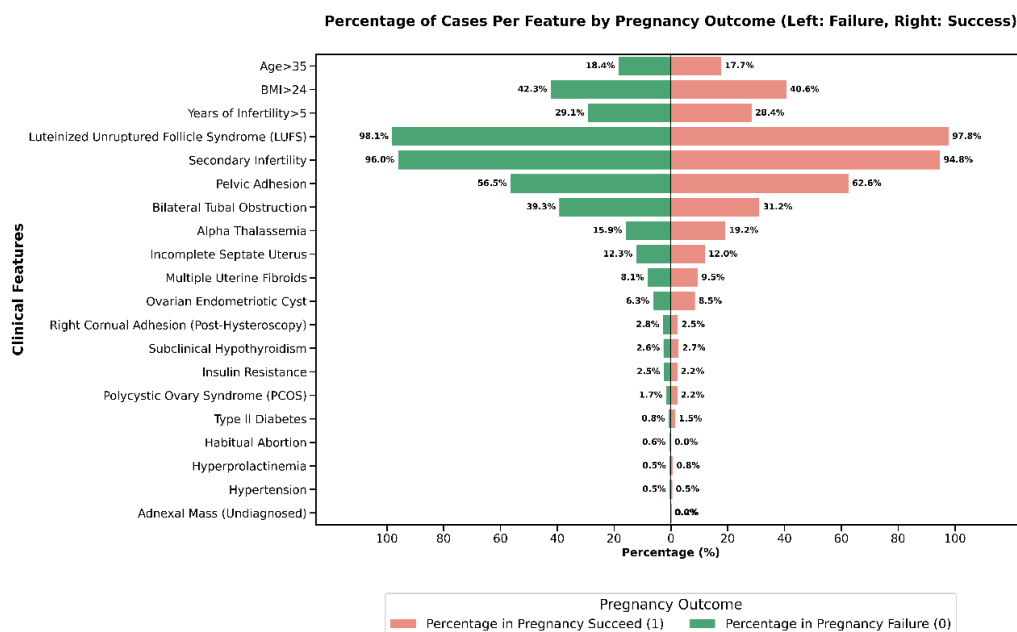


Figure 4. Comparative prevalence of clinical features in IVF cycles resulting in pregnancy success versus failure.

Conditions such as luteinized unruptured follicle syndrome (LUFS), secondary infertility, and pelvic adhesion were highly prevalent in both successful and unsuccessful outcomes, indicating their common occurrence among PCOS-IVF patients. Features that were notably more frequent in the pregnancy failure group included bilateral tubal obstruction (39.3% in failure vs. 31.2% in success) and age greater than 35 years (18.4% in failure vs. 17.7% in success). Conversely, some conditions, such as adnexal mass (undiagnosed), hyperprolactinemia, and hypertension, demonstrated extremely low overall prevalence, limiting their discriminatory power. These results demonstrate that certain clinical factors, particularly tubal obstruction and advanced age, can potentially serve as meaningful indicators for clinical pregnancy outcomes, emphasizing their importance in clinical evaluation and patient counseling prior to IVF treatment.

3.1.4. Correlation of Features with Pregnancy Outcome

Pearson correlation analysis was conducted to quantify relationships between individual clinical features and IVF pregnancy outcomes, as shown in Figure 5:

- (1) Bilateral tubal obstruction exhibited the strongest negative correlation (-0.064), aligning with the clinical understanding of impaired fertility associated with significant tubal pathology.
- (2) Pelvic adhesion (0.048) and adnexal mass (undiagnosed) (0.045) demonstrated small positive correlation with pregnancy success, an unexpected finding possibly influenced by confounding clinical management practices.
- (3) Habitual abortion (-0.033) and secondary infertility (-0.024) negatively correlated with pregnancy success, confirming their relevance as adverse prognostic factors.
- (4) Other factors, including BMI, insulin resistance, hypertension, and age, exhibited negligible or minimal correlations, suggesting limited predictive value when assessed individually in this cohort.

In sum, the correlation analysis indicated that no single factor exerted a significant effect on the clinical outcome of an IVF cycle in patients with PCOS. Further data mining techniques need to be implemented to discover the interactive and combined effects of different factors.

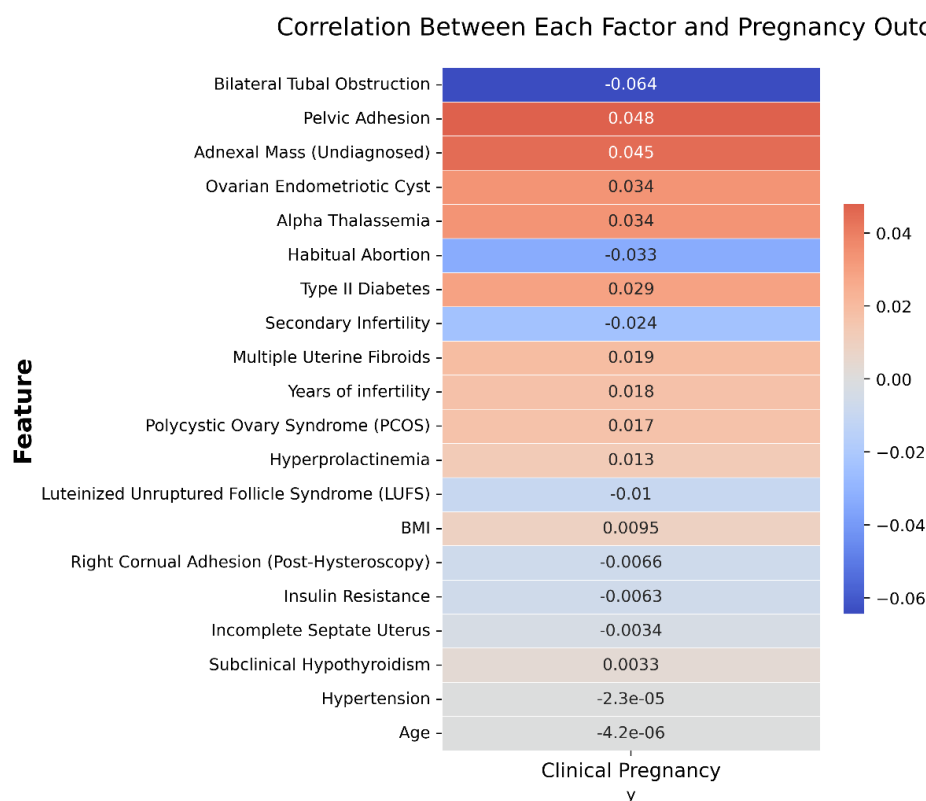


Figure 5. Heatmap of correlations between clinical features and clinical pregnancy outcomes.

3.2. Association Rule Mining Outcomes

Association rule mining (ARM) was employed to identify clinical diagnoses and patient characteristics significantly associated with an increased risk of IVF clinical pregnancy failure among patients with PCOS.

3.2.1. Overall Rule Summary

ARM analysis was specifically directed toward uncovering associations where the consequent was fixed as clinical pregnancy failure, thereby enabling the identification of clinical factors and conditions that elevate the risk of unsuccessful IVF outcomes. After applying stringent thresholds—minimum support ≥ 0.05 , confidence ≥ 0.60 , and lift > 1.0 —and statistical validation via the chi-square (χ^2) test, a total of 26 significant rules were generated, as shown in Figure 6. These rules frequently involved antecedents comprising ovarian dysfunction factors (e.g., luteinized unruptured follicle syndrome (LUFS)), structural uterine anomalies, tubal factors (e.g., bilateral tubal obstruction), and advanced maternal age.

The frequency and strength of associations of individual clinical features (single itemsets) with IVF clinical pregnancy failure in PCOS patients are summarized in Table 3:

- (1) The most frequent clinical features associated with IVF failure were luteinized unruptured follicle syndrome (LUFS) and secondary infertility, each identified in 13 instances.
- (2) Luteinized unruptured follicle syndrome (LUFS) had a high support rate (79.67%), indicating its high prevalence, although it demonstrated a relatively weak lift (1.0017), suggesting limited discriminative power when considered alone.
- (3) Bilateral tubal obstruction exhibited the strongest association with clinical pregnancy failure, showing the highest lift value (1.0422) and a confidence of 85.25%, emphasizing its significance as a risk factor.

- (4) Other important features, such as years of infertility >5 and BMI >24, demonstrated moderate frequencies and lifts (1.0104 and 1.0196, respectively), indicating their meaningful, though more limited, contributions as individual predictors.
- (5) Pelvic adhesion lacked calculated metrics in this analysis, limiting the interpretation of its independent predictive strength.

Table 3. Frequency and association metrics of individual clinical features.

Itemset	Frequency	Support	Confidence	Lift
Luteinized Unruptured Follicle Syndrome (LUFS)	13	0.7967	0.8194	1.0017
Secondary Infertility	13	0.2505	0.8251	1.0086
Pelvic Adhesion	7	-	-	-
Years of Infertility > 5	7	0.1751	0.8266	1.0104
Bilateral Tubal Obstruction	6	0.3122	0.8525	1.0422
BMI > 24	4	0.1711	0.8341	1.0196
Age > 35	4	0.1502	0.8234	1.0065

In summary, the statistical results revealed that, when considered in isolation, most features exhibited limited predictive value for clinical pregnancy outcomes in PCOS patients undergoing IVF. This suggests that single factors alone may not sufficiently explain treatment success or failure, highlighting the importance of analyzing multidimensional risk combinations.

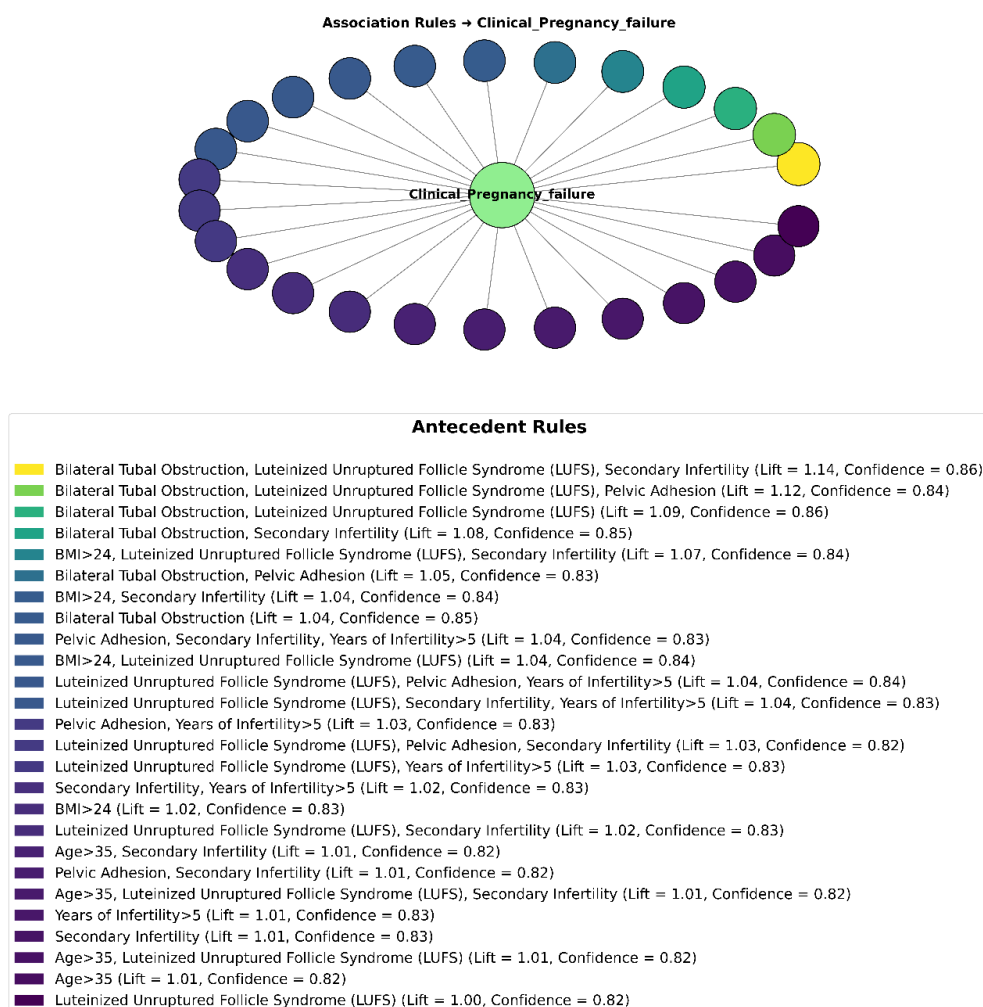


Figure 6. Effective rules leading to IVF clinical pregnancy failure in PCOS patients.

3.2.2. Top-Ranked Association Rules

The top 10 association rules, ranked by lift and confidence, are summarized in Figure 7. All rules share a common consequent—clinical pregnancy failure—and were identified using the Apriori algorithm with support ≥ 0.05 , confidence ≥ 0.60 , and lift > 1 .

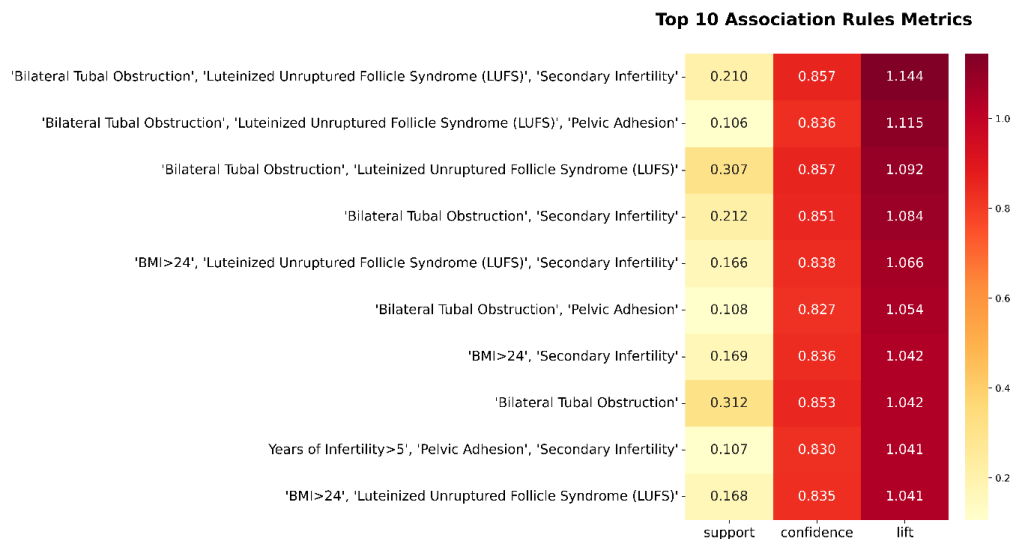


Figure 7. Top 10 association rules predicting clinical pregnancy failure in PCOS patients.

These top-ranked rules reveal that **combinatorial effects** of anatomical, ovarian, and demographic risk factors substantially heighten the likelihood of IVF failure:

- (1) Bilateral tubal obstruction appears in 7 of the top 10 rules, underscoring its centrality as a structural impediment to successful implantation or embryo transport.
- (2) Luteinized unruptured follicle syndrome (LUFS) frequently co-occurs with both tubal factors and metabolic risks (e.g., BMI > 24), suggesting that ovarian dysfunction and metabolic dysregulation synergistically compromise reproductive outcomes.
- (3) Secondary infertility and prolonged infertility (>5 years) repeatedly emerge in multifactorial rule sets, reflecting the compounded difficulty of achieving pregnancy in patients with a history of prior conception failure.

These insights highlight the clinical relevance of **multi-feature pattern recognition**, which offers a more nuanced risk stratification than single risk factors evaluated in isolation. They also provide a foundation for developing AI-assisted decision support tools to predict IVF failure and inform personalized treatment strategies.

4. Discussion

4.1. Comparative Analysis

4.1.1. Comparison with Previous Literature or Known Clinical Evidence

The findings of this research provide a **multidimensional perspective** on IVF outcomes of PCOS patients that complements prior knowledge from more reductionist analyses. Several of our key associations—such as the detrimental effects of hydrosalpinx (bilateral tubal obstruction) and prolonged infertility—are strongly supported by the literature. For instance, a study by Ou et al. [28] established that the presence of a hydrosalpinx can **diminish implantation rates and increase early pregnancy loss** in IVF due to embryotoxic fluid and poor endometrial receptivity. This aligns with the rule in this study that any PCOS patient with an uncorrected bilateral tubal blockage (likely hydrosalpinx) almost invariably fails to conceive via IVF unless the tube issue is addressed. The logical clinical action is to perform a salpingectomy or proximal tubal occlusion prior to IVF in such

patients, a recommendation echoed in many studies that showed improved IVF success after hydrosalpinx treatment [29,30].

The strong influence of infertility duration on IVF outcome in this study reinforces a consistent theme: *the sooner, the better*. An early meta-analysis [31] found a negative association between the duration of infertility and IVF success, and more recent analyses concur that beyond roughly 3–5 years of trying, each additional year is associated with lower pregnancy odds [32,33]. This study's PCOS-specific data suggest that even within a relatively young cohort, those who had been infertile for ≥ 5 years had significantly reduced success. This could be partly because a longer duration often correlates with older age and other factors; however, even after accounting for age in some of our combined rules, duration remained a factor. One interpretation is that long-term infertility may indicate underlying intractable issues (e.g., poor egg/embryo quality, endometrial dysfunction) that persist despite IVF. It might also reflect that these patients have undergone multiple prior treatments or IVF cycles without success, hinting at recurrent implantation failure scenarios. Clinically, this stresses that practitioners might consider escalating treatment or exploring adjunct therapies (immunological work-ups, use of donor gametes, etc.) when faced with a patient who has had many years of unexplained infertility.

This research reaffirms the critical impact of female age—which remains the single strongest determinant of IVF success across all populations. Advanced age (≥ 35 , especially ≥ 40) dramatically elevated failure risk in our PCOS cohort, consonant with general IVF outcomes. Likewise, obesity and metabolic factors are well-documented to impair fertility treatment outcomes [34]. A 2024 systematic review noted that, in women with PCOS, high BMI independently lowers clinical pregnancy and live birth rates and raises miscarriage risk. In this article, obesity featured in some rules, though not the top rule, implying that while obesity is indeed harmful, other factors, such as tubal status or duration, were even more dominant in our dataset. It is possible that because a majority of our PCOS patients were overweight, BMI did not differentiate outcomes as sharply—a type of range restriction effect. However, we did observe that *lean PCOS* patients had slightly better success rates than obese PCOS patients, aligning with the consensus that weight management can improve IVF outcomes in PCOS.

4.1.2. Unexpected Findings in the Current Analysis

The data-driven discovery of the LUFs + tubal factor combination as a high-failure profile appears to be a novel insight with limited direct precedent in the literature. LUFs is a subtle form of ovulatory dysfunction, and while it is known to cause infertility [35], it is not commonly discussed in the context of IVF outcomes because ovarian stimulation with an HCG trigger is expected to circumvent follicle rupture problems. However, the results suggest that some PCOS patients may still experience issues analogous to LUFs even in IVF (e.g., follicles that luteinize without yielding an egg). A recent study by Li et al. [36] noted that LUF cycles negatively affected pregnancy outcomes in natural-cycle FET, highlighting that luteinization without ovulation can disrupt timing and endometrial preparation. In stimulated IVF cycles, an argument can be made that if a patient has a tendency toward LUFs, careful monitoring and trigger timing are crucial—or alternatives such as a *dual trigger* (HCG + GnRH agonist) might be beneficial to ensure oocyte release. The combination with the tubal factor is likely a proxy for the overall *severity of infertility*: these patients effectively have two strikes against them, and indeed, our analysis shows that they fare poorly. While not previously reported as a combined risk in the literature, this finding is intuitive and underscores the importance of addressing *all* known factors in a multi-disciplinary management approach to give the best chance of success.

Another novel discovery is that our approach identified interactions that traditional multivariable models might miss or not emphasize. For instance, using logistic regression, Liu et al. [37] found that PCOS per se was not an independent predictor of live birth after adjusting for confounders, meaning that if age, BMI, etc., were controlled, PCOS patients performed as well as others. However, that same analysis showed that within the PCOS group, factors such as younger

age, shorter infertility, and good embryo quality were associated with higher live birth rates. The results in this study complement this by explicitly highlighting *combinations* (e.g., older age + fewer good embryos) that lead to failure. Essentially, ARM provides a human-readable set of rules that align with what an experienced clinician might surmise through years of practice. The benefit is that ARM can systematically scan through dozens of features to flag combinations that merit attention, possibly revealing less obvious patterns.

Besides, no individual predictor exhibited a robust, statistically significant effect. However, a small subset of variables—such as bilateral tubal obstruction—displayed modest associations, with lift=1.042.

4.2. The Strengths and Innovation of This Study

A strength of this study is the demonstration of how data mining techniques such as Apriori can be applied in reproductive medicine. This approach can handle many variables and uncover associations without the need to pre-specify an outcome model. The rules generated are intuitively understandable (“IF X and Y, THEN Z”), which could aid clinical decision-making more directly than a complex predictive model. To our knowledge, this is the first study to report association rules in the context of IVF outcomes for PCOS. It thus opens the door for using similar methods on larger IVF databases to perhaps discover phenotype-specific patterns (for example, does a combination of certain hormone levels predict ovarian hyperstimulation syndrome risk in PCOS?). Additionally, by focusing on **clinical pregnancy failure**, this research highlighted an outcome (failure to conceive) that is often less reported than success rates, yet is critically important when counseling patients about their prognosis and when planning interventions.

4.3. Limitations of This Study

Despite its insights, this research has several limitations. First, the study is retrospective and observational; **association** does not imply **causation**. The rules we found do not prove that, say, LUFs *causes* IVF failure—only that they occur together frequently. There could be underlying confounders (for example, perhaps women with LUFs also had poor ovarian reserve, which was the real driver of failure). We attempted to mitigate spurious findings by requiring relatively high support and by conducting statistical tests, but some associations might still be coincidental or due to bias in the dataset. Second, the dataset size (N ~300) is moderate; a larger sample would allow the detection of associations with lower support (rarer but potentially important scenarios). Our choice of a 10% support threshold means we likely missed rules involving very rare conditions (e.g., uncommon genetic factors or severe male factor cases)—these might still be clinically significant for individual patients but were not detectable in our analysis. Third, our feature encoding, while comprehensive, was limited to what was recorded. We did not include some potentially relevant variables such as **AMH levels, insulin resistance indices, or detailed embryo morphology scores**. The inclusion of such data might yield additional rules (for instance, a combination of low AMH + PCOS could predict poor ovarian response). Fourth, the analysis was confined to PCOS patients at a single center; thus, the rules reflect that specific population and practice (e.g., the stimulation protocols used, the prevalence of certain issues in that clinic). Caution is needed in generalizing the results to all PCOS patients or other IVF centers. What holds true in our data (for example, proportion of patients with hydrosalpinx) may differ elsewhere. Validation on external datasets would strengthen the confidence in these rules.

Part of the results showed inconsistent with the conclusions that have been confirmed by previous studies, for instance, the influence of BMI on the clinical outcomes of IVF was not reflected, while it is an axiom widely recognized by the most researchers[38–40]. The reason might be that over 80 % of patients were below BMI threshold, 24 kg m⁻² for Asian women standard by WHO, diluting its discriminatory power. In future, multi-centre datasets with broader BMI ranges may clarify obesity-specific gradients.

Finally, as with any data mining, there is a risk of **overfitting** or finding patterns that lack biological plausibility. We have tried to interpret only those rules that made clinical sense and matched some external evidence. It is reassuring that our top findings were aligned with known mechanisms (e.g., tubal fluid harming implantation, long infertility reflecting tougher cases). However, we remain careful not to overinterpret combinations that could be artifacts.

5. Conclusions

In conclusion, association rule mining was applied to identify **key combinations of factors associated with IVF failure in PCOS patients**. The results highlight that it is often the convergence of multiple adverse factors—such as **ovulatory dysfunction, tubal pathology, and long-standing infertility**—that dramatically lowers the chances of pregnancy in this high-risk group. These findings are largely in agreement with the existing literature on individual risk factors while also providing a *novel integrated view* of how these factors interact. For clinicians, the insights underscore the importance of **comprehensive infertility work-ups** in PCOS: a patient with both PCOS and another infertility factor (e.g., tubal obstruction) should be counseled about the lower success probability and the need to possibly correct the remediable factor before IVF. Similarly, aggressive management (or earlier transition to IVF) may be warranted for those with many years of infertility rather than prolonged attempts with lesser treatments.

This work demonstrates the utility of data-driven approaches in reproductive medicine. By uncovering patterns that might be overlooked by traditional analyses, association rule mining can generate hypotheses for further research (e.g., investigating the mechanistic link between LUFs and IVF outcomes in PCOS) and potentially inform clinical decision support systems. Future studies should validate these rules in larger, multi-center cohorts and assess their predictive value prospectively. It would also be worthwhile to extend this analysis to **IVF success** rules (profiles of patients who succeed) and to compare PCOS with non-PCOS infertile populations to see whether different rules apply. Ultimately, translating these findings into practice—for example, developing a risk score or checklist based on the presence of multiple factors—could help personalize IVF counseling and treatment for PCOS patients. In summary, women with PCOS are a heterogeneous group, and our study highlights that *the sum of their reproductive challenges* determines IVF outcomes. Recognizing and addressing each element of that sum holds promise for improving fertility success in this prevalent and challenging condition.

Author Contributions: Xuehong Zhu: Data curation; formal analysis ; investigation ; writing – original draft. Guanghui Dong: project administration (equal); writing – original draft . Zhong Lin: Supervision; Fund Provider. Lina Ge: Investigation, formal analysis. Feng Han: Conception; Framework Construction; writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number [82460309] and the APC was funded by Reproductive Hospital of Guangxi.

Data Availability Statement: All raw data and code are available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Azziz, R.; Carmina, E.; Chen, Z.; Dunaif, A.; Laven, J.S.; Legro, R.S.; Lizneva, D.; Natterson-Horowitz, B.; Teede, H.J.; Yildiz, B.O. Polycystic Ovary Syndrome. *Nat. Rev. Dis. Primer* **2016**, *2*, 1–18.
2. Bozdag, G.; Mumusoglu, S.; Zengin, D.; Karabulut, E.; Yildiz, B.O. The Prevalence and Phenotypic Features of Polycystic Ovary Syndrome: A Systematic Review and Meta-Analysis. *Hum. Reprod.* **2016**, *31*, 2841–2855.
3. Sunkara, S.K.; Rittenberg, V.; Raine-Fenning, N.; Bhattacharya, S.; Zamora, J.; Coomarasamy, A. Association between the Number of Eggs and Live Birth in IVF Treatment: An Analysis of 400 135 Treatment Cycles. *Hum. Reprod.* **2011**, *26*, 1768–1774.

4. McGee, E.A.; Hsueh, A.J. Initial and Cyclic Recruitment of Ovarian Follicles. *Endocr. Rev.* **2000**, *21*, 200–214.
5. Dumesic, D.A.; Oberfield, S.E.; Stener-Victorin, E.; Marshall, J.C.; Laven, J.S.; Legro, R.S. Scientific Statement on the Diagnostic Criteria, Epidemiology, Pathophysiology, and Molecular Genetics of Polycystic Ovary Syndrome. *Endocr. Rev.* **2015**, *36*, 487–525.
6. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116.
7. Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the Proceedings of the 1993 ACM SIGMOD international conference on Management of data; ACM: Washington D.C. USA, June 1993; pp. 207–216.
8. Wu, W.-T.; Li, Y.-J.; Feng, A.-Z.; Li, L.; Huang, T.; Xu, A.-D.; Lyu, J. Data Mining in Clinical Big Data: The Frequently Used Databases, Steps, and Methodological Models. *Mil. Med. Res.* **2021**, *8*, 44, doi:10.1186/s40779-021-00338-z.
9. Samuels, J. In *One-Hot Encoding and Two-Hot Encoding: An Introduction*; 2024.
10. World Health Organization *Standards for Maternal and Neonatal Care*; **2007**.
11. World Health Organization *WHO Recommendations on Antenatal Care for a Positive Pregnancy Experience*; **2016**.
12. Ombelet, W. WHO Fact Sheet on Infertility Gives Hope to Millions of Infertile Couples Worldwide. *Facts Views Vis. ObGyn* **2020**, *12*, 249.
13. World Health Organization *Infertility Prevalence Estimates, 1990–2021*; World Health Organization, 2023; ISBN 92-4-006831-7.
14. Coussa, A.; Hasan, H.A.; Barber, T.M. Impact of Contraception and IVF Hormones on Metabolic, Endocrine, and Inflammatory Status. *J. Assist. Reprod. Genet.* **2020**, *37*, 1267–1272, doi:10.1007/s10815-020-01756-z.
15. Herman, T.; Csehely, S.; Orosz, M.; Bhattoa, H.P.; Deli, T.; Torok, P.; Lagana, A.S.; Chiantera, V.; Jakab, A. Impact of Endocrine Disorders on IVF Outcomes: Results from a Large, Single-Centre, Prospective Study. *Reprod. Sci.* **2023**, *30*, 1878–1890, doi:10.1007/s43032-022-01137-0.
16. Vannuccini, S.; Clifton, V.L.; Fraser, I.S.; Taylor, H.S.; Critchley, H.; Giudice, L.C.; Petraglia, F. Infertility and Reproductive Disorders: Impact of Hormonal and Inflammatory Mechanisms on Pregnancy Outcome. *Hum. Reprod. Update* **2016**, *22*, 104–115.
17. Wang, L.; Yu, X.; Xiong, D.; Leng, M.; Liang, M.; Li, R.; He, L.; Yan, H.; Zhou, X.; Jike, E.; et al. Hormonal and Metabolic Influences on Outcomes in PCOS Undergoing Assisted Reproduction: The Role of BMI in Fresh Embryo Transfers. *BMC Pregnancy Childbirth* **2025**, *25*, 368, doi:10.1186/s12884-025-07480-9.
18. Harrison, R.F.; Bonnar, J.; Thompson, W. *Diagnosis and Management of Tubo-Uterine Factors in Infertility*; Springer Science & Business Media, 2012; Vol. 4;.
19. Ozgur, K.; Bulut, H.; Berkkanoglu, M.; Coetzee, K.; Kaya, G. ICSI Pregnancy Outcomes Following Hysteroscopic Placement of Essure Devices for Hydrosalpinx in Laparoscopic Contraindicated Patients. *Reprod. Biomed. Online* **2014**, *29*, 113–118.
20. Qiu, J.; Du, T.; Chen, C.; Lyu, Q.; Mol, B.W.; Zhao, M.; Kuang, Y. Impact of Uterine Malformations on Pregnancy and Neonatal Outcomes of IVF/ICSI–Frozen Embryo Transfer. *Hum. Reprod.* **2022**, *37*, 428–446.
21. Tournaye, H. Male Factor Infertility and ART. *Asian J. Androl.* **2011**, *14*, 103.
22. Solanki, S.K.; Patel, J.T. A Survey on Association Rule Mining. In Proceedings of the 2015 fifth international conference on advanced computing & communication technologies; IEEE, 2015; pp. 212–216.
23. Altaf, W.; Shahbaz, M.; Guergachi, A. Applications of Association Rule Mining in Health Informatics: A Survey. *Artif. Intell. Rev.* **2017**, *47*, 313–340, doi:10.1007/s10462-016-9483-9.
24. Pradhan, G.N.; Prabhakaran, B. Association Rule Mining in Multiple, Multidimensional Time Series Medical Data. *J. Healthc. Inform. Res.* **2017**, *1*, 92–118, doi:10.1007/s41666-017-0001-x.
25. Al-Maolegi, M.; Arkok, B. An Improved Apriori Algorithm for Association Rules. *ArXiv Prepr. ArXiv14033948* **2014**.
26. Hegland, M. THE APRIORI ALGORITHM – A TUTORIAL. In *Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore*; WORLD SCIENTIFIC, 2007; Vol. 11, pp. 209–262 ISBN 978-981-270-905-9.
27. Melo, A.S.; Ferriani, R.A.; Navarro, P.A. Treatment of Infertility in Women with Polycystic Ovary Syndrome: Approach to Clinical Practice. *Clinics* **2015**, *70*, 765–769, doi:10.6061/clinics/2015(11)09.

28. Ou, H.; Sun, J.; Lin, L.; Ma, X. Ovarian Response, Pregnancy Outcomes, and Complications Between Salpingectomy and Proximal Tubal Occlusion in Hydrosalpinx Patients Before in Vitro Fertilization: A Meta-Analysis. *Front. Surg.* **2022**, *9*, 830612, doi:10.3389/fsurg.2022.830612.
29. Capmas, P.; Suarthana, E.; Tulandi, T. Management of Hydrosalpinx in the Era of Assisted Reproductive Technology: A Systematic Review and Meta-Analysis. *J. Minim. Invasive Gynecol.* **2021**, *28*, 418–441, doi:10.1016/j.jmig.2020.08.017.
30. Xu, B.; Zhang, Q.; Zhao, J.; Wang, Y.; Xu, D.; Li, Y. Pregnancy Outcome of in Vitro Fertilization after Essure and Laparoscopic Management of Hydrosalpinx: A Systematic Review and Meta-Analysis. *Fertil. Steril.* **2017**, *108*, 84–95.e5, doi:10.1016/j.fertnstert.2017.05.005.
31. Zhang, L.; Cai, H.; Li, W.; Tian, L.; Shi, J. Duration of Infertility and Assisted Reproductive Outcomes in Non-Male Factor Infertility: Can Use of ICSI Turn the Tide? *BMC Womens Health* **2022**, *22*, 480, doi:10.1186/s12905-022-02062-9.
32. Huang, C.; Shi, Q.; Xing, J.; Yan, Y.; Shen, X.; Shan, H.; Sun, H.; Mei, J. The Relationship between Duration of Infertility and Clinical Outcomes of Intrauterine Insemination for Younger Women: A Retrospective Clinical Study. *BMC Pregnancy Childbirth* **2024**, *24*, 199, doi:10.1186/s12884-024-06398-y.
33. Wang, X.; Tian, P.; Zhao, Y.; Lu, J.; Dong, C.; Zhang, C. The Association between Female Age and Pregnancy Outcomes in Patients Receiving First Elective Single Embryo Transfer Cycle: A Retrospective Cohort Study. *Sci. Rep.* **2024**, *14*, 19216, doi:10.1038/s41598-024-70249-1.
34. Alenezi, S.A.; Khan, R.; Amer, S. The Impact of High BMI on Pregnancy Outcomes and Complications in Women with PCOS Undergoing IVF – A Systematic Review and Meta-Analysis. *J. Clin. Med.* **2024**, *13*, 1578, doi:10.3390/jcm13061578.
35. Azmoodeh, A.; Pejman Manesh, M.; Akbari Asbagh, F.; Ghaseminejad, A.; Hamzehgardeshi, Z. Effects of Letrozole-HMG and Clomiphene-HMG on Incidence of Luteinized Unruptured Follicle Syndrome in Infertile Women Undergoing Induction Ovulation and Intrauterine Insemination: A Randomised Trial. *Glob. J. Health Sci.* **2015**, *8*, 244, doi:10.5539/gjhs.v8n4p244.
36. Li, S.; Liu, L.; Meng, T.; Miao, B.; Sun, M.; Zhou, C.; Xu, Y. Impact of Luteinized Unruptured Follicles on Clinical Outcomes of Natural Cycles for Frozen/Thawed Blastocyst Transfer. *Front. Endocrinol.* **2021**, *12*, 738005.
37. Liu, S.; Mo, M.; Xiao, S.; Li, L.; Hu, X.; Hong, L.; Wang, L.; Lian, R.; Huang, C.; Zeng, Y.; et al. Pregnancy Outcomes of Women With Polycystic Ovary Syndrome for the First In Vitro Fertilization Treatment: A Retrospective Cohort Study With 7678 Patients. *Front. Endocrinol.* **2020**, *11*, 575337, doi:10.3389/fendo.2020.575337.
38. Dybciak, P.; Humeniuk, E.; Raczkiwicz, D.; Krakowiak, J.; Wdowiak, A.; Bojar, I. Anxiety and Depression in Women with Polycystic Ovary Syndrome. *Medicina (Mex.)* **2022**, *58*, 942, doi:10.3390/medicina58070942.
39. Rakic, D.; Joksimovic Jovic, J.; Jakovljevic, V.; Zivkovic, V.; Nikolic, M.; Sretenovic, J.; Nikolic, M.; Jovic, N.; Bicanin Ilic, M.; Arsenijevic, P.; et al. High Fat Diet Exaggerate Metabolic and Reproductive PCOS Features by Promoting Oxidative Stress: An Improved EV Model in Rats. *Medicina (Mex.)* **2023**, *59*, 1104, doi:10.3390/medicina59061104.
40. Kusuhara, S.; Kishimoto-Kishi, M.; Matsumiya, W.; Miki, A.; Imai, H.; Nakamura, M. Short-Term Outcomes of Intravitreal Faricimab Injection for Diabetic Macular Edema. *Medicina (Mex.)* **2023**, *59*, 665, doi:10.3390/medicina59040665.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.