

Original Research Article

Clinical Profile, Treatment Outcomes and Machine-Learning Prediction of Treatment Success in Couples with Primary Infertility: A Hospital-Based cross-sectional study of 567 couples from western India

Dr. Manan S. Popat¹, Dr. Rutvi K. Patel¹, Dr. Dhvani M. Patel¹, Dr. Tanay B. Patel¹, Dr. Nilesh Kanzariya¹

¹ Department of Pharmacy Practice, SAL Institute of Pharmacy, Gujarat Technological University, Opposite Science City, Sola Bhadaj Road, Ahmedabad 380060, Gujarat, India

*Corresponding author: Dr. Manan S. Popat, Department of Pharmacy Practice, SAL Institute of Pharmacy, Opposite Science City, Sola Bhadaj Road, Ahmedabad 380060, Gujarat, India.

Running title: Predicting infertility treatment success with machine learning

Ethics approval: Institutional Ethics Committee, SAL Hospital, Ahmedabad (protocol SAL/2023-IHEC/2308, approved 25 October 2023).

Acknowledgments: None.

ABSTRACT

Background: Infertility affects roughly one in six adults worldwide, yet couples entering assisted reproduction in India are rarely given individualized estimates of the likelihood that a given treatment will succeed. Machine learning has been proposed as a way of generating such estimates, but published reports rarely test whether the models outperform the simplest possible baseline.

Objectives: To describe the clinical profile and treatment outcomes of couples presenting with primary infertility at two fertility centers in Ahmedabad, and to evaluate whether three commonly used supervised learning algorithms can predict treatment success from routinely recorded clinical variables.

Methods: Cross-sectional study combining retrospective and prospective record review over six months at Sunflower Hospital and Gayatri Hospital, Ahmedabad. Couples in which the female partner was aged 18 to 45 years and presenting with primary infertility were included (n = 567). Demographic, lifestyle, endocrine, structural and semen variables were abstracted onto a structured case report form. Treatment outcome (success or failure, as documented in the record) was analyzed for ovulation induction (OI), intrauterine insemination (IUI) and in vitro fertilization with intracytoplasmic sperm injection (IVF plus ICSI). Decision tree, logistic regression and extreme gradient boosting (XGBoost) classifiers were trained on a single random split of each treatment cohort. In addition to accuracy, sensitivity, specificity, predictive values and balanced accuracy were derived from the confusion matrices, and accuracy was compared with the no-information rate, that is, the accuracy obtained by always predicting the majority class.

Results: Obesity was present in 45.3% of women and 51.7% of men. Polycystic ovary syndrome was documented in 35.6% of women, an abnormal thyroid stimulating hormone level in 10.4% and sonographic adenomyosis in 41.8%. Among men, 11.8% had azoospermia, 27.7% a sperm concentration below 15 million/mL and 51.2% sperm motility below 40%. Success was recorded in 64 of 330 couples treated with OI (19.4%), 54 of 276 treated with IUI (19.6%), 12 of 152 treated with IVF without ICSI (7.9%) and 302 of 429 treated with IVF plus ICSI (70.4%). Headline model accuracy ranged from 69.7% to 85.7%, but in eight of nine model-treatment combinations this did not exceed the corresponding no-information rate. Balanced accuracy ranged from 48.5% to 68.1%, and recall for the minority outcome was poor (sensitivity 8.3% to 58.3% for OI and IUI success; specificity 6.7% to 42.9% for IVF plus ICSI failure).

Conclusions: This cohort shows a high burden of modifiable and endocrine risk factors among couples with primary infertility in western India. It also demonstrates that apparently high classification accuracy in small, imbalanced fertility datasets is largely an artifact of outcome prevalence. Prediction models intended for fertility counseling should be reported with discrimination and calibration measures, validated by resampling and compared against a majority-class baseline before any clinical claim is made.

Keywords: primary infertility; assisted reproductive technology; machine learning; class imbalance; in vitro fertilization; intrauterine insemination; India

1. Introduction

Infertility, defined as the failure to achieve a clinical pregnancy after twelve months or more of regular unprotected intercourse, is one of the most prevalent yet least resourced conditions in reproductive medicine. The World Health Organization

estimates a lifetime prevalence of approximately 17.5% among adults, with only modest variation between high-income (17.8%) and low- and middle-income (16.5%) settings [1,2]. The condition carries consequences that extend well beyond fecundity, including psychological distress, marital strain, social exclusion

and, particularly in South Asian settings, disproportionate blame directed at women.

Primary infertility, in which a couple has never achieved a pregnancy, is conventionally distinguished from secondary infertility, which follows a previous conception [3,4]. The distinction is clinically useful because the causal mix differs: secondary infertility is particularly common where unsafe abortion and postpartum infection lead to tubal damage [5-7], whereas primary infertility more often reflects ovulatory, endocrine or male factor causes. Male factors are estimated to contribute to about half of all cases of couple infertility [8]. Contemporary management proceeds along a ladder of increasing invasiveness and cost, from ovulation induction (OI) [9] and intrauterine insemination (IUI) [10-12] through in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI) [13,14] to gamete donation [15,16].

Each step of that ladder carries a substantial probability of failure, and the counseling that precedes it is generally based on population-level averages rather than on the individual couple. This matters most in self-funded health systems, where an unsuccessful IVF cycle imposes a direct financial burden on the family. Machine learning (ML), a branch of artificial intelligence in which algorithms learn predictive patterns from data rather than from explicitly programmed rules [17], is increasingly applied across healthcare [18-20] and has attracted interest as a means of generating individualized probabilities of treatment success. Hassan and colleagues applied multilayer perceptrons, support vector machines and tree ensembles with feature selection to IVF outcome prediction [21]; Qiu and colleagues built an XGBoost model for live birth before a first IVF cycle in 7,188 women, reporting an area under the receiver operating characteristic curve (AUC) of 0.73 [22]; and Zeadna and colleagues predicted surgical sperm retrieval in non-obstructive azoospermia with an AUC of 0.80 [23]. The practical considerations for introducing such tools into fertility care have also been reviewed [24].

Two gaps motivated the present work. First, almost all published fertility prediction models address a single modality, usually IVF, and are derived from large registries in high-income countries; comparative data across the full treatment ladder in an Indian clinic population are scarce. Second, and more fundamentally, model performance in this literature is frequently summarized by overall accuracy alone. Because treatment success is an intrinsically imbalanced outcome, accuracy can be high while the model contributes nothing beyond predicting the commoner outcome for every couple. We therefore set out both to characterize a series of couples with primary infertility and to interrogate ML performance against an explicit majority-class baseline.

The objectives were: (i) to identify female and male risk factors documented in couples presenting with primary infertility; (ii) to quantify the success and failure rates of each treatment modality used in this population; and (iii) to evaluate the performance of decision tree, logistic regression and XGBoost classifiers in predicting the outcome of the three most frequently used modalities.

2. Materials and Methods

2.1 Study design and setting

This was a cross-sectional study with both retrospective and prospective components, conducted over six months at Sunflower Hospital and Gayatri Hospital, two fertility and

gynecology centers in Ahmedabad, Gujarat, India. Retrospective data were abstracted from the medical records department; prospective data were collected from couples attending the outpatient department during the study period.

2.2 Ethical approval

The protocol was reviewed and approved by the Institutional Ethics Committee, SAL Hospital, Ahmedabad (protocol number SAL/2023-IHEC/2308, dated 25 October 2023). Data at both sites were collected under the approved protocol from the medical records and outpatient departments, with the cooperation of the hospital management. The analysis used aggregate data only, and no personally identifiable information is reported in this article.

2.3 Participants

Couples attending either center for the evaluation or treatment of primary infertility were eligible. Primary infertility was defined as failure to conceive despite regular unprotected intercourse for at least twelve months, with no previous conception in the current partnership. Couples in which the female partner was aged 18 to 45 years were included. Couples in which the female partner was under 18 or over 45 years of age, and women with a cancer diagnosis undergoing IVF, were excluded.

2.4 Sample size

The minimum sample size was calculated using the standard formula for a single proportion, $n = Z^2 p (1 - p) / e^2$, with $Z = 1.96$ (95% confidence), an assumed prevalence of successful assisted reproductive treatment of $p = 0.30$ and an absolute margin of error $e = 0.05$, giving $n = 323$ couples. The achieved sample of 567 couples exceeded this target.

2.5 Data collection and variables

Data were recorded on a structured case report form (Supplementary File 1) covering: demographic details (age, height, weight, body mass index, area of residence); family and social history (duration of marriage, family structure, education, employment, exercise, smoking, smokeless tobacco, alcohol); comorbidities and current medication; female investigations (thyroid stimulating hormone, follicle stimulating hormone, luteinizing hormone, prolactin, anti-Müllerian hormone, antinuclear antibody, fasting and random blood glucose, menstrual cycle characteristics, documented polycystic ovary syndrome, structural abnormalities on imaging); male investigations (sperm concentration, sperm motility, sperm abnormality, structural abnormality); and the treatment received with its documented outcome. Forms were completed on paper or through an electronic form and were then transferred to a spreadsheet for cleaning and analysis.

Body mass index was categorized as underweight, normal, overweight or obese according to the categories used in the source records. Abnormal hormone values were those falling outside the reference intervals stated on the case report form (thyroid stimulating hormone 0.3 to 4.0 mIU/L; follicle stimulating hormone 4.7 to 21.5 IU/L; luteinizing hormone 5 to 25 IU/L; prolactin below 25 ng/mL; anti-Müllerian hormone 2 to 6.8 ng/mL). Oligozoospermia was defined as a sperm concentration below 15 million/mL and asthenozoospermia as sperm motility below 40%, corresponding to the lower reference limits for concentration and total motility in the fifth edition of the World Health Organization laboratory manual [25]. Treatment outcome was recorded as success or failure exactly as documented by the treating clinician in the case record; where a

couple underwent more than one modality, the outcome was recorded separately for each modality, so treatment cohorts overlap and do not sum to the total cohort (Figure 1).

2.6 Machine learning analysis

Supervised classification was performed in Python 3 using scikit-learn and the xgboost package. Three algorithms were evaluated: a classification and regression tree (DecisionTreeClassifier), logistic regression (LogisticRegression) and extreme gradient boosting (XGBClassifier). The outcome variable for each treatment cohort was binary (0 = failure, 1 = success), and the couple identifier was removed before modeling. For the decision tree and logistic regression, missing predictor values were imputed with the column median, and for logistic regression predictors were additionally standardized to zero mean and unit variance; imputation, scaling and model fitting were chained in a single pipeline fitted on the training partition only. XGBoost was fitted without imputation, using its native handling of missing values.

Each treatment cohort was partitioned once, at random with a fixed seed (random_state = 42), into training and test sets. An 80:20 split was used for seven of the nine models; for the decision tree applied to IUI and to IVF plus ICSI a 70:30 split was used. All algorithms were run with library default hyperparameters. No cross-validation, hyperparameter tuning,

class weighting, resampling or feature selection was performed, and the default classification threshold of 0.5 was used. These choices are stated explicitly because they bear directly on the interpretation of the results.

Performance was summarized by overall accuracy and by the confusion matrix. From each confusion matrix we derived sensitivity (recall for treatment success), specificity (recall for treatment failure), positive and negative predictive values and balanced accuracy, defined as the mean of sensitivity and specificity. Each accuracy value was compared with the no-information rate for its own test set, that is, the proportion of the majority outcome class, which is the accuracy achievable by a trivial classifier that predicts the same outcome for every couple. Confusion matrices were read in the scikit-learn convention, in which rows are true classes ordered failure then success, so the first cell is the count of true negatives.

2.7 Statistical analysis

Categorical variables are presented as counts and percentages of the relevant denominator. Where a variable was not recorded for the whole cohort, the denominator with available data is stated. No inferential comparison between treatment modalities was undertaken because allocation to modality was clinical rather than random and the groups differ systematically in indication and severity.

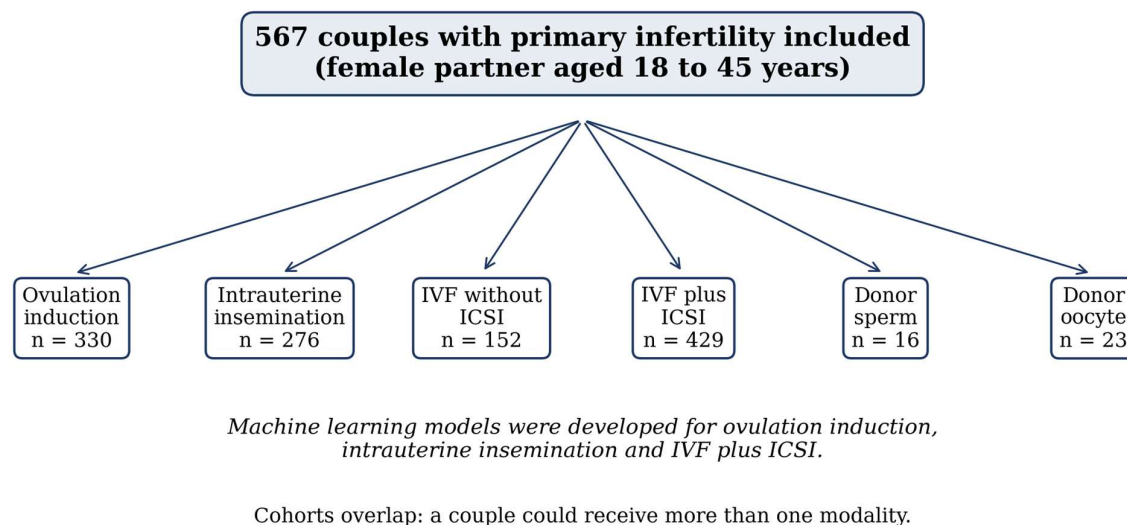


Figure 1. Couples included in the study and the number contributing to each treatment cohort. Because couples could undergo more than one modality, the treatment cohorts overlap. Machine learning models were developed for the three largest cohorts (ovulation induction, intrauterine insemination and IVF plus ICSI).

3. Results

3.1 Cohort characteristics

A total of 567 couples were included (Figure 1). Most female partners were aged 25 to 35 years (406, 71.6%), with 120 (21.2%) aged 35 to 45 years and 41 (7.2%) under 25 years. Male partners were older, with 190 (33.5%) in the 35 to 45 year band. Obesity was the single most frequent anthropometric category in both partners, affecting 257 women (45.3%) and 293 men (51.7%), while 37 women (6.5%) and 22 men (3.9%) were underweight. Duration of marriage was 0 to 5 years in 245 couples (43.2%), 5 to 10 years in 182 (32.1%) and more than 10 years in 140 (24.7%). Educational attainment was high in both partners, with graduate or postgraduate qualifications in 378 women (66.7%) and 415 men (73.2%). Family structure was documented for 128 couples, of whom 89 (69.5%) lived in a joint family. Full demographic data are given in Table 1.

Table 1. Demographic characteristics of 567 couples with primary infertility

Characteristic	Female partner, n (%)	Male partner, n (%)
Age, under 25 years	41 (7.2)	14 (2.5)
Age, 25 to 35 years	406 (71.6)	363 (64.0)
Age, 35 to 45 years	120 (21.2)	190 (33.5)
BMI, underweight	37 (6.5)	22 (3.9)
BMI, normal	145 (25.6)	106 (18.7)
BMI, overweight	103 (18.2)	94 (16.6)
BMI, obese	257 (45.3)	293 (51.7)
Education, none	17 (3.0)	9 (1.6)
Education, 12th standard or less	172 (30.3)	143 (25.2)
Education, graduate	252 (44.4)	302 (53.3)
Education, postgraduate or higher	126 (22.2)	113 (19.9)
Occupation, homemaker	386 (68.1)	0 (0.0)
Occupation, salaried employment	160 (28.2)	357 (63.0)
Occupation, business or self-employed	21 (3.7)	210 (37.0)

BMI, body mass index. Percentages are of 567.

Table 1 (continued). Couple-level characteristics

Characteristic	Couples, n (%)
Marriage duration 0 to 5 years	245 (43.2)
Marriage duration 5 to 10 years	182 (32.1)
Marriage duration over 10 years	140 (24.7)
Joint family (n = 128 with data)	89 (69.5)
Nuclear family (n = 128 with data)	39 (30.5)

Family structure was recorded for 128 of 567 couples.

Substance use showed a marked gender gradient (Table 2). Among men, 106 (18.7%) used smokeless tobacco, 76 (13.4%) consumed alcohol and 64 (11.3%) smoked. Among women, 3 (0.5%) smoked and 3 (0.5%) consumed alcohol, and no woman reported smokeless tobacco use.

Table 2. Substance use by partner

Substance	Female partner, n (%)	Male partner, n (%)
Cigarette smoking	3 (0.5)	64 (11.3)
Smokeless tobacco	0 (0.0)	106 (18.7)
Alcohol	3 (0.5)	76 (13.4)

Percentages are of 567 partners of each sex.

3.2 Comorbidities

Hypothyroidism was the commonest comorbidity in women (48, 8.5%) and was five times more frequent than in men (9, 1.6%). Hypertension (16, 2.8%) and type 2 diabetes mellitus (12, 2.1%) predominated in men. All other recorded conditions occurred in fewer than 1% of either partner (Table 3).

Table 3. Documented comorbidities

Condition	Female partner, n (%)	Male partner, n (%)
Hypothyroidism	48 (8.5)	9 (1.6)
Hypertension	8 (1.4)	16 (2.8)
Type 2 diabetes mellitus	5 (0.9)	12 (2.1)
Hyperthyroidism	5 (0.9)	2 (0.4)
Epilepsy	4 (0.7)	0 (0.0)
Depression	1 (0.2)	1 (0.2)
Arrhythmia	1 (0.2)	0 (0.0)
Systemic lupus erythematosus	1 (0.2)	0 (0.0)
Type 1 diabetes mellitus	1 (0.2)	0 (0.0)
Thalassemia minor	1 (0.2)	0 (0.0)
Ovarian fibroid	1 (0.2)	0 (0.0)
Asthma	0 (0.0)	1 (0.2)
Hepatitis B	0 (0.0)	1 (0.2)
Solitary kidney	0 (0.0)	1 (0.2)

Percentages are of 567 partners of each sex.

3.3 Female factors

An abnormal thyroid stimulating hormone concentration was the most frequent endocrine abnormality (59, 10.4%), followed by abnormal follicle stimulating hormone and luteinizing hormone (31 each, 5.5%) and hyperprolactinemia (23, 4.1%); antinuclear antibody was positive in 2 (0.4%). Polycystic ovary syndrome was documented in 202 women (35.6%). On imaging, adenomyosis was reported in 237 women (41.8%), pelvic congestion and adherent ovary in 27 each (4.8%), a vertical uterus in 18 (3.2%) and poor ovarian reserve in 10 (1.8%). Endometriosis, endometrial polyps and adherent uterus were each documented in 3 women (0.5%), and unilateral tubal block in 2 (0.4%) (Table 4).

3.4 Male factors

Azoospermia was present in 67 men (11.8%). Sperm concentration was below 15 million/mL in 155 of 560 men with a recorded value (27.7%), and sperm motility was below 40% in 283 of 553 men with a recorded value (51.2%). Reduced motility was therefore the commonest single semen abnormality in this cohort (Table 4).

Table 4. Female and male clinical findings

Finding	n	% of denominator
Female: abnormal thyroid stimulating hormone	59	10.4
Female: abnormal follicle stimulating hormone	31	5.5
Female: abnormal luteinizing hormone	31	5.5
Female: raised prolactin	23	4.1
Female: positive antinuclear antibody	2	0.4
Female: polycystic ovary syndrome	202	35.6
Female: adenomyosis on imaging	237	41.8
Female: pelvic congestion	27	4.8
Female: adherent ovary	27	4.8
Female: vertical uterus	18	3.2
Female: poor ovarian reserve	10	1.8
Female: endometriosis	3	0.5
Female: endometrial polyps	3	0.5
Female: adherent uterus	3	0.5
Female: unilateral tubal block	2	0.4
Female: uterine fibroid	1	0.2
Female: arcuate uterus	1	0.2
Female: tilted uterus	1	0.2
Male: azoospermia	67	11.8
Male: sperm concentration below 15 million/mL	155	27.7 (of 560)
Male: sperm motility below 40%	283	51.2 (of 553)

Denominator is 567 unless stated. Semen denominators reflect the number of men with a recorded value.

3.5 Treatment outcomes

Couples frequently underwent more than one modality, either sequentially or after failure, so the modality cohorts overlap. Success was documented in 64 of 330 couples treated with ovulation induction (19.4%) and in 54 of 276 treated with intrauterine insemination (19.6%). Among couples treated with IVF without ICSI, 12 of 152 succeeded (7.9%), whereas 302 of 429 couples treated with IVF plus ICSI succeeded (70.4%). Donor gamete cycles were few: 7 of 16 donor sperm cycles (43.8%) and 15 of 23 donor oocyte cycles (65.2%) were successful (Table 5, Figure 2).

The contrast between IVF alone and IVF plus ICSI is striking but should not be read as a treatment effect. The two groups were not randomized and differ in indication (ICSI being selected predominantly for male factor infertility, which in this cohort was common), and the IVF-alone group is small and may be enriched for couples in whom conventional insemination had already failed. No adjustment for female age, ovarian reserve or cycle number was possible with the recorded data.

Table 5. Treatment outcomes by modality

Modality	Couples treated	Success, n (%)	Failure, n (%)
Ovulation induction	330	64 (19.4)	266 (80.6)
Intrauterine insemination	276	54 (19.6)	222 (80.4)
IVF without ICSI	152	12 (7.9)	140 (92.1)
IVF plus ICSI	429	302 (70.4)	127 (29.6)
Donor sperm	16	7 (43.8)	9 (56.2)
Donor oocyte	23	15 (65.2)	8 (34.8)

ICSI, intracytoplasmic sperm injection; IVF, in vitro fertilization. Couples could receive more than one modality, so the cohorts overlap and do not sum to 567. Outcome is per couple per modality, not per cycle.

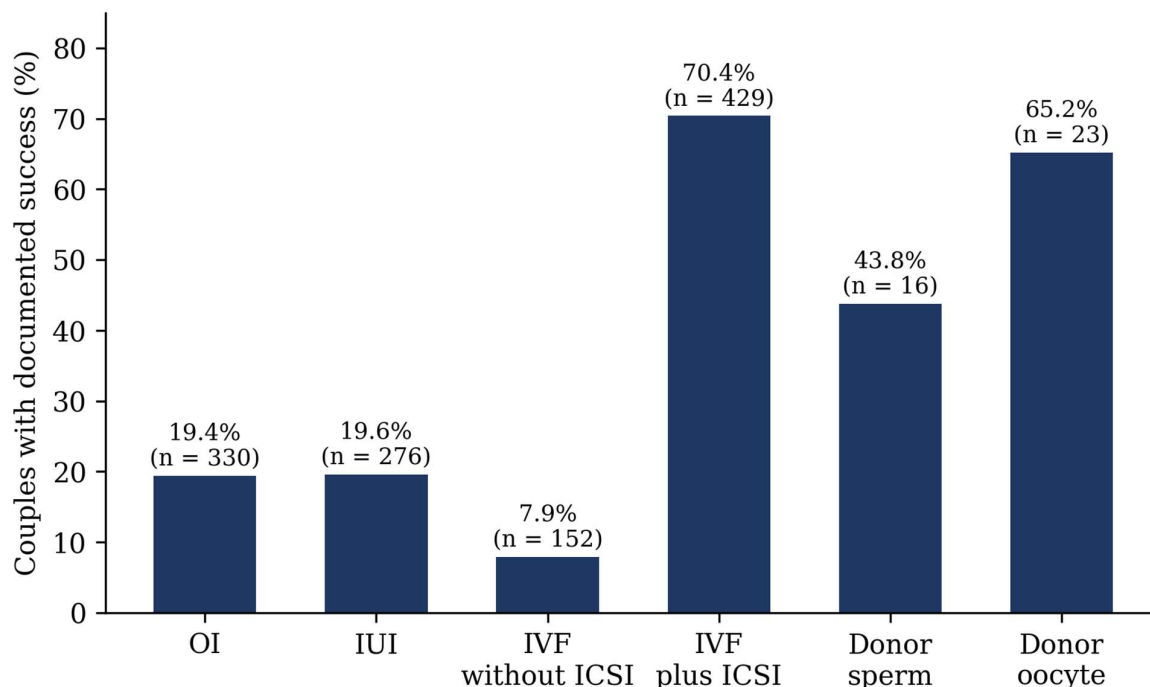


Figure 2. Documented treatment success by modality among 567 couples with primary infertility. Bars show the percentage of couples with a successful outcome, with the number of couples treated shown above each bar. OI, ovulation induction; IUI, intrauterine insemination; IVF, in vitro fertilization; ICSI, intracytoplasmic sperm injection.

3.6 Machine learning performance

Headline accuracies appeared favorable, ranging from 69.7% (decision tree, ovulation induction) to 85.7% (XGBoost, intrauterine insemination). When each accuracy was placed alongside the no-information rate of its own test set, however, the picture changed substantially (Table 6, Figure 3). For ovulation induction the majority outcome (failure) accounted for 81.8% of the test set, and none of the three models reached this figure: logistic regression came closest at 80.3%. For intrauterine insemination the no-information rate was 84.3% for the 70:30 split and 85.7% for the 80:20 split; the decision tree exceeded its baseline by 1.2 percentage points and XGBoost exactly matched its baseline. For IVF plus ICSI, where success was the majority outcome, all three models fell below the corresponding baseline.

Class-specific performance confirmed this interpretation. In the two treatments in which success was uncommon, sensitivity for success was low: 25.0%, 41.7% and 58.3% for the decision tree, logistic regression and XGBoost respectively in ovulation induction, and 23.1%, 8.3% and 25.0% in intrauterine insemination. Positive predictive value for success ranged from 16.7% to 60.0%. Conversely, in IVF plus ICSI, where failure was the minority outcome, the models identified success well (sensitivity 80.0% to 91.9%) but failure poorly (specificity 6.7% for the decision tree, 21.7% for logistic regression and 42.9% for XGBoost), with negative predictive values of 20.0%, 45.5% and 40.9% respectively. Balanced accuracy, which weights both classes equally, ranged from 48.5% to 68.1%; the value of 48.5% for logistic regression in intrauterine insemination is below the 50% expected by chance. Across all nine models the highest balanced accuracy was 68.1% (XGBoost, ovulation induction).

Table 6. Model performance against the majority-class baseline

Treatment	Model	Test n	Accuracy %	No-information rate %	Sens. %	Spec. %	Balanced acc. %
Ovulation induction	Decision tree	66	69.7	81.8	25.0	79.6	52.3
Ovulation induction	Logistic regression	66	80.3	81.8	41.7	88.9	65.3
Ovulation induction	XGBoost	66	74.2	81.8	58.3	77.8	68.1
Intrauterine insemination	Decision tree	83	85.5	84.3	23.1	97.1	60.1
Intrauterine insemination	Logistic regression	56	71.4	78.6	8.3	88.6	48.5
Intrauterine insemination	XGBoost	56	85.7	85.7	25.0	95.8	60.4
IVF plus ICSI	Decision tree	129	72.1	76.7	91.9	6.7	49.3
IVF plus ICSI	Logistic regression	86	72.1	73.3	90.5	21.7	56.1
IVF plus ICSI	XGBoost	86	70.9	75.6	80.0	42.9	61.4

Sens., sensitivity (recall for treatment success); Spec., specificity (recall for treatment failure); Balanced acc., mean of sensitivity and specificity. The no-information rate is the proportion of the majority outcome in that test set, that is, the accuracy of always predicting the commoner outcome. Test partitions were 80:20 except for the decision tree applied to intrauterine insemination and to IVF plus ICSI (70:30). Values are derived from the confusion matrices read in the scikit-learn row order (true negatives first).

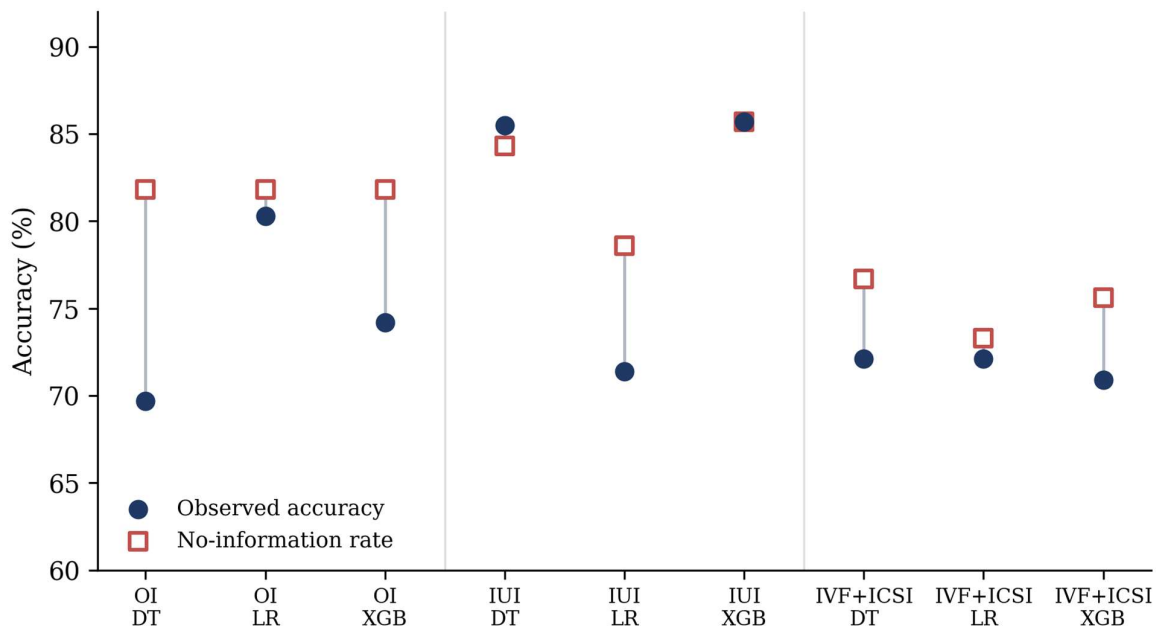


Figure 3. Observed accuracy (filled circles) and no-information rate (open squares) for each of the nine model-treatment combinations. Vertical lines join each pair. Where the circle lies on or below the square, the model offers no gain over always predicting the majority outcome. DT, decision tree; LR, logistic regression; XGB, XGBoost.

4. Discussion

4.1 Principal findings

In 567 couples presenting with primary infertility at two Ahmedabad fertility centers, we found a high prevalence of modifiable and endocrine risk factors: obesity in about half of both partners, polycystic ovary syndrome in more than a third of women, thyroid dysfunction in one in ten, and impaired sperm motility in half of the men assessed. Treatment success per couple was modest for ovulation induction and intrauterine insemination (about one in five) and considerably higher for IVF combined with ICSI. Three supervised learning algorithms achieved accuracies that superficially appeared clinically useful, but in eight of nine model-treatment combinations they did not exceed the accuracy of simply predicting the majority outcome for every couple, and their balanced accuracy was close to chance.

4.2 Cohort profile in context

Obesity affected 45.3% of women and 51.7% of men in this cohort, and adiposity is a recognized contributor to ovulatory dysfunction, insulin resistance and impaired spermatogenesis [26,27]. Because the sample is drawn from couples seeking fertility care rather than from the general population, the finding should be read as a description of the treated population and not as an estimate of causal contribution. Nevertheless, it identifies weight optimization as the single most widely applicable pre-treatment intervention in this group.

Documented polycystic ovary syndrome in 35.6% of women [28] is more than three times the pooled community prevalence of about 10% by Rotterdam criteria reported for Indian women [29], as would be expected in a population selected for infertility. Hypothyroidism in 8.5% of women, together with an abnormal thyroid stimulating hormone value in 10.4%, reinforces the case for thyroid screening at the first infertility consultation; thyroid dysfunction also affects male reproductive function [30,31]. The reported frequency of adenomyosis (41.8%) is considerably higher than in most published infertility series [32,33] and almost certainly reflects liberal sonographic reporting of myometrial heterogeneity rather than histologically confirmed disease. By contrast, uterine fibroids, the commonest benign uterine tumor [34], were documented in only one woman. The adenomyosis figure should therefore be interpreted with caution and is best regarded as a marker of imaging practice at the participating centers.

Among men, azoospermia in 11.8% falls within the 10% to 15% range usually reported among infertile men [35]. Reduced sperm motility was the dominant semen abnormality, present in half of those tested. Taken with the frequency of smokeless tobacco and alcohol use among male partners, this supports routine male lifestyle counseling as part of couple-level management rather than treating the male partner as a passive contributor.

4.3 Interpretation of the prediction models

The central methodological message of this study is that accuracy is an unsafe summary statistic for fertility outcome prediction. When one outcome accounts for 80% of a cohort, a model that always predicts that outcome achieves 80% accuracy while providing no information about any individual couple. Our models approached, but with one marginal exception did not surpass, that trivial benchmark. Balanced accuracy and class-specific recall exposed the same problem from the opposite direction: in cohorts where success was rare the models rarely identified success, and in the cohort where failure was rare they rarely identified failure. The clinically decisive question, namely which couples will not benefit from the proposed treatment, was answered poorly by every model tested.

This is not an argument against machine learning in reproductive medicine. It is an argument for the reporting standards already adopted in the better studies in this field. Qiu and colleagues reported an AUC of 0.73 with a stated accuracy, and Zeadna and colleagues reported an AUC of 0.80 with sensitivity and specificity and a formally derived cut-off, both of which permit a reader to judge discrimination independently of prevalence [22,23]. Hassan and colleagues combined feature selection with several classifiers and reported performance across them [21]. By contrast, single-split accuracy from default-parameter models on a few hundred records, as presented here, is a proof of feasibility and not evidence of clinical utility.

Several specific design features limited performance and are readily remediable. A single random train-test split makes every estimate dependent on one partition, and with test sets of 56 to 129 cases the confidence intervals around each accuracy are wide. Default hyperparameters and an unweighted loss function push tree-based and regression models towards the majority class. The classification threshold was left at 0.5, whereas in counseling the useful output is a calibrated probability rather than a hard label. Finally, several predictors that carry most of the prognostic weight in the published literature, notably anti-Müllerian hormone, antral follicle count, number of mature oocytes retrieved and embryo grade, were incompletely recorded and could not contribute.

4.4 Clinical implications

For clinicians, the practical yield of this study lies in its descriptive component. Roughly four in five couples undergoing ovulation induction or intrauterine insemination in this setting did not conceive from that treatment, a figure that should be communicated plainly before treatment begins so that expectations, and the financial planning that follows from them, are realistic. The prominence of obesity, thyroid dysfunction and male lifestyle factors identifies concrete pre-treatment targets. For those developing prediction tools, the study is a caution consistent with wider concerns about validation and accountability of artificial intelligence in healthcare [20]: before a model is offered to a couple, it should be shown to beat the majority-class baseline, to discriminate as measured by AUC, to be calibrated, and to hold up under cross-validation and preferably external validation.

4.5 Strengths and limitations

Strengths include a sample well above the calculated requirement, coverage of the whole treatment ladder rather than a single modality, uniform data capture through a structured case report form, and a transparent re-analysis of model performance against an explicit baseline.

The limitations are substantial and are stated in full. First, the study is observational and confined to two centers in one city, so the findings may not generalize. Second, the outcome variable was abstracted as recorded by the treating clinician and was not standardized as biochemical pregnancy, clinical pregnancy or live birth; live birth, the outcome that matters most to couples, was not captured. Third, outcomes were analyzed per couple per modality rather than per cycle, so couples undergoing repeated cycles are not distinguished from those undergoing one, and modality cohorts overlap. Fourth, comparisons between modalities are confounded by indication and cannot support causal inference. Fifth, several key predictors were missing or incompletely recorded, and missing values were imputed by the median for two of the three models, which attenuates variance. Sixth, no cross-validation, tuning, class weighting or external validation was performed, and no calibration or AUC was computed, so discrimination cannot be reported. Seventh, the retrospective component is subject to documentation bias, and the six-month window limits the duration of follow-up.

4.6 Future directions

A logical next step is a prospective, multicenter cohort with a predefined outcome of live birth per initiated cycle, complete capture of ovarian reserve and embryology variables, and a prespecified modeling plan using repeated stratified cross-validation, class-weighted or resampled training, probability calibration and reporting of AUC with confidence intervals alongside decision-curve analysis. Adherence to the TRIPOD reporting guideline for prediction model studies [36] would allow direct comparison with the international literature and would make any resulting tool auditable before it reaches the consultation room.

5. Conclusion

Couples presenting with primary infertility in this western Indian cohort carried a heavy burden of modifiable and endocrine risk: obesity in about half of both partners, polycystic ovary syndrome in over a third of women, thyroid dysfunction in one in ten and impaired sperm motility in half of the men. Roughly one in five couples conceived following ovulation induction or intrauterine insemination, and the majority conceived following IVF combined with ICSI, although that comparison is confounded by indication. Decision tree, logistic regression and XGBoost classifiers trained on routinely recorded variables produced accuracies of 69.7% to 85.7% that did not, with one marginal exception, exceed the accuracy obtainable by always predicting the commoner outcome. Machine learning may yet contribute usefully to fertility counseling, but only when models are evaluated against a majority-class baseline and reported with discrimination, calibration and validation measures rather than accuracy alone.

Declarations

Ethics approval and consent to participate. Approved by the Institutional Ethics Committee, SAL Hospital, Ahmedabad (SAL/2023-IHEC/2308, 25 October 2023). Data collection at both sites was carried out under the approved protocol with the cooperation of the hospital management.

Consent for publication. Not applicable; no individually identifiable data are reported.

Availability of data and materials. The de-identified dataset and the analysis scripts are available from the corresponding author on reasonable request.

Competing interests. The authors declare that they have no competing interests.

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions. MSP, RKP, DMP and TBP carried out the study, including data collection, data curation and analysis, under the guidance and supervision of NK. MSP drafted the manuscript. All authors reviewed the manuscript and approved the final version.

Acknowledgments. The authors thank Dr. R. G. Patel and Ms. Maya Bohra (Sunflower Hospital) and Dr. Aesha Patel (Gayatri Hospital) for facilitating data access, and the Institutional Ethics Committee, SAL Hospital, for its review.

References

- World Health Organization. Infertility [fact sheet]. Geneva: WHO; 2023.
- World Health Organization. Infertility prevalence estimates, 1990-2021. Geneva: WHO; 2023.
- World Health Organization. Revised glossary on assisted reproductive terminology (ART). Geneva: WHO.
- Magdum M, Chowdhury MAT, Begum N, Riya S. Types of infertility and its risk factors among infertile women: a prospective study in Dhaka city. *J Biosci Med.* 2022;10(4):158-68.
- Vander Borgh M, Wyns C. Fertility and infertility: definition and epidemiology. *Clin Biochem.* 2018;62:2-10.
- Buckett W, Bentick B. The epidemiology of infertility in a rural population. *Acta Obstet Gynecol Scand.* 1997;76(3):233-7.
- McCormack WM. Pelvic inflammatory disease. *N Engl J Med.* 1994;330(2):115-9.
- Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. *Reprod Biol Endocrinol.* 2015;13:37.
- Legro RS, Brzyski RG, Diamond MP, Coutifaris C, Schlaff WD, Casson P, et al. Letrozole versus clomiphene for infertility in the polycystic ovary syndrome. *N Engl J Med.* 2014;371(2):119-29.
- Bhattacharya S, Harrild K, Mollison J, Wordsworth S, Tay C, Harrold A, et al. Clomifene citrate or unstimulated intrauterine insemination compared with expectant management for unexplained infertility: pragmatic randomised controlled trial. *BMJ.* 2008;337:a716.
- Cantineau AEP, Janssen MJ, Cohlen BJ. Synchronised approach for intrauterine insemination in subfertile couples. *Cochrane Database Syst Rev.* 2010;(4):CD006942.
- van Rumste MME, Custers IM, van der Veen F, van Wely M, Evers JH, Mol BWJ. The influence of the number of follicles on pregnancy rates in intrauterine insemination with ovarian stimulation: a meta-analysis. *Hum Reprod Update.* 2008;14(6):563-70.
- Boulet SL, Mehta A, Kissin DM, Warner L, Kawwass JF, Jamieson DJ. Trends in use of and reproductive outcomes associated with intracytoplasmic sperm injection. *JAMA.* 2015;313(3):255-63.
- Centers for Disease Control and Prevention. ART success rates. Atlanta: CDC.
- Martinez F, Racca A, Rodriguez I, Polyzos NP. Ovarian stimulation for oocyte donation: a systematic review and meta-analysis. *Hum Reprod Update.* 2021;27(4):673-96.
- Allen CP, Marconi N, McLernon DJ, Bhattacharya S, Maheshwari A. Outcomes of pregnancies using donor sperm compared with those using partner sperm: systematic review and meta-analysis. *Hum Reprod Update.* 2021;27(1):190-211.
- El Naqa I, Murphy MJ. What is machine learning? In: El Naqa I, Li R, Murphy MJ, editors. *Machine learning in radiation oncology*. Cham: Springer; 2015. p. 3-11.
- Habeh H, Gohel S. Machine learning in healthcare. *Curr Genomics.* 2021;22(4):291-300.
- Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: features, pillars and applications. *Int J Intell Netw.* 2022;3:58-73.
- Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Abbas J, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomed Mater Devices.* 2023;1(2):731-8.
- Hassan MR, Al-Insaf S, Hossain MI, Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. *Neural Comput Appl.* 2020;32(7):2283-97.
- Qiu J, Li P, Dong M, Xin X, Tan J. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. *J Transl Med.* 2019;17(1):317.
- Zeadna A, Khateeb N, Rokach L, Lior Y, Har-Vardi I, Harlev A, et al. Prediction of sperm extraction in non-obstructive azoospermia patients: a machine-learning perspective. *Hum Reprod.* 2020;35(7):1505-14.
- Swain J, VerMilyea MT, Meseguer M, Ezcurra D. AI in the treatment of fertility: key considerations. *J Assist Reprod Genet.* 2020;37(11):2817-24.
- World Health Organization. WHO laboratory manual for the examination and processing of human semen. 5th ed. Geneva: WHO; 2010.
- International Institute for Population Sciences (IIPS), ICF. National Family Health Survey (NFHS-5), 2019-21: India. Mumbai: IIPS; 2021.
- Deyhoul N, Mohamaddoost T, Hosseini M. Infertility-related risk factors: a systematic review. *Int J Womens Health Reprod Sci.* 2017;5(1):24-9.
- Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. *Nat Rev Endocrinol.* 2018;14(5):270-84.
- Bharali MD, Rajendran R, Goswami J, Singal K, Rajendran V. Prevalence of polycystic ovarian syndrome in India: a systematic review and meta-analysis. *Cureus.* 2022;14(12):e32351.
- Krajewska-Kulak E, Sengupta P. Thyroid function in male infertility. *Front Endocrinol (Lausanne).* 2013;4:174.
- Sengupta P, Dutta S, Karkada IR, Chinni SV. Endocrinopathies and male infertility. *Life (Basel).* 2022;12(1):10.
- Campo S, Campo V, Benagiano G. Infertility and adenomyosis. *Obstet Gynecol Int.* 2012;2012:786132.
- Munro MG. Uterine polyps, adenomyosis, leiomyomas, and endometrial receptivity. *Fertil Steril.* 2019;111(4):629-40.
- Stewart EA. Uterine fibroids. *Lancet.* 2001;357(9252):293-8.
- Koh E, Sin HS, Fukushima M, Namiki M. Azoospermia factor and male infertility. *Reprod Med Biol.* 2010;9(3):129-37.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594.

Supplementary Material

Supplementary File 1. Case report form used for data collection.

Supplementary File 2. Python scripts for the logistic regression, decision tree and XGBoost models.