JENRS

Received: 11 April 2026, Revised: 17 July 2026, Accepted: 01 August 2026, Online: 12 August 2026

DOI: https://doi.org/10.55708/js0508001

A Step Toward the Use of AI for Polyendocrine Metabolic Ovarian Syndrome (PMOS): Cost-Aware Risk Stratification with Uncertainty-Aware Triage and Conformal Prediction

Adewale Alex Adegoke[1], Idris Babalola[*2], Peter Adebayo Odesola[3]

[1] Office for National Statistics, United Kingdom

[2] Department of Health and Social Care, London, United Kingdom

[3] Southampton Solent University, Southampton, United Kingdom

Email(s): adegokeaa44@gmail.com (A. A. Adegoke), peterodes27@gmail.com (P. A. Odesola)

*Corresponding author: Idris Babalola, Department of Health and Social Care, London, United Kingdom, eidreiz01@gmail.com

ABSTRACT: Polyendocrine metabolic ovarian syndrome (PMOS) previously known as Polycystic Ovary Syndrome (PCOS) is a common endocrine disorder affecting women of reproductive age, yet its diagnosis remains challenging because accurate assessment often depends on investigations that are costly, less accessible, and difficult to scale in resource-constrained clinical settings. We develop and assess a cost-aware modelling framework for PCOS risk stratification, comparing an inexpensive-feature model based on demographic, symptom, anthropometric, and routine clinical variables with an augmented model that additionally incorporates laboratory and ultrasound features when fuller diagnostic work-up is available. Logistic Regression (LR) and Random Forest (RF) models are evaluated using discrimination and calibration metrics, including AUC, accuracy, F1, precision, recall, Brier score, and expected calibration error (ECE), alongside clinical utility through decision curve analysis (DCA). To support safer decision-making under constrained capacity, we further integrate conformal prediction to provide finite-sample coverage guarantees and controlled abstention.

Across train-test and out-of-fold evaluations, the augmented model showed consistent performance gains over the inexpensive-feature model, with AUC increasing by 6.9% for LR and 7.4% for RF. LR exhibited more favourable calibration in most settings, while RF achieved higher precision. Decision curve analysis showed higher net benefit for the augmented model across clinically relevant threshold regions, although feature sensitivity analysis indicated that a compact subset of inexpensive predictors preserved over 80% of maximal AUC, supporting the practical value of lower-burden screening. Conformal prediction achieved 94.5% overall coverage with a 41.3% abstention rate, maintaining near-nominal validity across age and BMI subgroups. Under constrained referral capacity, prioritising highest-risk cases maximised net benefit at lower capacity levels, while uncertainty-aware deferral became more compatible with decision utility as available capacity increased. These findings support a proof-of-concept framework in which inexpensive information provides meaningful early risk stratification, while added laboratory and ultrasound inputs offer incremental value when more resource-intensive assessment is feasible.

KEYWORDS: PMOS, cost-aware risk stratification, inexpensive-feature model, uncertainty-aware triage, conformal prediction, clinical utility

1. Introduction

1.1. Background

In accordance with the 2026 international consensus renaming of polycystic ovary syndrome, this manuscript adopts the term polyendocrine metabolic ovarian syndrome (PMOS) [1]. The abbreviation PMOS is used in the title and abstract to reflect this updated terminology. However, PCOS is retained in the main text where appropriate to maintain consistency with earlier literature, existing diagnostic guidelines, and source dataset labels.

Polycystic ovary syndrome (PCOS) stands as the most common endocrine disorder affecting reproductive-aged women worldwide, with prevalence estimates ranging from 4% to 21% in earlier studies, depending on diagnostic criteria used [2]. Later investigations have reported higher estimates ranging 5% to 26% [3]. This trend demonstrates the increasing attention of the condition and investigation into the global burden of PCOS The syndrome is a leading cause of anovulatory infertility, accounting for over 75% of cases related to ovulation disorders [4].

Beyond reproductive health, PCOS is associated with significant metabolic and cardiometabolic risks, including insulin resistance, type 2 diabetes, obesity and cardiovascular disease [5], [6]. These comorbidities threaten long-term individual health and place considerable socioeconomic burden on healthcare systems. Additionally, delayed diagnosis often because of heterogeneous symptom profiles and limitations of conventional diagnostic criteria restrains early intervention strategies aimed at mitigating long-term outcomes [3]. An effective predictive framework that can swiftly identify high-risk individuals presents a critical opportunity for clinical intervention, resource optimisation and improved patient outcomes.

In PCOS, artificial intelligence and machine learning are increasingly being explored to improve diagnostic and classification performance and support earlier detection [7]. In this study, the practical question is not whether a model performs well under full-information conditions, but how much useful decision support can already be obtained from information that is inexpensive, routinely available and feasible to collect at scale. This consideration is especially important in PCOS, where confirmatory laboratory and ultrasound assessments may be delayed, unevenly available or financially burdensome, thereby motivating a framework that compares an inexpensive-feature model with an augmented model that incorporates more resource-intensive diagnostics when available.

1.2. Gaps in research

While machine learning (ML) applications for PCOS risk prediction are growing in number, they often focus primarily on maximizing diagnostic accuracy using standard performance metrics such as AUC, precision and recall without assessing whether these models deliver meaningful clinical value. A systematic review noted high reported performance in PCOS detection (AUCs between 73% and 100%, diagnostic accuracy up to 100% and positive predictive value (PPV)/negative predictive value (NPV) frequently exceeding 90%) [7]. However, these studies typically lack evaluations based on decision-analytic frameworks that reflect real clinical trade-offs. Moreover, existing work often neglects probabilistic calibration, critical when translating predictions into actionable risk thresholds. Methods to handle predictive uncertainty remain largely unexplored in PCOS contexts, although conformal prediction has shown promise for delivering reliable set-valued predictions in healthcare [8]. Likewise, Decision Curve Analysis (DCA) which allows quantification of net clinical benefit across threshold strategies is notably absent from PCOS-focused ML studies [9].

Another important but under-addressed area is operational realism. In many clinical settings, capacity constraints limit how many patients can receive intensive evaluation or screening. Yet, prior PCOS prediction models rarely consider triage strategies under capacity limitations, which can significantly influence utility and practicality. There is limited emphasis on feature cost-efficiency; few studies assess whether omitting low-importance features compromises performance, a key consideration for low-resource implementation. Recent PCOS prediction with machine learning has progressed across three data modalities: self-reported/clinical features, electronic health records (EHRs) and imaging; yet most studies emphasize discrimination (accuracy/AUC) over calibration, decision-analytic value, or uncertainty aware deployment. In a large EHR study [10], gradient boosting, logistic regression, SVM and random forests on 30,601 at-risk outpatients and reported AUCs of 0.87-0.91 for prevalent PCOS and 0.82-0.84 for incident PCOS, highlighting scalability to routine care but without decision curve analysis (DCA) or coverage-guaranteed uncertainty estimates.

On imaging, [11] built a deep-learning pipeline on ovarian ultrasound features and achieved 99.89% accuracy in a 1,000-image benchmark, demonstrating near-ceiling discrimination while leaving calibration and clinical utility assessment unexplored. A hospital-based multimodal study [12] fused ovarian ultrasound images with patient-level clinical data and achieved 82.46% classification accuracy, but evaluation remained centered on conventional predictive performance metrics, without reported probability calibration, decision-curve net benefit, or subgroup reliability analysis. Another work by [13] pursued “non-invasive” diagnosis from clinical plus ultrasound features, with XGBoost reaching AUC = 0.995 (Clinical+USG+AMH) and even perfect test-set

performance (AUC = 1.00) on a small external set, which the authors themselves caution may reflect overfitting/data leakage, yet calibration, DCA and abstention policies were not explored.

Furthermore, several works develop patient-facing or clinician-support tools from non-invasive or questionnaire-like variables [14] proposed "machine-aided self-diagnostic" PCOS models from anthropometrics and symptoms to enable pre-consultation screening; they prioritized convenience and explainability but did not report probability calibration, DCA, or abstention policies. In a study by [15], they used correlation-based feature selection to reduce inputs and trained LR/NB/SVM, the SVM achieved 94.44% accuracy on the clinical subset, while a VGG16 modelling pipeline on ultrasound achieved 98.29% validation accuracy again without calibration or decision-analytic evaluation. In [16], the authors built hormone-based diagnostic models combining AMH and steroid/estrogen panels, the logistic classifier achieved test-set ROC-AUC = 0.934 (specificity = 0.947; F1 = 0.864), while XGBoost maximized accuracy (0.905) and sensitivity (0.870), yet no calibration checks, decision curve analysis, or uncertainty triage were reported.

A substantial stream of studies relies on the widely used 541-patient Kerala clinical dataset [17] (the same open dataset used in our study), typically targeting binary PCOS classification with conventional metrics. In [18], the authors introduced an "explainable multi-stacking" framework and reported ~98% accuracy using feature-selected clinical predictors and SHAP-based explanations, but did not assess calibration, capacity-constrained triage, or uncertainty aware abstention. In [19], the authors built a web-based pipeline on the same dataset, where AdaBoost and Random Forest each reached 94% accuracy with F1 up to 0.91; despite a clear engineering contribution, calibration analysis and DCA were not performed. A comparative study by [20] evaluated seven supervised learners on the same Kerala PCOS dataset and reported Logistic Regression as best overall (accuracy = 91.7%, precision = 96.0%, ROC-AUC = 96.8%), but it did not assess calibration, decision-analytic utility, uncertainty, or subgroup reliability leaving open questions central to cost-aware deployment and safe abstention. A Smart Polycystic Ovary Syndrome diagnostic system (SPOSDS) optimized feature selection and classifiers [21] was conducted on the same dataset and reported ~93.25% accuracy yet omitted probability calibration diagnostics and decision-analytic evaluation.

In addition, [22] systematically compared traditional and ensemble classifiers on the Kerala dataset, underscoring the value of feature selection but again without calibration reporting or uncertainty aware triage. [23] previously evaluated logistic regression and random forests on the dataset in IEEE TENSYMP, reporting ~91% accuracy but no assessment of reliability (e.g., calibration curves, Brier score) or clinical net benefit. In [24], the authors examined eight traditional and ensemble models on the 541-record dataset, highlighting very high headline scores (e.g., up to 99.78% for LR/SVM in their web supplement) without external validation, calibration, decision-curve, or uncertainty aware triage, highlighting overfitting risks for clinical translation. In [25], the researchers proposed an optimized stacked ensemble (WaOEL) for "early, inexpensive" PCOS screening from symptomatic, attaining accuracy = 92.8% and AUC = 0.93 on a blended low-cost feature set that included the Kaggle PCOS resource; however, the study emphasized accuracy without calibration, capacity-constrained triage, or formal error-rate control. In [26], the authors introduced a stacked-learning pipeline with ADASYN/SMOTE rebalancing on the same Kaggle dataset and reported 97% accuracy, but again did not quantify probability calibration, net benefit under operational caps, or coverage guarantees. Overall, prior work rarely evaluates models under capacity constraints (e.g., fixed slots for further testing), seldom implements triage policies that integrate model confidence and almost never use formal uncertainty quantification with coverage guarantees in PCOS.

Conformal prediction (CP) has emerged in clinical AI as a distribution-free framework to deliver set-valued predictions with finite-sample coverage, enabling principled abstention; however, CP has not been incorporated into PCOS risk tools. A recent study [27] introduced practical CP methods with coverage guarantees but did not target PCOS specifically. A clinical review in [8] catalogued CP applications across medical areas and argued for CP to quantify per-case reliability, yet no PCOS application was identified.

Methodological elements needed to bridge research to practice such as DCA for net benefit and calibration for trustworthy probabilities also remain underused in PCOS modeling. DCA is a standard way to quantify whether model-guided action improves decisions over "treat all" or "treat none but PCOS studies rarely report it [28]. Even where discrimination is strong (e.g., ultrasound or stacked clinical models), miscalibration can undermine threshold-based actionability [7], a concern not addressed in the above works.

In summary, existing PCOS prediction studies report strong discrimination across datasets and modalities, but they generally leave several clinically important questions unresolved. First, they rarely evaluate whether lower-burden clinical information can provide sufficient first-line risk stratification before escalation to laboratory and ultrasound assessment, limiting their usefulness in resource-constrained settings. Second, they seldom assess

probability calibration or decision-analytic value, even though poor calibration can misalign risk thresholds and weaken clinical actionability. Third, uncertainty-aware triage under fixed capacity constraints is rarely examined, despite its relevance when only a subset of patients can be referred for more intensive evaluation. Finally, conformal prediction has not been incorporated into PCOS risk tools to maintain coverage while allowing controlled abstention across patient subgroups. These gaps motivate our integrated framework, which compares an inexpensive-feature model with an augmented model, evaluates clinical utility under constrained capacity, and embeds conformal prediction as the uncertainty-control layer of the decision pathway.

1.3. Methodological Advances Needed

To address these gaps, an improved framework for PCOS risk stratification should combine discrimination, calibration, uncertainty quantification, decision-analytic evaluation, and feature cost-efficiency within a coherent workflow. Model performance should not be judged by accuracy alone. In settings where clinical action depends on threshold probabilities, probability quality matters, and calibration metrics such as the Brier score and expected calibration error (ECE) become essential for assessing whether predicted risks are trustworthy [29]. Likewise, DCA provides a clinically interpretable way to determine whether model-guided action offers greater net benefit than default strategies such as treating all or treating none [28].

In addition, PCOS assessments often take place under constraints in which laboratory assays and ultrasound are not immediately available for all patients [2], [4]. This makes it important to evaluate how much predictive value can be obtained from inexpensive, routinely available information before escalation becomes necessary, and what incremental value is gained when more resource-intensive diagnostics are added. Feature importance analysis grouped feature reduction and targeted ablation therefore have an important role in identifying which variables support lower-burden screening and which added diagnostics provide meaningful marginal benefit [13], [18], [22].

Finally, uncertainty-aware prediction frameworks such as conformal prediction provide a principled way to generate prediction sets with finite-sample coverage guarantees and to defer ambiguous cases when confidence is insufficient [27]. When combined with capacity-constrained triage policies, this allows the evaluation of not only which model performs best, but how model predictions may be translated into safer and more operationally realistic referral decisions. These elements define the methodological basis for the framework developed in this study.

1.4. Research Question

This study is guided by three core research questions that together address predictive performance, operational feasibility, and safety in AI-assisted PCOS risk stratification:

1. How can a cost-aware modelling framework that compares an inexpensive-feature model with an augmented model using inexpensive plus expensive features support improvements in predictive performance for PCOS while maintaining practical relevance under resource constraints?
2. Under real-world capacity constraints, how do risk-based, utility-based, and uncertainty-aware triage strategies differ in terms of net clinical benefit, and under what capacity conditions is uncertainty-aware referral most useful?
3. How can conformal prediction maintain high coverage and controlled abstention rates across PCOS patient subgroups, and how can this contribute to safer and more trustworthy AI-assisted decision-making?

These questions form the backbone of our investigation, linking model development with downstream clinical utility, uncertainty handling, and responsible use.

1.5. Contribution and Novelty of This Work

This study presents a proof-of-concept framework for AI-assisted PCOS risk stratification that integrates performance evaluation, cost considerations, uncertainty handling, and clinical utility within a single workflow. The contribution of this work lies in the operational integration of these components to assess how PCOS risk models may function under realistic diagnostic and capacity constraints, with emphasis on practical implementation, clinical utility, and deployment considerations.

The contributions to our work as follows:

1. By comparing an inexpensive-feature model with an augmented model that adds laboratory and ultrasound variables, we quantify how much predictive value can be obtained from lower-burden clinical information and what incremental benefit is gained when more resource-intensive diagnostics are available. Feature importance and sensitivity analyses further show that a compact subset of inexpensive predictors preserves over 80% of maximal AUC, supporting cost-efficient early risk stratification [30], [31].
2. We extend model evaluation beyond conventional accuracy metrics by incorporating Decision Curve Analysis (DCA) to quantify net clinical benefit across plausible decision thresholds. This enables predictive

performance to be interpreted in terms of its potential clinical consequences and decision-making value, an approach that remains rarely applied in PCOS-focused modelling studies [32].

3. We evaluate capacity-constrained triage policies that compare risk-based, utility-based, and uncertainty-aware referral strategies. This provides an operational view of how predictive models may be used when only a subset of patients can receive more intensive assessment.
4. We embed conformal prediction within the decision pathway to produce set-valued predictions with finite-sample coverage guarantees and explicit abstention. This enables structured deferral of ambiguous cases and supports more transparent uncertainty handling across subgroups such as age and BMI, contributing to safer AI-assisted decision support [33].

The following section describes the data, modelling approaches, evaluation metrics, and decision-analytic methods used to address these questions.

2. Methods

2.1. Data Description

The dataset used in this study is a publicly available Polyendocrine metabolic ovarian syndrome (PCOS) dataset comprising records collected from 10 hospitals across Kerala, India. It contains 541 patient records and 44 variables, including a binary target variable indicating the presence of PCOS (1), with 177 participants (32.7%), or the absence of PCOS (0), with 364 participants (67.3%) as shown in Fig 1. The variables capture demographic information, anthropometric measurements, lifestyle habits, menstrual history, biochemical assay results, and ultrasound findings. These features provide a broad clinical representation of factors associated with PCOS and support predictive modelling of the condition.

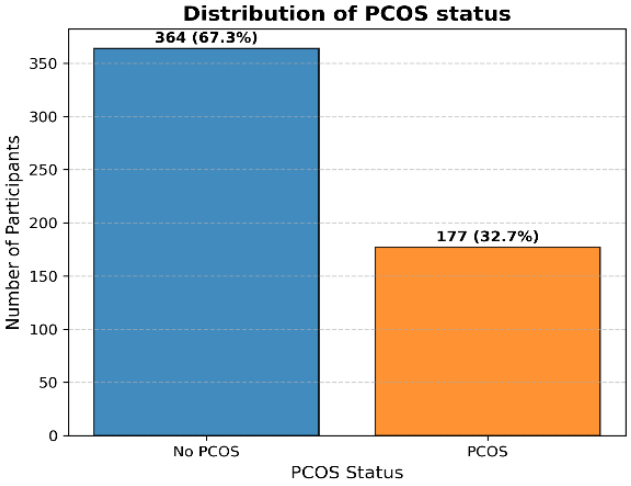

Figure 1: Distribution of PCOS Status

Table 1: Variables with their data types and brief descriptions.

S/N	Feature Name	Data Type	Description
1	Age (years)	Numeric	Age of the patient in years
2	Weight (Kg)	Numeric	Body weight measured in kilograms
3	Height (Cm)	Numeric	Height of the patient in centimetres
4	BMI	Numeric	Body mass index, calculated from weight and height
5	Blood Group	Categorical	Patient's blood group, encoded numerically
6	Pulse rate(bpm)	Numeric	Heart rate measured in beats per minute
7	RR (breaths/min)	Numeric	Respiratory rate measured in breaths per minute
8	Hb(g/dl)	Numeric	Haemoglobin concentration in grams per decilitre
9	Cycle(R/I)	Categorical	Menstrual cycle regularity or irregularity indicator
10	Cycle length(days)	Numeric	Length of menstrual cycle in days
11	Marriage Status (Years)	Numeric	Number of years since marriage
12	Pregnant(Y/N)	Binary	Pregnancy status
13	No. of abortions	Numeric	Number of abortions or miscarriages
14	I beta-HCG (mIU/mL)	Numeric	Serum beta-HCG hormone level (first measurement)
15	II beta-HCG (mIU/mL)	Numeric	Serum beta-HCG hormone level (second measurement)
16	FSH (mIU/mL)	Numeric	Follicle-stimulating hormone level
17	LH (mIU/mL)	Numeric	Luteinising hormone level
18	FSH/LH	Numeric	Ratio of FSH to LH
19	Hip(inch)	Numeric	Hip circumference in inches
20	Waist(inch)	Numeric	Waist circumference in inches
21	Waist: Hip Ratio	Numeric	Waist-to-hip circumference ratio
22	TSH (mIU/L)	Numeric	Thyroid-stimulating hormone level

S/N	Feature Name	Data Type	Description
23	AMH (ng/mL)	Numeric	Anti-Müllerian hormone level
24	PRL (ng/mL)	Numeric	Prolactin hormone level
25	Vit D3 (ng/mL)	Numeric	Vitamin D3 level
26	PRG (ng/mL)	Numeric	Progesterone hormone level
27	RBS (mg/dl)	Numeric	Random blood sugar level
28	Weight gain(Y/N)	Binary	Indicator for recent weight gain
29	hair growth(Y/N)	Binary	Indicator for excessive hair growth
30	Skin darkening (Y/N)	Binary	Indicator for skin hyperpigmentation
31	Hair loss(Y/N)	Binary	Indicator for hair loss
32	Pimples(Y/N)	Binary	Indicator for presence of acne
33	Fast food (Y/N)	Binary	Indicator for regular fast-food consumption
34	Reg.Exercise (Y/N)	Binary	Indicator for regular exercise
35	BP Systolic (mmHg)	Numeric	Systolic blood pressure measurement
36	BP Diastolic (mmHg)	Numeric	Diastolic blood pressure measurement
37	Follicle No. (L)	Numeric	Number of ovarian follicles on the left ovary
38	Follicle No. (R)	Numeric	Number of ovarian follicles on the right ovary
39	Avg. F size (L) (mm)	Numeric	Average follicle size in the left ovary
40	Avg. F size (R) (mm)	Numeric	Average follicle size in the right ovary
41	Endometrium (mm)	Numeric	Endometrial thickness in millimetres
42	PCOS (Y/N)	Binary	Presence or absence of PCOS

2.2. Data Pre-processing

All column names were stripped of extra spaces to ensure consistency. A missing value check identified two variables with incomplete records:

- Marriage Status (Years) - 1 missing value, handled using median imputation.
- Fast food (Y/N) - 1 missing value, imputed with the mode.

We selected these imputation methods due to the minimal proportion of missing data and their ability to preserve the central tendency of the variables with minimal distortion [34].

Duplicate record checks across all variables returned zero duplicates. Binary categorical variables such as Pregnant(Y/N), Weight gain(Y/N) were identified by the (Y/N) suffix in their names and recorded as integers (0 = No, 1 = Yes) to facilitate direct use in analysis.

After these steps, the dataset contained 541 complete observations and 42 variables, with no missing values or duplicate records.

2.3. Feature Partitioning

For modelling purposes, we separated all predictor variables into two clinically motivated categories based on method of acquisition, availability and relative cost. The target variable, PCOS (Y/N), is a binary indicator of diagnosis. This partitioning was designed to compare an inexpensive-feature model, built from variables that are routinely obtainable without laboratory or imaging procedures, with an augmented model using inexpensive plus expensive features, in which laboratory assays and ultrasound findings are added when fuller diagnostic work-up is available.

- Initial evaluation using inexpensive and readily obtainable information such as patient history, anthropometric measurements, lifestyle indicators, and routine vital signs;
- further assessment using more resource-intensive investigations such as hormonal assays and pelvic ultrasound when additional diagnostic clarification is needed.

Figure 2 presents an overview of the proposed cost-aware PCOS risk stratification framework. Inexpensive features are used for first-line risk estimation, while uncertainty and clinical decision rules guide escalation to more resource-intensive assessment. The workflow emphasises efficient resource use, transparency, and clinically informed decision support, with algorithmic complexity kept secondary to practical deployment considerations.

The dataset was partitioned enabling comparison between a lower-burden model based on inexpensive information and an augmented model that incorporates laboratory and ultrasound features. Table 2 summarises the principles used to define these feature groups, and Table 3 presents the resulting partitioning.

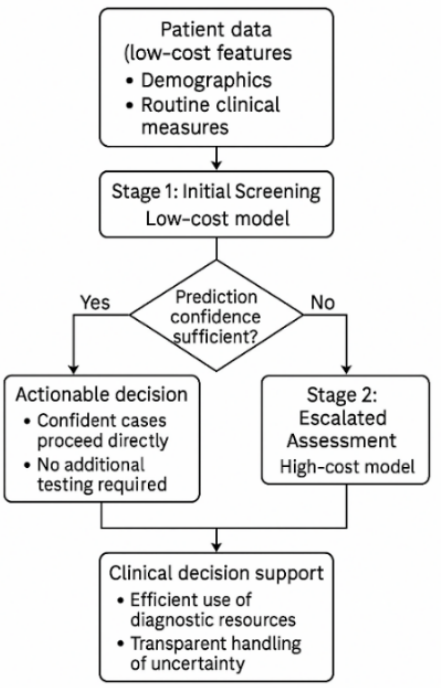

Figure 2: Proposed workflow of the PCOS modelling

Table 2: Principles of Feature Partitioning

Model/Feature Group	Inclusion Criteria	Examples
Inexpensive-feature model	Self-reported, physical examination, or routine vital signs; no laboratory or imaging required; minimal acquisition cost.	Age, BMI, Blood Pressure, Lifestyle indicators
Expensive additions in the augmented model	Laboratory analysis (e.g., hormonal assays) or imaging procedures (e.g., pelvic ultrasound); require specialist facilities and higher financial cost.	FSH, LH, AMH, Ultrasound follicle counts

Table 3. Feature Categories after partitioning with Justification

Feature Group	Category	Feature Name	Partition	Justification
Inexpensive features	Demographics	Blood Group	inexpensive	Often known or on medical record
		Marriage Status (Yrs), Pregnant (Y/N), Age (yrs)	inexpensive	Self-reported
		No. of aborptions	inexpensive	Medical history
	Anthropometrics	Weight (Kg), Height (Cm)	inexpensive	Routine measure
		BMI, Waist: Hip Ratio	inexpensive	Derived
		Hip (inch), Waist (inch)	inexpensive	Measured with tape
	Vitals	Pulse rate (bpm), RR (breaths/min), BP Systolic (mmHg), BP Diastolic (mmHg)	inexpensive	Measured at clinic
	Menstrual History	Cycle (R/I), Cycle length (days)	inexpensive	Patient-reported
	Lifestyle / Symptoms	Weight gain (Y/N), hair growth (Y/N), Skin darkening (Y/N), Hair loss (Y/N), Pimples (Y/N), Fast food (Y/N), Reg.Exercise (Y/N)	Inexpensive	Patient-reported
Augmented features	Hormonal Assays	Hb(g/dl)	Expensive	Requires lab blood test
		I beta-HCG (mIU/mL), II beta-HCG (mIU/mL), FSH (mIU/mL), LH (mIU/mL), TSH (mIU/L), AMH (ng/mL), PRL (ng/mL), Vit D3 (ng/mL), PRG (ng/mL), RBS (mg/dl)	Expensive	Requires lab assay
		FSH/LH	Expensive	Derived from lab results
	Ultrasound Finding	Follicle No. (L), Follicle No. (R), Avg. F size (L) (mm), Avg. F size (R) (mm), Endometrium (mm)	Expensive	Requires ultrasound

Table 3 details the variables assigned to the inexpensive feature group and the additional expensive variables incorporated into the augmented model. The inexpensive feature group comprised 24 features while the expensive additions in the augmented features were 17 features. This split reflects real-world diagnostic economics; inexpensive feature groups are low-cost or free to collect, while the augmented features require specialized lab or imaging services. This partition supports three related objectives in the present study:

1. Evaluation of an inexpensive-feature model for first-line PCOS risk stratification under limited resource conditions.
2. Comparison with an augmented model that adds laboratory and ultrasound variables when fuller diagnostic assessment is available.
3. Support for downstream triage and uncertainty-aware escalation, where early predictions from inexpensive information may guide referral for more intensive follow-up.

2.4. Baseline Modelling Framework

2.4.1. Modelling Configuration

To evaluate the trade-off between predictive performance and diagnostic burden, we developed two baseline modelling configurations. The first was an inexpensive-feature model, trained using demographic, symptom, anthropometric, lifestyle, menstrual, and routine clinical variables that can be obtained without specialised testing. The second was an augmented model using inexpensive plus expensive features, in which laboratory and ultrasound variables were added to the inexpensive feature set to represent fuller diagnostic work-up when these investigations are available.

The inexpensive-feature model was intended to reflect a first-line risk stratification setting in which rapid, accessible, and lower-burden information is used to estimate PCOS risk. The augmented model was included to quantify the incremental value gained when more resource-intensive diagnostics are added. This comparison supports the question of the study, namely how much predictive signal can be captured before escalation becomes necessary, and how much additional value is obtained when fuller testing is feasible.

2.4.2. Baseline Algorithms

To provide transparent and interpretable benchmark models for the inexpensive-feature and augmented configurations, we selected Logistic Regression (LR) and Random Forest (RF) as the baseline classifiers.

1. Logistic Regression (LR), serving as an interpretable linear model, with standardisation applied to all numeric predictors. Its transparency allows direct inspection of coefficient magnitudes and directions, supporting clinical interpretability [31].
2. Random Forest (RF), an ensemble of decision trees capable of modelling non-linear relationships and higher-order feature interactions, with intrinsic robustness to multicollinearity and noise [35].

LR and RF were evaluated across both the inexpensive-feature model and the augmented model to compare discrimination, calibration, and downstream decision utility under different levels of diagnostic information availability instead of comparing a broad set of algorithms. This focus allowed us to investigate the value of the framework itself while keeping computation lightweight and practical for potential clinical use. The training approaches for this work in summarized in Table 4.

2.4.3. Evaluation Metrics

We assessed model performance using a combination of statistical and clinical evaluation measures. Discrimination ability was quantified via Area Under the Receiver Operating Characteristic Curve (AUC), Accuracy, F1-score, Precision and Recall [36]. Calibration quality was measured using Brier score [37] and Expected Calibration Error (ECE) [38], ensuring predicted probabilities aligned with observed event rates. In addition, decision curve analysis in [28] was retained in the evaluation plan to estimate clinical net benefit across a range of threshold probabilities, enabling direct translation of model outputs into decision-focused frameworks explored in Section 2.5.

2.4.4. Integrated Feature Importance and Selection

To understand how predictive value was distributed across feature groups, we conducted integrated feature importance and sensitivity analysis across the two baseline model configurations. For the inexpensive-feature model, we examined permutation importance [39] and grouped feature reduction to determine how much predictive performance could be preserved using a smaller and lower-burden set of variables.

For the augmented model, we performed targeted ablation of the added expensive variables to assess which laboratory and ultrasound features contributed incremental value beyond the inexpensive baseline.

The choice of distinct feature-selection approaches reflects the different aims of the two grouped feature models and how they work in a cost-sensitive workflow. Inexpensive-feature model aims to identify the minimal set of low-cost features that support early triage, making permutation-based ranking and grouped removal appropriate for evaluating redundancy. Addition of

expensive features in the augmented model requires determining which costly tests meaningfully enhance discrimination beyond inexpensive-feature, for which single-feature ablation provides the clearest estimate of marginal gain. The two analyses therefore offer complementary insights; feature importance within the inexpensive-feature model isolates informative low-cost variables for initial PCOS risk assessment, while ablation of expensive additions in the augmented model identifies high-cost features that justify escalation.

2.4.5. *Preventing Data Leakage*

All preprocessing, feature selection and model fitting procedures were performed strictly within the training folds during cross-validation, with transformations applied to validation and test folds only after fitting. This ensured no information from the validation or test data influenced model training, thereby preserving the validity of out-of-sample performance estimates [40]. The outputs from the baseline modelling phase serve three primary functions:

1. Identification of high-yield inexpensive features suitable for lower-burden first-line screening.
2. Quantification of the marginal value of expensive diagnostic features within the augmented model.
3. Provision of benchmark performance metrics for subsequent decision curve analysis, triage evaluation, and conformal prediction.

By quantifying feature contributions, calibration behaviour, and performance trade-offs under different levels of diagnostic burden, the baseline framework provides the empirical foundation for the decision-focused analyses that follow.

2.4.6. *Baseline Model Experimental Settings*

To ensure reproducibility, we conducted all experiments with fixed parameters as summarized in Table 5. These parameters detail the data partitioning strategy, cross-validation configuration, calibration settings and algorithm-specific parameters.

2.5. *Decision-Focused Learning with Conformal Prediction*

In this phase of the study, we extended the baseline predictive framework into a decision-focused setting by integrating clinical utility analysis with class-conditional split conformal prediction [41], [42]. Whereas the inexpensive-feature and augmented models quantify performance trade-offs under different levels of diagnostic information, the conformal component serves as the uncertainty-control layer of the framework, allowing ambiguous cases to be identified and deferred when model confidence is insufficient for a definitive binary decision. This enables evaluation of predictive accuracy and how model outputs may be translated into safer and more operationally realistic triage under constrained capacity.

2.5.1. *Clinical Utility Assessment and Capacity-Constrained Referral Optimisation*

In this study, we evaluated the clinical utility of the PCOS risk models and their operational use under constrained diagnostic capacity using a decision-analytic framework. Figure 3 illustrates the flow from baseline risk estimation using the inexpensive-feature and augmented models to decision curve analysis, net benefit evaluation across thresholds, and comparison of risk-based, utility-based, and uncertainty-aware referral strategies under fixed referral-capacity constraints.

Table 4: Model training framework

Approach	**Training Strategy**	**Cross validation used**	**Evaluation set**	**Predictions used for metrics**
Train-test split evaluation	Model trained on a single training subset	No	Fixed test set (30%)	Test set predictions
Cross validated evaluation (OOF)	Model trained across folds	Yes (5 folds)	Out of fold samples	OOF predictions

Table 5: Baseline Model Experimental Settings

Parameter	**Value**
Train/test split ratio	70% training / 30% testing (stratified by target)
Cross-validation folds (OOF)	5 folds (StratifiedKFold)
Random seed	42 (fixed across all experiments)
ECE bin count	10 bins
Random forest trees (baseline models)	500 trees
Random forest trees (Permutation importance)	200 trees
Scaling method	StandardScaler (applied only to LR numeric features)
Evaluation metric for model selection	AUC (ROC area)

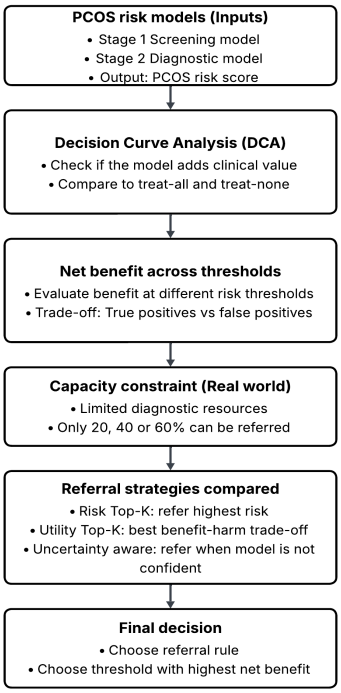

Figure 3: Clinical utility and capacity-constrained decision framework for our PCOS risk models.

Predicted PCOS risk scores from both the inexpensive-feature model and the augmented model were used as inputs to this analysis. We first assessed model usefulness using Decision Curve Analysis (DCA) [28], which quantifies clinical value in terms of net benefit across a range of decision thresholds. For each threshold probability, model-based referral decisions were compared with two non-informative reference strategies, namely treating all individuals as high risk and treating none. Net benefit was computed as:

$$NB = \frac{TP}{N} - \frac{FP}{N} \cdot \frac{p_t}{1 - p_t}$$

where TP = True Positive, FP = False Positive, N = patient population, p_t= Decision thresholds, the relative trade-off between false positive and false negative decisions.

We then extended the analysis to reflect real-world limits on diagnostic capacity. Specifically, we considered referral capacities of 20%, 40%, and 60%, representing scenarios in which only a subset of women can be referred for further assessment. Within each capacity setting and at each decision threshold, we evaluated three referral strategies.

First, we considered a Risk Top-K policy, in which referrals were allocated to individuals with the highest predicted PCOS risk probabilities. This reflects a straightforward risk-ranking approach based solely on estimated disease likelihood.

Second, we evaluated a Utility Top-K policy, in which individuals were ranked according to an expected utility score that incorporates the threshold-specific trade-off between false positive and false negative decisions.

$$si = pi - w(1 - pi), \quad w = pt/(1 - pt)$$

Third, we evaluated an Uncertainty-Aware policy, in which individuals for whom the model could not make a confident classification were prioritised for referral, and any remaining referral capacity was then allocated using the same utility-based ranking. In the present framework, this strategy is intended to ensure that diagnostically ambiguous cases are not automatically deprioritised under fixed capacity constraints, while still preserving alignment with threshold-specific decision utility.

For each referral capacity, we computed net benefit across the full threshold grid for all three strategies and identified the threshold that maximised net benefit. This analysis defines the optimal capacity-constrained referral rules for PCOS risk stratification and provides the decision-analytic foundation for the uncertainty-aware triage framework used in this study. These policies allow comparison of how the inexpensive-feature model and the augmented model may support referral decisions under scarcity, while clarifying the operational role of conformal abstention within the broader decision pathway.

2.5.2. *Data Partitioning for Conformal Prediction*

We partitioned the dataset into training, calibration and test sets using a stratified split to preserve class balance. In line with established conformal methodology [41], [43], we allocated 60% of the data for training, 20% for calibration and 20% for testing. We used the calibration set exclusively for determining nonconformity thresholds, thereby ensuring an unbiased evaluation on the held-out test set.

2.5.3. *Model Training and Nonconformity Scoring*

We trained a classifier on the training set presented [35]. Using the calibration set, we obtained class-conditional probability estimates and computed nonconformity scores: for positive cases, (1 - p_1); for negative cases, (p_1), where p_1 denotes the predicted probability of the positive label. We then calculated class-conditional (Mondrian) quantiles using the finite-sample formula:

$$q\alpha = \text{Quantile}[(n+1)\ (1-\alpha)]/n \text{ with } \alpha = 0.05,$$

yielding separate thresholds t_1 and t_0 for the positive and negative classes.

2.5.4. *Prediction Set Construction and Subgroup Analysis*

For each test instance, we constructed the conformal prediction set according to inclusion rules: we included class 1 if $p_1 >= 1 - t_1$ and class 0 if $p_1 <= t_0$. When both conditions were satisfied, we labelled the case as abstained, triggering a deferred decision. We then evaluated coverage and abstention rates across clinically relevant subgroups, including age quartiles and BMI

categories defined by World Health Organisation (WHO) guidelines.

2.5.5. *Coverage and Confidence Interval Estimation*

We computed coverage, abstention rates, singleton set rates and polarity-specific singleton proportions both overall and within subgroups. To quantify uncertainty in these estimates, we applied the Clopper-Pearson exact method which provided reliable confidence intervals even for smaller subgroup sizes [44].

2.6. *Comparative Evaluation and Statistical Analysis*

2.6.1. *Evaluation Strategy*

We designed a unified evaluation framework to compare baseline classifiers, decision-focused learning models and their conformal prediction-augmented variants. Our analysis jointly assessed predictive performance, calibration reliability, clinical decision utility and uncertainty quantification, with all statistical inference grounded in out-of-fold (OOF) predictions. These OOF predictions served as the common input for all subsequent statistical tests, calibration analysis, decision curve estimations and bootstrap confidence interval computations.

2.6.2. *Performance, Calibration and Decision Utility Analysis*

We evaluated discrimination using the area under the ROC curve (AUC), with statistical differences between paired AUCs tested via the DeLong method for correlated ROC curves [45]. Sensitivity and specificity were compared using the McNemar test for paired binary classifications, applied at a fixed decision threshold of 0.5. Calibration was assessed through both the Brier score and the expected calibration error (ECE), the latter computed with 10 equal-width probability bins. ECE estimates were accompanied by bootstrap-derived 95 % confidence intervals using 500 stratified resamples to preserve class proportions. Paired differences in Brier scores between models were quantified using bootstrap mean difference estimation (1,000resamples).

Table 6: Experimental Settings for decision curve analysis & conformal prediction

	Parameter	**Implementation Details**	**Notes**
Decision Curve Analysis	Referral capacities	0.20, 0.40, 0.60	Fraction of test cohort
	Decision thresholds (pt)	0.05 to 0.50	Step size = 0.01
	Triage policies	uncertainty aware, Risk Top-K, Utility Top-K	See Section 2.5.1
	Net benefit computation	Vickers & Elkin (2006)	Normalised by test set size
Conformal Prediction	Feature set	Inexpensive-feature model	Demographic, lifestyle and basic clinical measures
	Train / Calibration / Test split	60% / 20% / 20%	Stratified by PCOS label
	Confidence level (α)	0.05	Target coverage = 95%
	Conformal type	Mondrian split conformal (Class-conditional thresholds)	Separate nonconformity distributions for positive and negative calibration samples
	Coverage CI method	Clopper-Pearson (exact)	Applied overall and within subgroups
	Subgroup definitions	Age quartiles, BMI categories	BMI cut-points per WHO guidelines
	Class conditional nonconformity scores	For PCOS: ($A_1 = 1 - p$). For non PCOS: ($A_0 = p$)	Lower values indicate better conformity with a class
	Rules for assigning labels to new patients (Prediction set construction)	For predicted probability (p): include PCOS if ($p \geq 1 - t_1$). Include non PCOS if ($p \leq t_0$)	Can produce {1}, {0}, or {0,1}, depending on uncertainty
	Prediction Set (Possible output)	{1} confident PCOS. {0} confident non PCOS. {0,1} ambiguous	Prediction set size reflects certainty
	Identification of uncertain cases	Abstention occurs when the prediction set is {0,1}	Empty sets do not occur under this method

For uncertainty quantification, we examined empirical coverage rates from the conformal prediction sets both overall and within clinically relevant strata (PCOS-positive vs. negative, age quartiles and BMI categories). Deviations from nominal coverage were tested using exact binomial confidence intervals (Clopper-Pearson) and abstention rates were compared between subgroups via the Chi-square test. To assess clinical utility, we applied decision curve analysis [28] and simulated diagnostic capacity constraints allowing advanced testing for 20 %, 40 % and 60 % of patients. For each scenario, we computed net benefit (NB) and the number of unnecessary advanced tests avoided. Pairwise NB differences were evaluated via bootstrap confidence intervals using 2,000 stratified patient-level resamples.

2.6.3. Statistical Inference and Multiple Comparisons

All hypothesis tests were two-sided, and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) control procedure. Statistical significance was defined as an adjusted $p < 0.05$. Effect sizes and CIs were prioritised over raw p-values in interpretation to align with modern statistical reporting recommendations [46].

Bootstrapping, exact tests and non-parametric alternatives (Wilcoxon signed-rank) were applied when distributional assumptions could not be verified via the Shapiro-Wilk test. All analysis were implemented in Python using NumPy, pandas and custom-written functions for DeLong, McNemar and bootstrap procedures, ensuring complete reproducibility without reliance on opaque third-party wrappers.

Table 7: Experimental Settings for Statistical analysis

Setting	Value
ECE computation	10 probability bins, bootstrap CIs from 500 resamples
Brier score differences	Paired bootstrap with 1,000 resamples
Net benefit comparisons	Bootstrap with 2,000 stratified patient-level resamples
Conformal coverage evaluation	Nominal confidence level per method (reported in results)

3. Results

3.1. Baseline Modelling Performance

In this study, we evaluated both the train-test performance (baseline) and the out-of-fold (OOF) cross-validation estimates for the inexpensive-feature model and the augmented model (inexpensive + expensive features), reporting discrimination (AUC, Accuracy, F1-score, Precision, Recall) and pre-calibration metrics (Brier score and ECE) summarized in Table 8, following TRIPOD guidelines for transparent reporting of prediction models [47]. This comparison was designed to quantify how much predictive value can already be obtained from inexpensive variables alone, and how much additional gain is achieved when laboratory and ultrasound features are added.

For the inexpensive-feature model, the train-test split results showed broadly comparable performance for Logistic Regression (LR) and Random Forest (RF): LR achieved 84.94% AUC, 83.44% Accuracy, 73.27% F1, 77.08% Precision and 69.81% Recall with Brier = 0.1335 and ECE = 0.0669, while RF reached 85.75% AUC, 80.98% Accuracy, 67.37% F1, 76.19% Precision and 60.38% Recall with Brier = 0.1357 and ECE = 0.0884. The corresponding OOF estimates were slightly higher for discrimination and accuracy and exhibited better pre-calibration, with LR at 87.60% AUC, 83.36% Accuracy, 73.84% F1, 76.05% Precision, 71.75% Recall, Brier = 0.1229, ECE = 0.0494 and RF at 88.57% AUC, 80.96% Accuracy, 68.31% F1, 75.00% Precision, 62.71% Recall, Brier = 0.1258, ECE = 0.0629 as presented in Table 8. LR provided a more balanced sensitivity-recall profile and more favourable raw calibration, whereas RF showed slightly stronger discrimination but tended to prioritise precision at the expense of recall. These findings indicate that the inexpensive-feature model already provides meaningful first-line PCOS risk stratification using lower-burden clinical information.

When moving to the augmented model, where laboratory and ultrasound features were added to the inexpensive baseline, performance improved for both algorithms in both evaluation settings. On the train-test approach, LR obtained 93.53% AUC, 85.89% Accuracy, 78.50% F1, 77.78% Precision, 79.25% Recall, with Brier = 0.0939 and ECE = 0.0513, while RF achieved 94.97% AUC, 89.57% Accuracy, 82.11% F1, 92.86% Precision, 73.58% Recall, Brier = 0.0902, ECE = 0.1113. The OOF estimates showed the same qualitative pattern: LR reached 94.47% AUC, 89.46% Accuracy, 83.76% F1, 84.48% Precision, 83.05% Recall, with Brier = 0.0782 and ECE = 0.0334, while RF achieved 96.00% AUC, 90.39% Accuracy, 84.34% F1, 90.32% Precision, 79.10% Recall, with Brier = 0.0845 and ECE = 0.1013. As in the inexpensive-feature setting, LR exhibited more favourable raw probability calibration, whereas RF delivered the strongest discrimination and precision once fuller diagnostic information was available.

These results illustrate the expected pattern in a cost-aware diagnostic framework, where the inexpensive-feature model provides moderately high discrimination with minimal acquisition burden, while the augmented model benefits from specialised tests to deliver stronger classification performance and improved discrimination.

Table 8: Train-test split performance and out-of-fold performance metrics (AUC, F1, Precision, Recall, Brier, ECE) for inexpensive model and augmented model

Evaluation Type	Feature Group	Algorithm	AUC (%)	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	Brier Score	ECE
Train-test split Prediction (Baseline)	Inexpensive	LR	84.94	83.44	73.27	77.08	69.81	0.1335	0.0669
		RF	85.75	80.98	67.37	76.19	60.38	0.1357	0.0884
	Augmented	LR	93.53	85.89	78.50	77.78	79.25	0.0939	0.0513
		RF	94.97	89.57	82.11	92.86	73.58	0.0902	0.1113
Out of fold Prediction (Cross validation)	Inexpensive	LR	87.60	83.36	73.84	76.05	71.75	0.1229	0.0494
		RF	88.57	80.96	68.31	75.00	62.71	0.1258	0.0629
	Augmented	LR	94.47	89.46	83.76	84.48	83.05	0.0782	0.0334
		RF	96.00	90.39	84.34	90.32	79.10	0.0845	0.1013

The baseline results are also consistent with the OOF estimates: adding laboratory and ultrasound variables materially improves predictive performance, while the inexpensive-feature model still retains substantial screening value. Between algorithms, RF showed the strongest overall discrimination and precision, particularly in the augmented setting, whereas LR yielded more balanced recall and better raw probability calibration. On this basis, we carried forward Random Forest for the subsequent feature sensitivity, ablation, and downstream decision-focused experiments, while retaining the LR comparison here to characterise the baseline performance-calibration trade-off.

Next, we plot the ROC curves used to compare the inexpensive-feature model and the augmented model using the OOF estimates in Fig 4. ROC curves evaluate the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across varying decision thresholds [48]. In the inexpensive-feature model, the Random Forest (RF) achieved an AUC of 88.60%, slightly outperforming Logistic Regression (LR) with an AUC of 87.6%. This indicates both models have strong discriminative power, but RF captures marginally more non-linear interactions in the limited feature set.

In the augmented feature model, incorporating expensive diagnostic features alongside the inexpensive set, both models improved markedly. RF rose to an AUC of 96.00%, while LR increased to 94.47%. The improvement is consistent with prior findings that richer feature sets enhance classification performance, particularly for ensemble methods [49]. The steeper rise of the ROC curve near the y-axis in the augmented feature model reflects a reduction in false positives at high sensitivity levels. While LR's AUC gains from inexpensive feature model to the augmented feature model were substantial, RF retained a consistent advantage at both configurations. This aligns with the known robustness of tree-based ensembles to complex feature interactions and heterogeneous data scales [50], [51] and supports the use of RF as the main baseline model for the subsequent feature reduction, ablation, and decision-focused analyses.

We also assessed how well predicted probabilities matched observed outcomes using reliability diagrams built from OOF predictions shown in Fig 5. Probabilities were partitioned into 10 equal-width bins and each point was annotated with its bin count to make sampling variability explicit. In the inexpensive feature model, both Random Forest and Logistic Regression track the diagonal closely across the bulk of the probability range. Small departures appear around the mid-probability bins (≈0.3-0.6), but the curves remain near the ideal line. This pattern indicates usable probabilistic outputs already at the inexpensive-feature configuration; any minor mid-range bias could be addressed later with lightweight post-hoc calibration if deployment requires tighter probability estimates.

In the augmented feature model, discrimination improved and calibration remained strong overall, with a slight over-confidence tendency for Random Forest in the upper-probability bins (>0.7) and a mild under-confidence for Logistic Regression in the mid-high range (≈0.5-0.7) before converging near the diagonal at the top end. These are typical model-specific patterns documented in the calibration literature: tree ensembles can become over-confident and monotonic transformations (e.g., Platt scaling) or non-parametric isotonic regression are standard remedies when strict probability calibration is required [52]. Considered along with the baseline metrics

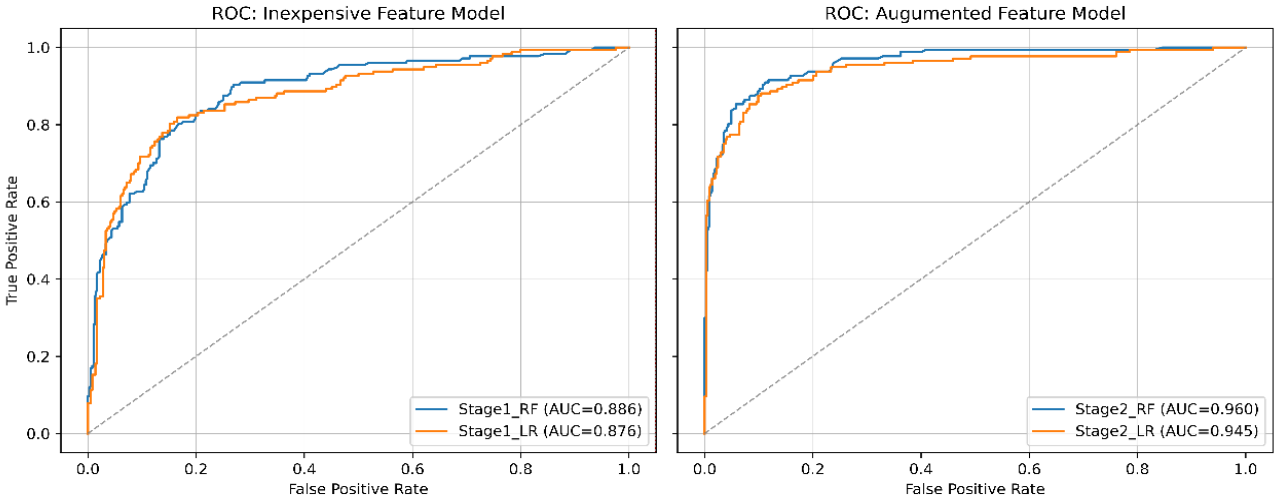

Figure 4: ROC curves for Inexpensive feature model and augmented models. Curves plotted from OOF predictions

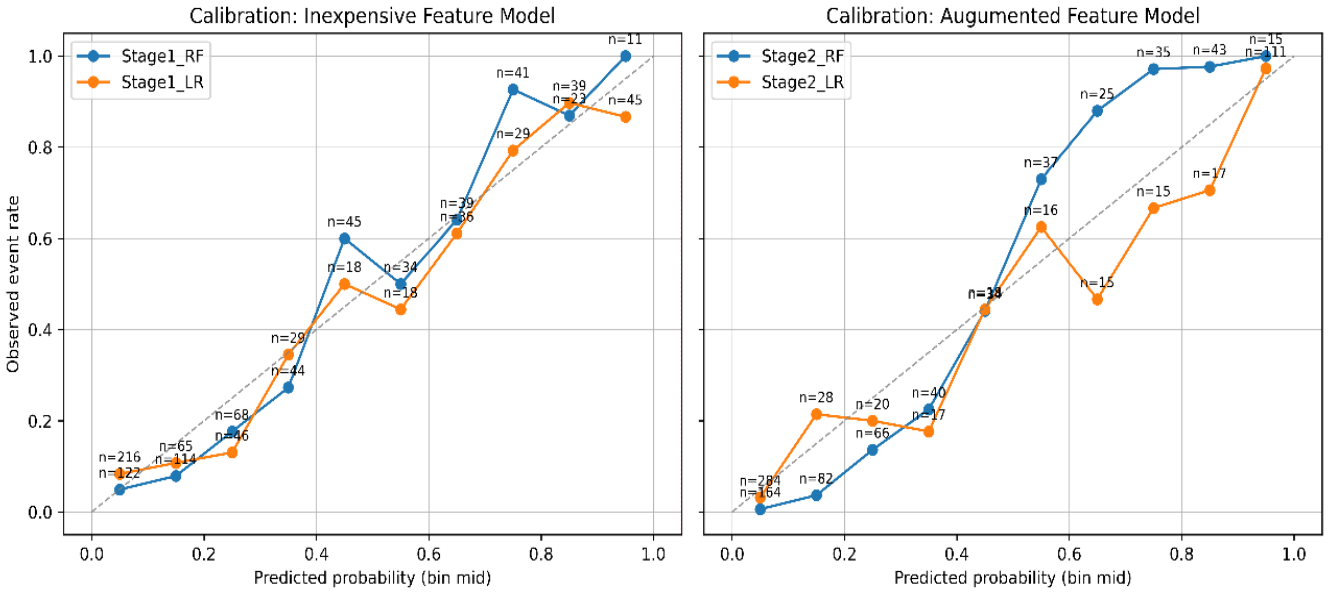

Figure 5: Reliability Diagram for inexpensive feature model and augmented models. Observed event rate vs. bin-mid predicted probability; 10 equal-width bins; per-bin n annotated.

in Table 8, these calibration patterns show that the augmented model improves discrimination, while the inexpensive-feature model still offers usable probabilistic outputs for early risk stratification and referral support when fuller diagnostic testing is not immediately available.

3.2. Feature Importance and Selection

To further examine the cost-performance trade-off identified in Section 3.1, we carried out feature importance and sensitivity analyses using Random Forest, which showed the strongest overall discrimination and precision in the baseline comparison and was therefore carried forward for the remaining experiments. For the inexpensive-feature model, we applied permutation importance and grouped feature removal to identify a lower-burden subset that could preserve performance. For the augmented model, we conducted one-at-a-time ablation of the added expensive features to assess their marginal contribution beyond the inexpensive baseline.

Over the full set of 24 "inexpensive" predictors ranked Hair growth (Y/N), Skin darkening (Y/N) and Fast food (Y/N) as the most influential features, with mean importance scores of 0.044, 0.031 and 0.020 respectively. In contrast, variables such as Hair loss (Y/N), No. of abortions, Blood group, BP Systolic (mmHg) and Height (cm) were ranked lowest, in some cases contributing negligible or slightly negative mean importance as shown in Fig 6. This ranking highlights a clear separation between high and low-impact inexpensive predictors, providing an interpretable basis for targeted feature reduction when time, cost, or data availability constrain first-line screening.

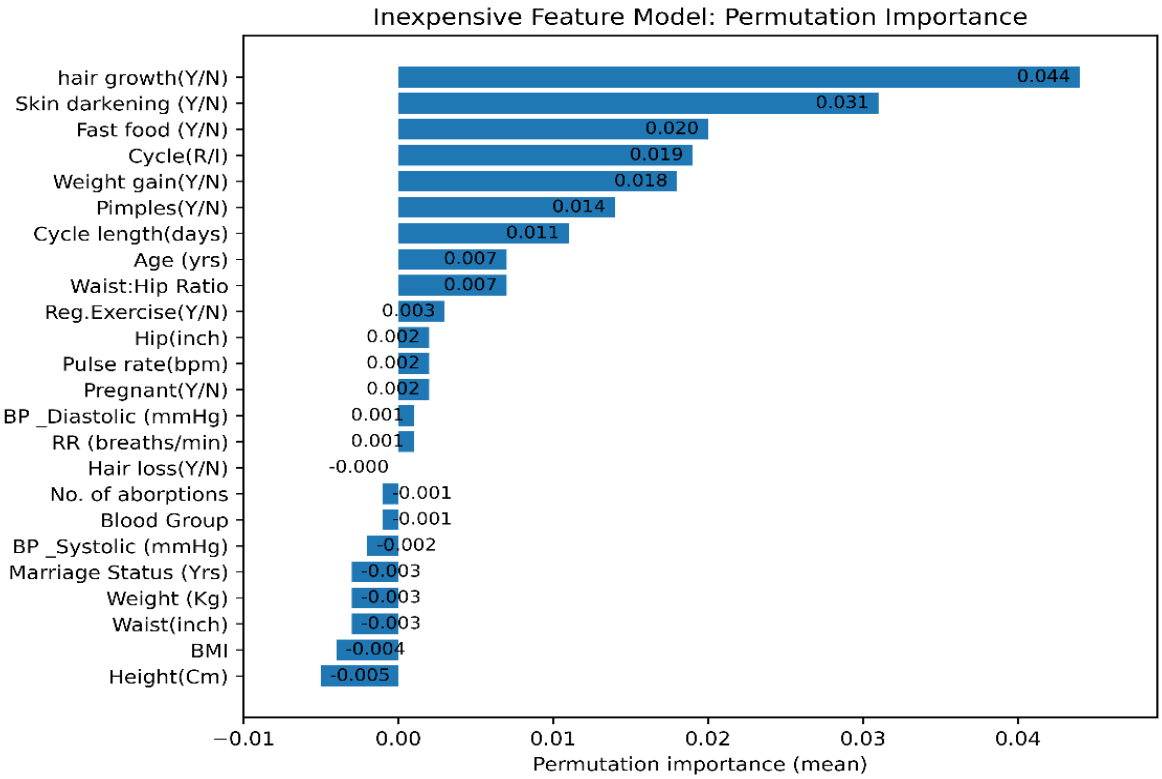

Figure 6: Ranked list of features by mean permutation importance (using inexpensive feature model – RF Baseline)

In addition, we conducted a feature drop sensitivity analysis (Fig 7) for the inexpensive-feature RF model under baseline evaluation. Removing the bottom 3 to 8 least important features, as summarised in Table 9, resulted in no deterioration in discrimination and, in fact, moderate AUC gains: from a baseline of 85.8% to 87.1% (k = 3) and 87.4% (k = 5), while removing the bottom 8 features yielded the highest observed AUC (88.6%) as shown in Fig 8. The associated metric changes revealed small improvements in recall (+7.55% at k = 8) but minor decreases in precision (-4.2% at k = 8) as seen in Fig 7. Brier score and Expected Calibration Error (ECE) changes were negligible across all k values. These results suggest that several of the lowest-ranked inexpensive predictors may contribute noise or redundancy, and that a reduced inexpensive-feature model can maintain, and in some cases slightly improve, predictive performance without sacrificing raw calibration quality.

Table 9: Least importance features guided by the permutation importance

K- removed	**Feature dropped**
3	Height (cm), BMI, Waist (inch)
5	Height (cm), BMI, Waist (inch), Weight (kg), Marriage Status (yrs)
8	Height (cm), BMI, Waist (inch), Weight (kg), Marriage Status (yrs), BP Systolic (mmHg), No. of abortions, Blood Group

The AUC slope plot (Fig 8) illustrated a consistent upward trend in performance as low-value predictors were dropped, with diminishing returns beyond the top ~16 features. This supports a modelling approach in which a compact inexpensive-feature subset can be deployed at lower operational cost while maintaining, or even slightly enhancing, predictive performance for early PCOS risk stratification.

To complement the permutation importance and feature-drop sensitivity analysis results from the inexpensive-feature model, we performed an ablation analysis on the augmented Random Forest model, where each added expensive feature was removed individually to measure its marginal contribution to performance. The change in AUC (ΔAUC) from the full model was used as the sensitivity metric. The results (Table 10) show that Follicle Number (R) had the largest performance drop when removed (ΔAUC = -2.49%), followed by Follicle Number (L) (-1.04%) and PRG (ng/mL) (-0.79%). Several features, including PRL (ng/mL) and Endometrium (mm), had minimal impact (ΔAUC ≤ -0.35%), suggesting that they contribute comparatively little additional value once stronger predictors are present.

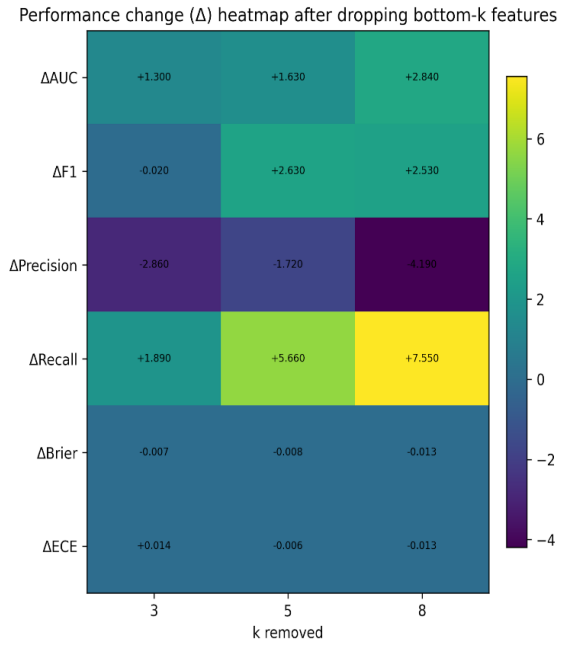

Figure 7: Heatmap of performance metric changes (ΔAUC, ΔF1, ΔPrecision, ΔRecall, ΔBrier, ΔECE) after dropping the k least important features

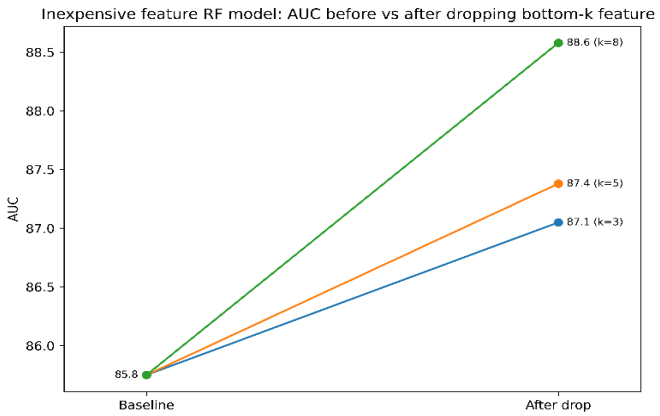

Figure 8: plot of baseline vs. post-drop AUC for k = 3, 5, 8 feature removals

The analysis highlights that a small subset of expensive features drives most of the incremental performance gain in the augmented model. Features with negligible ΔAUC could potentially be deprioritised in a cost-sensitive setting without substantial loss in discrimination. When considered alongside the permutation importance and drop-sensitivity results from the inexpensive-feature model, these findings support the central practical motivation of this study: meaningful PCOS risk stratification can be achieved using lower-burden information, while a selected set of additional diagnostics provides incremental value when escalation is feasible

Table 10: Ablation results for augmented features

Feature Removed	AUC (%)	ΔAUC (%)
Follicle No. (R)	92.48	-2.49
Follicle No. (L)	93.93	-1.04
PRG (ng/mL)	94.18	-0.79
AMH (ng/mL)	94.19	-0.78
LH (mIU/mL)	94.24	-0.73
FSH/LH	94.33	-0.64
II beta-HCG (mIU/mL)	94.34	-0.63
RBS (mg/dl)	94.37	-0.60
Vit D3 (ng/mL)	94.37	-0.60
Avg. F size (R) (mm)	94.45	-0.52
TSH (mIU/L)	94.49	-0.48
I beta-HCG (mIU/mL)	94.54	-0.43
PRL (ng/mL)	94.63	-0.34
Hb(g/dl)	94.64	-0.33
FSH (mIU/mL)	94.68	-0.29
Avg. F size (L) (mm)	94.74	-0.23
Endometrium (mm)	94.83	-0.14

3.3. Decision Curve Analysis

We carried forward the OOF prediction estimates for the remaining experiments and decision-focused analyses, as these provide the most stable out-of-sample basis for comparing clinical utility under the unified framework described in Section 2.5. We designed a decision curve analysis with 2,000 stratified bootstrap resamples and compared the net benefit of the inexpensive-feature and augmented models across threshold probabilities from 0.05 to 0.50. As shown in Fig 9, both model configurations deliver positive clinical net benefit across the examined thresholds relative to the "treat none" strategy, which remains flat at zero. They also outperform the "treat all" strategy over nearly the entire range, with the treat-all curve declining rapidly as threshold increases and crossing zero at approximately one-third risk. Comparing the two feature configurations, the augmented model consistently lies above the inexpensive-feature model, indicating incremental decision value from adding laboratory and ultrasound variables.

The gain is most evident in the 0.15 to 0.35 threshold region, where many screening and referral decisions are likely to be made. At very low thresholds, around 0.05 to 0.10, the curves are closer, but the augmented model still retains a slight advantage; at higher thresholds, net benefit tapers for all models, yet the augmented model continues to remain higher. Between algorithms, the differences are modest and the bootstrap confidence intervals overlap across most thresholds. RF is slightly higher than LR in the augmented setting, consistent with its stronger precision, while LR tracks closely and at times leads in the inexpensive-feature setting, consistent with its higher recall and more favourable raw calibration. Clinically, this supports a practical strategy in which inexpensive variables can be used for lower-burden initial screening, while escalation to the augmented model provides additional net benefit when more resource-intensive diagnostics are available.

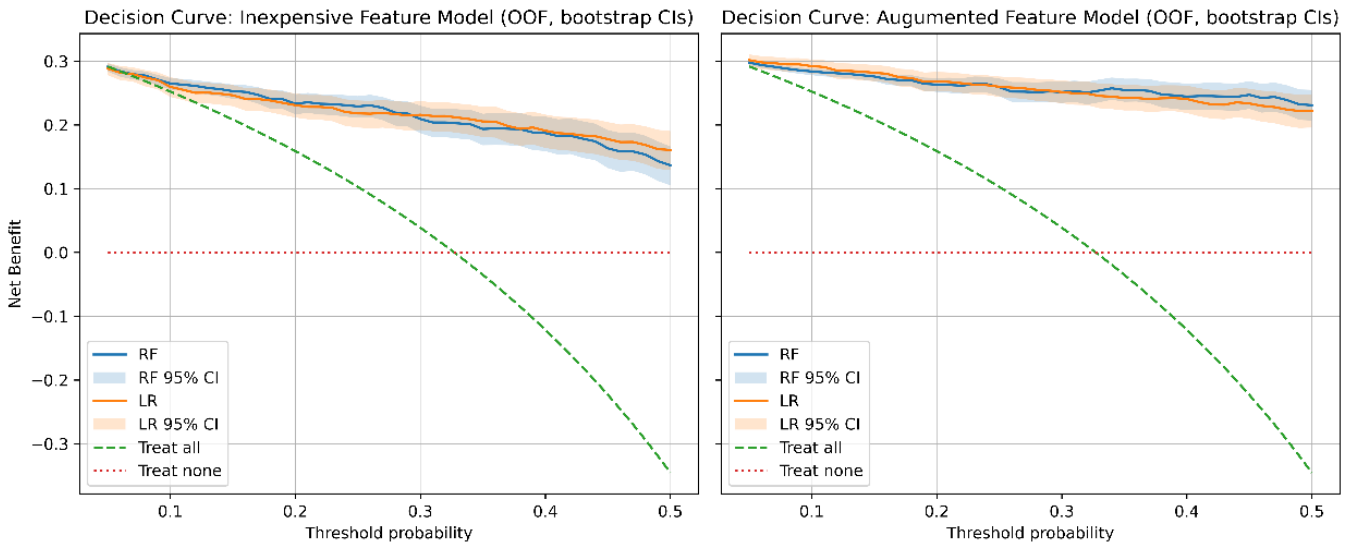

Figure 9: Decision curve plots for the inexpensive-feature model (left) and the augmented model (right) using OOF predictions. Net benefit is plotted against threshold probability with 95% bootstrap confidence intervals.

3.4. Capacity-Constrained Triage

To extend the decision curve analysis into a more operational setting, we next evaluated how referral performance changes when diagnostic capacity is limited. Using the OOF risk estimates from the baseline models and the triage policies defined in Section 2.5.1, we compared Risk Top-K, Utility Top-K, and Uncertainty-Aware referral under fixed referral capacities of 20%, 40%, and 60% as summarized in Table 11. The aim was to identify the highest net benefit strategy, to examine how uncertainty-based deferral behaves when referral slots are scarce [28], [53].

At all capacities (20%, 40%, 60%), the two baselines (Risk Top-K and Utility Top-K) are essentially indistinguishable: the curves lie on top of each other across thresholds as seen in Fig 10,11 and 12. That matches our definition of the utility score i.e si=pi-w(1-pi) with a fixed threshold pt: because si is a monotone transform of pi under symmetric costs, ranking by utility selects the same patients as ranking by risk. The uncertainty-aware policy (prioritising abstentions first, then filling to capacity by utility) underperforms the baselines at every capacity across the shown threshold range (0.05-0.50).

At 20% capacity, the uncertainty-aware policy yields lower net benefit than Risk Top-K across most thresholds as presented in Figure 10, indicating that when referral slots are highly scarce, prioritising the very highest-risk cases is more effective than diverting capacity toward abstained or ambiguous cases. The gap is largest at lower thresholds, where the uncertainty-aware curve is even outperformed by the treat-all reference up to approximately pt ≈ 0.22, showing how costly this diversion can be under severe scarcity. In this setting, abstaining on borderline-confidence cases reduces true positive yield more than it reduces false positives. This pattern is consistent with selective-classification theory, which predicts that abstention becomes disadvantageous when effective coverage is tight [54].

At 40% capacity, risk-based selection continues to maintain higher net benefit than the uncertainty-aware policy across most thresholds, while the Utility Top-K curve again lies almost exactly on top of the Risk Top-K curve as shown in Fig 11. The gap therefore persists at this intermediate capacity: prioritising abstentions still displaces higher-value, higher-risk cases that would otherwise contribute more to net benefit. At the same time, this 40% setting provides the most informative balance between case yield and overtreatment in the present study, because it delivers a substantial increase in detected true positives relative to the 20% scenario without the markedly higher false-positive burden observed at 60% capacity.

Table 11: Best net benefit by policy and capacity

Capacity (% of cohort triaged)	NB (uncertainty aware)	NB (Risk Top-K)	NB (Utility Top-K)	ΔNB (uncertainty aware - Risk Top-K)	ΔNB (uncertainty aware - Utility Top-K)
20	0.125	0.154	0.154	-0.029	-0.029
40	0.162	0.270	0.270	-0.106	-0.106
60	0.297	0.297	0.297	0.000	0.000

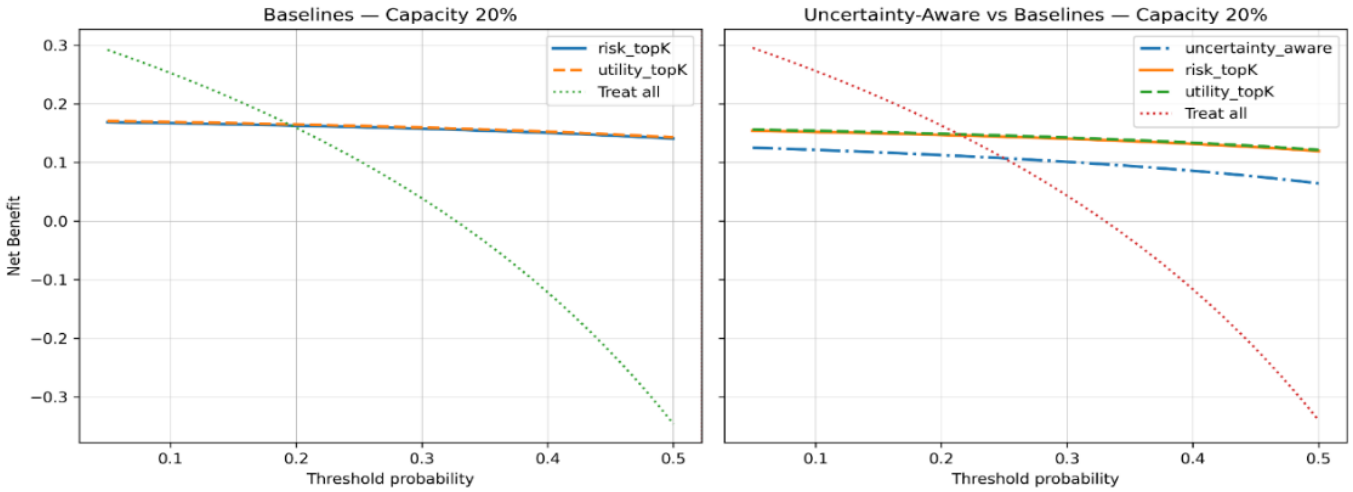

Figure 10: Triage Net Benefit baselines (Risk Top-K vs Utility Top-K) and uncertainty aware, 20% capacity.

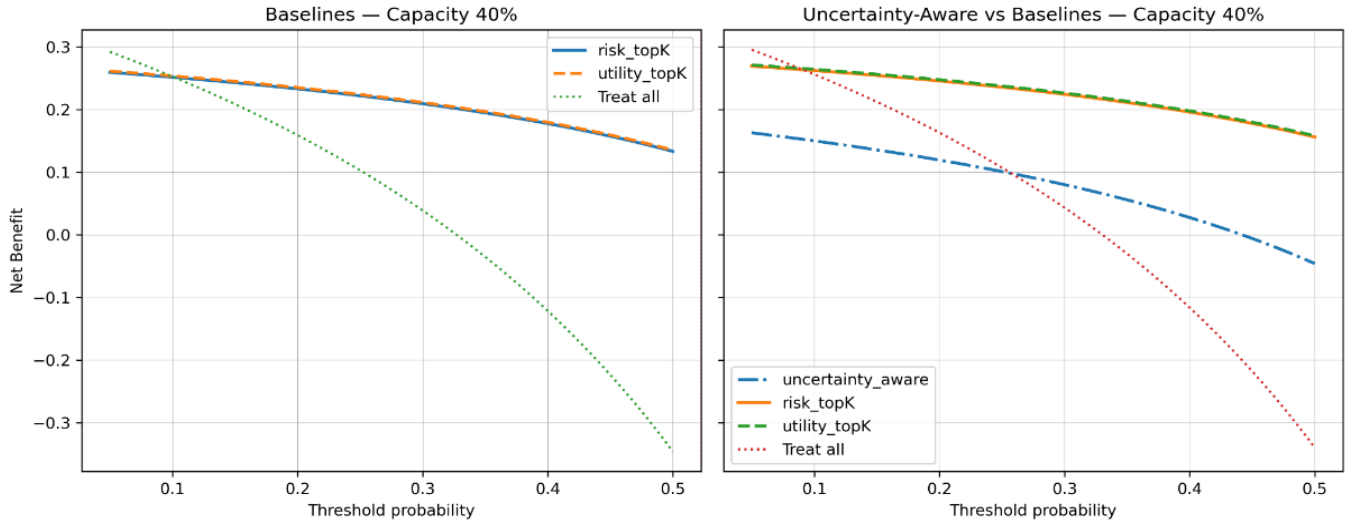

Figure 11: Triage Net Benefit baselines (Risk Top-K vs Utility Top-K) and uncertainty aware, 40% capacity.

At 60% capacity, risk-based selection continues to remain above the treat-all reference, while the Utility Top-K curve again tracks it almost exactly as shown in Fig 12. At this more generous capacity level, all three policies converge and produce nearly identical net benefit across thresholds. Once referral capacity is sufficiently high, the top-K sets selected by the different rules overlap strongly, and the effect of prioritising abstentions becomes much smaller. In practical terms, uncertainty-aware referral becomes less costly to decision utility as available capacity increases, even though it does not surpass risk-based selection in net benefit.

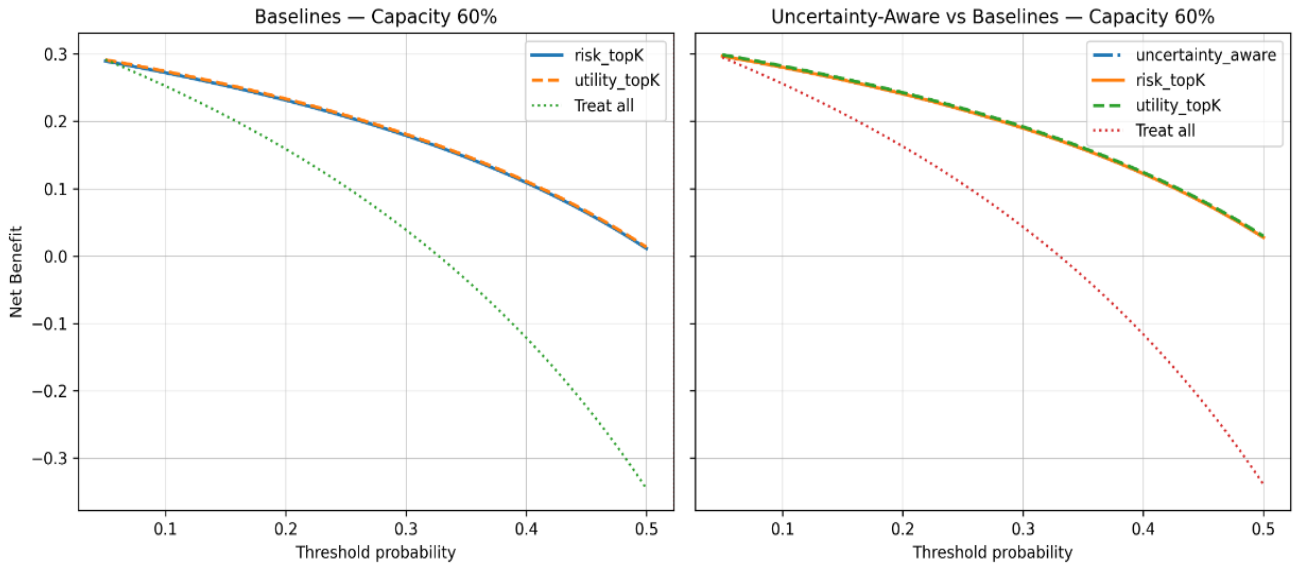

Figure 12: Triage Net Benefit baselines (Risk Top-K vs Utility Top-K) and uncertainty aware, 60% capacity.

Across capacity settings, the net-benefit curves decline monotonically with threshold, so the maximum net benefit occurs at the lowest tested threshold, approximately 0.05, for each policy. In this dataset and policy design, rank-by-risk selection remains preferable at constrained capacities, while uncertainty-first selection does not improve net benefit and often reduces it when referral slots are limited. Only when capacity becomes more generous do the three strategies begin to behave similarly. These plot findings are consistent with the summary statistics in Table 12 and indicate that uncertainty-aware referral should be interpreted primarily as a cautious decision-control mechanism for ambiguous cases; the findings do not support its interpretation as a universally superior triage rule.

In the present setup, the utility function reduces to a monotone transformation of predicted risk under the same threshold-induced trade-off used in standard decision curve analysis. Under that symmetry, Utility Top K and Risk Top-K pick the same set of patients at the (empirically optimal) threshold, so their net benefits coincide. This is expected when the policy's "value" is a monotone transform of risk under fixed class costs and a shared threshold [28].

3.5. Conformal Prediction: coverage and abstention by age/BMI

We applied class-conditional conformal prediction ($\alpha = 0.05$) to the inexpensive-feature Random Forest model using OOF predictions from Table 8 and summarised performance overall and by age and BMI subgroups in Table 13. In the context of the present framework, this conformal layer serves as the final uncertainty-control component, allowing the model either to issue a confident singleton prediction or to abstain and defer cases that remain ambiguous under lower-burden first-line information. For the overall cohort, the conformal predictor achieved 94.5% coverage (95% CI: 88.4-98.0%), close to the nominal 95% target. The abstention rate was 41.3% (95% CI: 32.0-51.1%), while the singleton rate was 58.7% (95% CI: 48.9-68.1%), indicating that more than half of predictions were issued as confident single-class outputs while maintaining near-nominal validity.

Across age quartiles, as shown in Fig 13, coverage remained high but varied alongside abstention. Age Q1 achieved 88.6% coverage (95% CI: 73.3-96.8%) with an abstention rate of 34.3%. Age Q2 and Age Q4 both achieved 100.0% coverage (95% CI: 85.8-100.0% and 86.8-100.0%, respectively), accompanied by abstention rates of 41.7% and 42.3%. Age Q3 recorded 91.7% coverage (95% CI: 73.0-99.0%) with the highest abstention rate at 50.0%, alongside a matched singleton rate of 50.0%. These results suggest that coverage is preserved across age strata largely through variation in abstention, with higher deferral in groups where the inexpensive-feature model is less decisive.

Patterns by BMI subgroup were similar. Participants with normal BMI showed 94.3% coverage (95% CI: 84.3-98.8%), the lowest abstention rate (32.1%), and the highest singleton rate (67.9%). The overweight group achieved 92.7% coverage with an abstention rate of 43.9%. Both obese and underweight groups achieved 100.0% coverage; however, this was accompanied by substantially higher abstention rates (75.0% and 57.1%, respectively) and wide confidence intervals due to small sample sizes. This pattern indicates that near-nominal validity can be retained in smaller or less data-rich subgroups, but often at the cost of greater abstention and reduced decisiveness.

Table 13 further disaggregates singleton predictions into positive and negative assignments and stratifies results by PCOS status.

Overall, 58.7% of predictions were singletons, comprising 15.6% positive singletons and 43.1% negative singletons, indicating that the conformal predictor more frequently issued confident negative predictions at the inexpensive-feature screening stage. Stratification by outcome revealed clear differences. Among individuals with PCOS, abstention increased to 52.8% and the singleton rate fell to 47.2%, with singletons dominated by positive assignments (38.9%) while negative ones (8.3%). In contrast, individuals without PCOS exhibited lower abstention (35.6%) and a higher singleton rate (64.4%), with negative singletons predominating (60.3%) and positive singletons occurring infrequently (4.1%). Across age and BMI subgroups, negative singletons were more common in larger, data-rich strata, whereas smaller subgroups showed higher abstention and fewer confident assignments. This pattern suggests that the conformal mechanism adjusts its decisiveness according to both outcome distribution and subgroup sample density.

Table 12: Best threshold and net benefit by referral capacity, with corresponding true-positive and false-positive trade-offs.

Capacity	Optimal Threshold	NB@best (Utility Top K)	NB@best (Risk Top K)	TP@best (Utility Top K)	FP@best (Utility Top K)
20%	0.05	0.168	0.168	92	16
40%	0.05	0.259	0.259	144	72
60%	0.05	0.289	0.289	165	159

Table 13: Conformal prediction overall and subgroup metrics for the inexpensive-feature Random Forest model ($\alpha = 0.05$).

Group	Count	Coverage (%)	Coverage 95% CI (%)	Abstention (%)	Abstention 95% CI (%)	Singleton rate (%)	Singleton rate 95% CI (%)	Positive Singleton (%)	Negative Singleton (%)
Overall	109	94.5	88.4-98.0	41.3	32.0-51.1	58.7	48.9-68.1	15.6	43.1
With PCOS	36	91.7	77.5-98.2	52.8	35.5-69.6	47.2	30.4-64.5	38.9	8.3
Without PCOS	73	95.9	88.5-99.1	35.6	24.7-47.7	64.4	52.3-75.3	4.1	60.3
Age Quartile 1	35	88.6	73.3-96.8	34.3	19.1-52.2	65.7	47.8-80.9	34.3	31.4
Age Quartile 2	24	100.0	85.8-100.0	41.7	22.1-63.4	58.3	36.6-77.9	16.7	41.6
Age Quartile 3	24	91.7	73.0-99.0	50.0	29.1-70.9	50.0	29.1-70.9	4.2	45.8
Age Quartile 4	26	100.0	86.8-100.0	42.3	23.4-63.1	57.7	36.9-76.6	0.0	57.7
BMI Normal	53	94.3	84.3-98.8	32.1	19.9-46.3	67.9	53.7-80.1	13.2	54.7
BMI Overweight	41	92.7	80.1-98.5	43.9	28.5-60.3	56.1	39.7-71.5	19.5	36.6
BMI Obese	8	100.0	63.1-100.0	75.0	34.9-96.8	25.0	3.2-65.1	25.0	0.0
BMI Underweight	7	100.0	59.0-100.0	57.1	18.4-90.1	42.9	9.9-81.6	0.00	42.9

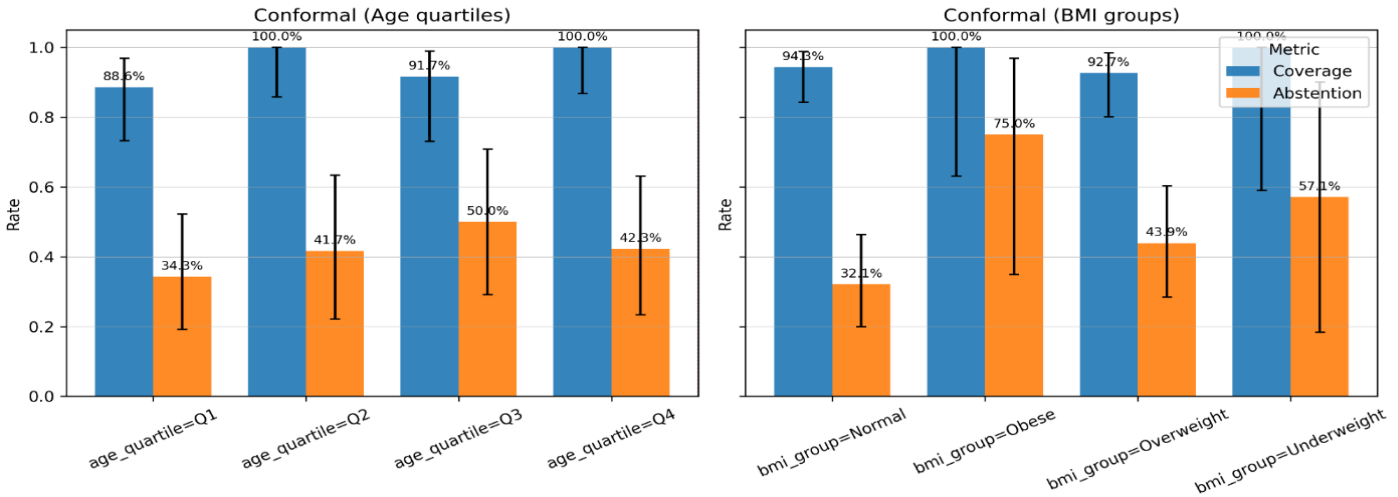

Figure 13: Grouped bars with error bars for coverage and abstention by (left) age quartile and (right) BMI group

Because abstention alone does not show how reliable the actioned predictions are, we next examined performance conditional on singleton decisions and the distribution of abstentions across subgroups. Table 14 summarizes abstention counts, the number of actioned cases, selective accuracy on singletons, and error rates among actioned predictions. Overall, 64 of 109 cases were actioned as singletons, with a selective accuracy of 90.6%. Among actioned predictions, false negatives occurred in 17.6% of positive cases, while false positives among negative cases were limited to 6.4%. Across age groups, selective accuracy varied from 82.6% in Age Quartile 1 to 100.0% in Quartiles 2 and 4. The largest abstention gap was observed in Age Quartile 3, where selective accuracy dropped to 83.3% and false-negative rates among actioned positives were elevated. Similar patterns were observed across BMI groups, with Normal BMI serving as the lowest-abstention reference. Obese and Underweight subgroups exhibited higher abstention gaps, consistent with small sample sizes and wider uncertainty. These results show that the conformal layer are controls coverage and also concentrates many uncertain cases into the deferred set, leaving the actioned singleton predictions comparatively accurate, while still requiring caution in smaller strata.

3.6. Statistical Comparisons

A series of statistical tests were conducted to evaluate whether the observed differences in model performance across metrics and settings were statistically significant. The results are summarised in Tables 15 to 18. These tests provide formal support for the descriptive trends reported in Sections 3.1 to 3.5, including the discrimination gains of the augmented model, the calibration advantage of LR, and the stronger downstream decision profile of RF.

Table 14: Selective performance metrics under conformal prediction across subgroups

Group	Abstention count, Set Size >= 2	Actioned (singletons), Set size = 1	Selective accuracy on singletons (%)	FN among actioned positives (%)	FP among actioned negatives (%)	Abstention gap vs lowest (%)	Lowest abstention reference group
Overall	45	64	90.6	17.6	6.4	NaN	NaN
With PCOS	19	17	82.4	17.6	NaN	NaN	NaN
Without PCOS	26	47	93.6	NaN	6.4	NaN	NaN
Age Quartile 1	12	23	82.6	10	23.1	0	Q1
Age Quartile 2	10	14	100	0	0	7.4	Q1
Age Quartile 3	12	12	83.3	66.7	0	15.7	Q1
Age Quartile 4	11	15	100	NaN	0	8	Q1
Normal	17	36	91.7	25	3.6	0	Normal
Overweight	18	23	87	14.3	12.5	11.8	Normal
Obese	6	2	100	0	NaN	42.9	Normal
Underweight	4	3	100	NaN	0	25.1	Normal

Note: NaN indicates "not estimable," either because the subgroup has zero actioned positives (FN) or zero actioned negatives (FP) among singleton cases, or because the abstention gap and reference group are only defined within a multi-level stratum (age quartiles or BMI groups) and therefore do not apply to Overall or PCOS-status rows.

First, DeLong's test confirmed that both augmented models significantly outperformed their inexpensive-feature counterparts in AUC (Table 15). The augmented RF model versus the inexpensive-feature RF model showed a ΔAUC of 7.4% ($p = 0.001$), while the augmented LR model versus the inexpensive-feature LR model yielded a ΔAUC of 6.9% ($p = 0.005$). In contrast, the within-configuration comparisons between RF and LR were not statistically significant, with $p = 0.544$ for the inexpensive-feature models and $p = 0.334$ for the augmented models. These findings formally support the discrimination gains already observed in Table 8 and are consistent with the stronger net-benefit curves seen for the augmented models in Section 3.3.

Bootstrap resampling showed consistent improvements in Brier scores when moving from the inexpensive-feature models to the augmented models (Table 16). The augmented RF model achieved a Brier reduction of -0.041 relative to the inexpensive-feature RF model, while the augmented LR model achieved a Brier reduction of -0.045 compared with the inexpensive-feature LR model. In both cases, the confidence intervals excluded zero, indicating that the gains were not attributable to sampling variation alone. These reductions suggest that adding laboratory and ultrasound features improved overall probability quality, complementing the discrimination gains already shown in Table 15.

Calibration analysis (Table 17) further highlighted the augmented LR model's more favourable raw probability calibration, with the lowest Brier score (0.078) and ECE (0.033). By contrast, although the augmented RF model achieved the highest discrimination, its ECE remained higher (0.101), consistent with the reliability diagrams discussed in Section 3.1. These findings reinforce the distinction already established in the baseline comparison: LR provided the most stable raw calibration, whereas RF offered the strongest overall discrimination and precision for downstream decision support.

McNemar's test (Table 18) supported the classification-level improvements observed for the augmented models. The augmented RF versus inexpensive-feature RF comparison revealed a highly significant imbalance in discordant predictions ($\chi^2 = 26.882$, $p < 0.001$), while the augmented LR versus inexpensive-feature LR comparison was also significant ($\chi^2 = 13.299$, $p = 0.001$). In contrast, LR and RF did not differ significantly within either the inexpensive-feature setting or the augmented setting. This indicates that the gains associated with adding laboratory and ultrasound variables reflect genuine reclassification improvement and are unlikely to be explained by chance fluctuations in paired predictions.

These statistical comparisons corroborate the patterns observed throughout the Results section (3.1-3.5). The augmented models provide statistically significant gains in discrimination and probability quality relative to the inexpensive-feature models, while LR retains the most favourable raw calibration profile and RF remains the strongest overall baseline for discrimination, precision, and downstream decision-focused analysis. In this sense,

Table 15: Pairwise AUC comparisons between inexpensive-feature & augmented models using DeLong's test.

Model 1	Model 2	Model 1 (AUC %)	Model 2 (AUC %)	Δ AUC (%)	Z Statistic	P value
Inexpensive-feature RF	Inexpensive-feature LR	88.6	87.6	1.0	0.607	0.544
Augmented RF	Augmented LR	96.0	94.5	1.5	0.965	0.334
Augmented RF	Inexpensive-feature RF	96.0	88.6	7.4	3.181	0.001
Augmented LR	Inexpensive-feature LR	94.5	87.6	6.9	2.828	0.005

Table 16: Bootstrap differences in Brier scores between inexpensive-feature and augmented models.

Model 1	Model 2	Δ Brier (1 - 2)	95% CI (Low)	95% CI (High)
Inexpensive-feature RF	Inexpensive-feature LR	0.003	-0.006	0.011
Augmented RF	Augmented LR	0.006	-0.004	0.017
Augmented RF	Inexpensive-feature RF	-0.041	-0.054	-0.029
Augmented LR	Inexpensive-feature LR	-0.045	-0.062	-0.027

Table 17: Calibration scores with 95% confidence intervals for inexpensive-feature and augmented models.

Model	Brier	ECE	ECE 95% CI (Low)	ECE 95% CI (High)
Inexpensive-feature RF	0.126	0.063	0.048	0.097
Inexpensive-feature LR	0.123	0.049	0.038	0.083
Augmented RF	0.084	0.101	0.084	0.126
Augmented LR	0.078	0.033	0.025	0.067

Table 18: McNemar's test results comparing inexpensive-feature and augmented models.

Model 1	Model 2	0→1 Errors	1→0 Errors	Chi- Square	P- Value
Inexpensive-feature RF	Inexpensive-feature LR	25	12	3.892	0.143
Augmented RF	Augmented LR	16	21	0.432	0.806
Augmented RF	Inexpensive-feature RF	21	72	26.882	0.000
Augmented LR	Inexpensive-feature LR	22	55	13.299	0.001

the statistical evidence supports the central practical conclusion of the study: meaningful PCOS risk stratification is already achievable from inexpensive information, while a limited set of additional diagnostics yields significant incremental value when escalation is feasible.

4. Discussion of Result

The cost-aware modelling framework in this study demonstrated that predictive performance can be meaningfully enhanced by moving from the inexpensive-feature model to the augmented model that adds laboratory and ultrasound variables when fuller diagnostic work-up is available. Across both Random Forest (RF) and Logistic Regression (LR), the augmented models achieved significant AUC gains over their inexpensive-feature counterparts (ΔAUC = 7.4% for RF, p = 0.001; ΔAUC = 6.9% for LR, p = 0.005) as presented in Table 15. At the same time, the inexpensive-feature model retained substantial predictive value, showing that lower-burden clinical information can already support meaningful first-line PCOS risk stratification before escalation becomes necessary.

In this work, feature importance and sensitivity analysis revealed that over 80% of the predictive performance of the inexpensive-feature model could be retained with the top 16 inexpensive features, potentially reducing administrative workload and unnecessary data collection while maintaining screening utility. Removing low-ranked inexpensive features resulted in consistent gains in discrimination, with exclusion of the bottom three, five, and eight features each producing incremental improvements in AUC (Figs 7 and 8). This suggests that several routine variables introduced noise as against signal. These findings indicate that a reduced inexpensive-feature model may offer a more stable and efficient first-line approach to PCOS risk assessment, particularly in settings where rapid and accessible screening is clinically

advantageous. In addition, correlated body-size measures indicated that multicollinearity may reduce their marginal contribution to prediction in this study, even though such variables remain clinically relevant for subgroup definition and downstream analysis.

Similarly, ablation of the expensive additions in the augmented model produced ΔAUC values ranging from approximately 0.14% to 2.49% as shown in Table 10. The largest losses followed removal of Follicle Number (R), Follicle Number (L), and PRG, indicating that a relatively small subset of laboratory and ultrasound variables accounts for much of the incremental performance gain beyond the inexpensive baseline. Other features contributed more modestly, suggesting that not every additional diagnostic measurement provides the same practical value once stronger predictors are already present. This reinforces the central motivation of the framework: lower-burden information may support useful first-line stratification, while a selected subset of more resource-intensive diagnostics provides targeted incremental value when escalation is feasible. The baseline comparison also revealed a consistent trade-off between algorithms. LR provided more favourable raw probability calibration across both feature settings, while RF delivered the strongest overall discrimination and precision, particularly in the augmented setting. This distinction is important because it explains why RF was carried forward for the subsequent feature sensitivity, decision curve, triage, and conformal analyses, while LR remained useful as a benchmark for calibration-aware comparison. In other words, the downstream analyses build on the model that offered the strongest decision-facing profile, while still acknowledging that calibration remains a critical consideration when probability estimates are used directly for threshold-based action.

Decision curve analysis further demonstrated that the augmented model provides higher net benefit than the inexpensive-feature model across clinically relevant threshold regions, particularly between approximately 0.15 and 0.35, where many screening and referral decisions are likely to be made. This finding indicates that the additional laboratory and ultrasound information has practical decision value, not only statistical value. However, the inexpensive-feature model also maintained positive net benefit across the examined thresholds, which is important in settings where specialised diagnostics are delayed, unavailable, or financially burdensome. The clinical implication is therefore not that the inexpensive-feature model should replace fuller diagnostic work-up, but that it may support lower-burden first-line risk stratification while preserving the option to escalate when additional diagnostic value is warranted.

Furthermore, a key observation from this study concerns the behaviour of uncertainty-aware triage under different referral capacities. At 20% and 40% capacity, the uncertainty-aware policy did not outperform risk-based selection; across threshold regions, its net benefit was lower because prioritising abstained or borderline-confidence cases reduced true-positive yield more than it reduced false positives. In the PCOS context, this suggests that when referral slots are highly constrained, diverting capacity toward uncertainty alone may delay identification of individuals with clinically meaningful risk profiles. Only at higher capacity (60%), where the referral sets selected by different policies overlap much more strongly, did all three strategies converge to similar net benefit. These findings suggest that uncertainty-aware triage is best interpreted not as a universally superior referral rule, but as a cautious decision-control mechanism whose operational cost becomes smaller as available capacity increases.

Interestingly, at the 40% capacity level, the model delivers the most informative balance between case yield and over treatment. Net benefit rises by 54% (from 0.168 to 0.259) in Table 12, and when contrasted against the uncertainty aware strategy in Table 11, it produces an incremental gain of 0.106. This operating point captures a substantial increase in true positives relative to the 20% setting, which yielded only 0.029, while maintaining a far more acceptable false positive burden than the 60% configuration, which delivered no incremental gain. In this work, the 40% capacity therefore represents the most clinically efficient region for PCOS treatment allocation, providing meaningful improvement in detected cases without a disproportionate rise in unnecessary interventions. Within the present dataset and policy design, the 40% capacity setting appeared to provide the most informative balance between true-positive yield and overtreatment, offering a substantial gain over the 20% scenario without the broader false-positive burden observed at 60%.

Building on the triage framework, the conformal prediction analysis clarifies how uncertainty can be operationalized within the inexpensive-feature screening stage. Overall marginal coverage reached 94.5%, close to the nominal 95% target, while the abstention rate of 41.3% indicates that a substantial proportion of cases are explicitly deferred when lower-burden first-line information is insufficient for confident classification. The conformal layer separates cases that can be actioned as singleton predictions from those that warrant escalation for fuller assessment, thereby avoiding forced binary decisions when prediction uncertainty remains high. The additional selective-performance results further show that actioned singleton predictions were comparatively accurate overall, while many uncertain cases were concentrated into the deferred set. This supports the interpretation of conformal prediction as the final decision-control layer of the framework, governing when

inexpensive first-line screening is sufficient and when escalation is more appropriate.

The subgroup results reinforce this interpretation while highlighting the limits imposed by small subgroup sizes. Coverage remained high across age and BMI strata, including subgroups with smaller counts, but in several of these cases this was achieved through increased abstention rather than more confident classification. This pattern was most pronounced in the obese and underweight BMI groups, where abstention exceeded 57% and confidence intervals widened because of limited counts. In this setting, higher abstention is better interpreted as model caution than as model failure: when data are sparse or subgroup patterns are less stable, the conformal mechanism defers as against forcing potentially unreliable early decisions. At the same time, these subgroup findings should be interpreted cautiously and viewed as exploratory, given the limited sample sizes.

Although the individual components used in this work are not novel in isolation, the contribution of this study lies in their operational integration within a single PCOS-oriented framework. Specifically, the work combines cost-aware feature evaluation, baseline comparison between inexpensive and augmented models, decision curve analysis, capacity-constrained triage, and conformal prediction with subgroup-level abstention analysis. In this sense, the novelty is not algorithmic, but methodological and practical: the framework shows how established tools can be assembled to evaluate PCOS risk prediction in a way that reflects diagnostic burden, uncertainty, and constrained clinical capacity more directly than conventional accuracy-focused studies. Overall, this discussion supports a consistent interpretation of the findings. The inexpensive-feature model provides meaningful lower-burden first-line stratification, the augmented model yields significant incremental gains when fuller diagnostics are available, RF offers the strongest downstream decision-facing profile and conformal prediction provides a structured mechanism for deferring ambiguous cases under operational constraints. These strengths, however, should be interpreted within the limits of the study design and dataset, which are discussed below.

5. Limitations & Future Works

This study has several limitations that should be considered when interpreting the findings, which reflect the exploratory and proof-of-concept nature of the work. The results are intended to illustrate the potential of a cost-aware, uncertainty-aware AI framework for more efficient and clinically grounded PCOS risk stratification under operational constraints, while recognising that further validation is required before deployment guidance can be established.

First, the dataset size was relatively modest (n = 541), which limits statistical power for subgroup analysis, particularly in smaller BMI categories, and constrains the generalisability of the results. External validation on larger and more diverse PCOS cohorts will be essential to assess robustness across populations and care settings. Second, the modelling and operational analyses were conducted on retrospective data. Prospective evaluation will be required to determine how uncertainty-aware triage and escalation based on inexpensive first-line screening perform in real clinical workflows, including their effects on diagnostic timing, clinician workload, referral efficiency, and patient outcomes. Third, this work deliberately focused on a limited set of well-established algorithms and subgroup definitions to support a proof-of-concept evaluation. Future studies could expand the framework to include additional modelling approaches, comparative analysis across algorithms, post hoc calibration, and alternative uncertainty-handling strategies. Similarly, subgroup analysis could be extended beyond age and BMI to incorporate richer clinical and socio-demographic factors relevant to PCOS heterogeneity and equity considerations.

Furthermore, the PCOS outcome label was derived from a dataset for which detailed clinical diagnostic criteria are not publicly available. As a result, label validity cannot be independently verified and may reflect heterogeneous diagnostic standards. This represents a key limitation because it may influence transportability beyond the study dataset and complicate interpretation of the absolute performance levels reported. Future work should therefore prioritise clinically curated datasets with explicit diagnostic definitions and external validation across settings.

In addition, although the framework is designed as an escalation pipeline, the augmented model was evaluated independently for decision curve analysis and capacity-constrained triage rather than being trained and assessed exclusively on cases escalated from the inexpensive-feature model. In real-world deployment, escalation would introduce selection effects and partial observability, particularly when laboratory and ultrasound assessments are unavailable for a substantial proportion of screened patients. Explicit modelling of this selection mechanism remains an important direction for future work and will be necessary to assess end-to-end pipeline behaviour more realistically.

Finally, decision curve analysis, triage policies, and conformal efficiency are inherently prevalence dependent. The reported results therefore reflect the prevalence structure of the study dataset and may not generalise directly to settings with different baseline PCOS prevalence. Future work should evaluate prevalence reweighting, recalibration strategies, and scenario-based

sensitivity analyses to support deployment in more specific clinical contexts. These limitations indicate that the present study should be interpreted as a structured proof-of-concept evaluation of how cost-aware modelling, uncertainty-aware triage, and conformal prediction may be combined in PCOS risk stratification, as oppose to a definitive deployment-ready diagnostic system.

6. Conclusion

This study presents a proof-of-concept, cost-aware AI framework designed to support PCOS risk stratification under operational and capacity constraints. By integrating targeted feature selection, decision-focused evaluation, uncertainty-aware triage, and conformal prediction, we demonstrate how predictive performance, reliability, and transparency can be considered jointly within a clinically motivated workflow. Instead of introducing new modelling techniques, the framework shows how well-established methods can be structured to prioritise inexpensive information, manage uncertainty explicitly, and guide escalation to more resource-intensive assessment when warranted.

Across the framework, we observed that the inexpensive-feature model provided meaningful first-line discrimination, while a limited set of added laboratory and ultrasound variables contributed significant incremental value in the augmented model. Feature reduction analysis showed that much of the predictive signal could be retained using a compact inexpensive subset, supporting lower-burden screening under constrained settings. Decision curve analysis illustrated how model-guided utility changes across thresholds and referral capacities, and conformal prediction provided near-nominal coverage with interpretable abstention across age and BMI subgroups. Together, these findings show how model evaluation can move beyond accuracy alone to reflect operational feasibility, uncertainty handling, and subgroup stability.

While the findings are not intended to support immediate clinical deployment or broad generalisation, they illustrate the potential of structured, uncertainty-aware AI frameworks to inform diagnostic decision-making in resource-constrained settings. In the context of PCOS, where diagnostic pathways are often delayed, heterogeneous, and resource intensive, this work contributes a practical methodological step toward AI-assisted triage strategies that may ultimately support more timely, transparent, and efficient assessment in women's health.

Conflict of Interest

The authors declare that no funding was received from any affiliated institution for this research. The work was conducted independently and the views expressed are solely those of the authors.

Data and Code Availability

The datasets analyzed during the current study is publicly available at Kaggle(https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data). The code supporting the findings of this study is available from the corresponding author on reasonable request.

References

[1] H. J. Teede, M. Bahri Khomami, R. Morman, J. S. E. Laven, A. E. Joham, M. F. Costello, M. Patil, D. A. Rees, L. Berry, M. G. Cree, H. Zhao, R. J. Norman, A. Dokras, and T. T. Piltonen, "Polyendocrine metabolic ovarian syndrome, the new name for polycystic ovary syndrome: A multistep global consensus process," *The Lancet*, vol. 407, pp. 2329–2339, 2026, doi: 10.1016/S0140-6736(26)00717-8.

[2] D. Lizneva, L. Suturina, W. Walker, S. Brakta, L. Gavrilova-Jordan, and R. Azziz, "Criteria, prevalence, and phenotypes of polycystic ovary syndrome," *Fertility and Sterility*, vol. 106, no. 1, pp. 6–15, 2016, doi: 10.1016/j.fertnstert.2016.05.003.

[3] P. D'Souza, D. E. Rodrigues, R. G. Kaipangala, L. K. Chacko, and J. D. Almeida, "Correlation between the physical signs, clinical parameters, and the quality of life in young women with polycystic ovarian syndrome: An explorative study," *Journal of South Asian Federation of Obstetrics and Gynaecology*, vol. 14, no. 1, pp. 17–21, 2022, doi: 10.5005/jp-journals-10006-1993.

[4] A. Ghafari, M. Maftoohi, M. E. Samarin, S. Barani, M. Banimohammad, and R. Samie, "The last update on polycystic ovary syndrome (PCOS), diagnosis criteria and novel treatment," *Endocrine and Metabolic Science*, vol. 17, Art. no. 100228, 2025, doi: 10.1016/j.endmts.2025.100228.

[5] C. Long, H. Feng, W. Duan, X. Chen, Y. Zhao, Y. Lan, and R. Yue, "Prevalence of polycystic ovary syndrome in patients with type 2 diabetes: A systematic review and meta-analysis," *Frontiers in Endocrinology*, vol. 13, Art. no. 980405, 2022, doi: 10.3389/fendo.2022.980405.

[6] R. S. Hussein, S. Bin Dayel, and O. Abahussein, "Polycystic ovary syndrome and reproductive health: A comprehensive review," *Clinical and Experimental Obstetrics & Gynecology*, vol. 51, no. 12, Art. no. 269, 2024, doi: 10.31083/j.ceog5112269.

[7] F. J. Barrera, E. D. L. Brown, A. Rojo, J. Obeso, H. Plata, E. P. Lincango, N. Terry, R. Rodríguez-Gutiérrez, J. E. Hall, and S. Shekhar, "Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: A systematic review," *Frontiers in Endocrinology*, vol. 14, Art. no. 1106625, 2023, doi: 10.3389/fendo.2023.1106625.

[8] J. Vazquez and J. C. Facelli, "Conformal prediction in clinical medical sciences," *Journal of Healthcare Informatics Research*, vol. 6, no. 3, pp. 241–252, 2022, doi: 10.1007/s41666-021-00113-8.

[9] X. Xu, Y. Jiang, J. Du, H. Sun, X. Wang, and C. Zhang, "Development and validation of a prediction model for suboptimal ovarian response in polycystic ovary syndrome (PCOS) patients undergoing GnRH-antagonist protocol in IVF/ICSI cycles," *Journal of Ovarian Research*, vol. 17, no. 1, Art. no. 116, 2024, doi: 10.1186/s13048-024-01437-w.

[10] Z. Zad, V. S. Jiang, A. T. Wolf, T. Wang, J. J. Cheng, I. C. Paschalidis, and S. Mahalingaiah, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," *Frontiers in Endocrinology*, vol. 15, Art. no. 1298628, 2024, doi: 10.3389/fendo.2024.1298628.

[11] S. A. Suha and M. N. Islam, "An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image," *Scientific Reports*, vol. 12, no. 1, Art. no. 17123, 2022, doi: 10.1038/s41598-022-21724-0.

[12] A. Alamoudi, I. U. Khan, N. Aslam, N. Alqahtani, H. S. Alsaif, O. Al Dandan, M. Al Gadeeb, and R. Al Bahrani, "A deep learning fusion approach to diagnosis the polycystic ovary syndrome (PCOS)," *Applied Computational Intelligence and Soft Computing*, vol. 2023, Art. no. 9686697, pp. 1–15, 2023, doi: 10.1155/2023/9686697.

[13] M. Agirsoy and M. A. Oehlschlaeger, "A machine learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features," *Scientific Reports*, vol. 15, Art. no. 33638, 2025, doi: 10.1038/s41598-025-10453-9.

[14] A. Zigarelli, Z. Jia, and H. Lee, "Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: Observational study," *JMIR Formative Research*, vol. 6, no. 3, Art. no. e29967, 2022, doi: 10.2196/29967.

[15] K. Shanmugavadivel, M. D. M. S., M. T. R., T. Al-Shehari, N. A. Alsadhan, and T. E. Yimer, "Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Art. no. 281, 2024, doi: 10.1186/s12911-024-02688-9.

[16] C. Tong, Y. Wu, Z. Zhuang, and Y. Yu, "A diagnostic model for polycystic ovary syndrome based on machine learning," *Scientific Reports*, vol. 15, Art. no. 9821, 2025, doi: 10.1038/s41598-025-92630-4.

[17] P. Kottarathil, "Polycystic ovary syndrome (PCOS)—Kerala dataset," *Kaggle*. Accessed: Jul. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos/data

[18] V. V. Khanna, K. Chadaga, N. Sampathila, S. Prabhu, V. Bhandage, and G. K. Hegde, "A distinctive explainable machine learning framework for detection of polycystic ovary syndrome," *Applied System Innovation*, vol. 6, no. 2, Art. no. 32, 2023, doi: 10.3390/asi6020032.

[19] M. M. Rahman, A. Islam, F. Islam, M. Zaman, M. R. Islam, M. S. Alam Sakib, and H. M. Hasan Babu, "Empowering early detection: A web-based machine learning approach for PCOS prediction," *Informatics in Medicine Unlocked*, vol. 47, Art. no. 101500, 2024, doi: 10.1016/j.imu.2024.101500.

[20] R. Taha, H. Zain El Abdin, and T. Musleh, "Comparative analysis of supervised machine learning models for PCOS prediction using clinical data," *Journal of Engineering Research and Sciences*, vol. 4, no. 6, pp. 16–26, 2025, doi: 10.55708/js0406003.

[21] S. Tiwari, L. Kane, D. Koundal, A. Jain, A. Alhudhaif, K. Polat, A. Zaguia, F. Alenezi, and S. A. Althubiti, "SPOSDS: A smart polycystic ovary syndrome diagnostic system using machine learning," *Expert Systems with Applications*, vol. 203, Art. no. 117592, 2022, doi: 10.1016/j.eswa.2022.117592.

[22] H. Danaei Mehr and H. Polat, "Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques," *Health and Technology*, vol. 12, pp. 137–150, 2022, doi: 10.1007/s12553-021-00613-y.

[23] S. Bharati, P. Podder, and M. R. H. Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, doi: 10.1109/TENSYMP50017.2020.9230932.

[24] K. M. Mohi Uddin, M. T. A. Bhuiyan, M. M. Rahman, M. M. Islam, and M. A. Uddin, "Early PCOS detection: A comparative analysis of traditional and ensemble machine learning models with advanced feature selection," *Engineering Reports*, vol. 7, no. 2, Art. no. e70008, 2025, doi: 10.1002/eng2.70008.

[25] B. Panjwani, J. Yadav, V. Mohan, N. Agarwal, and S. Agarwal, "Optimized machine learning for the early detection of polycystic ovary syndrome in women," *Sensors*, vol. 25, no. 4, Art. no. 1166, 2025, doi: 10.3390/s25041166.

[26] H. M. Emara, W. El-Shafai, N. F. Soliman, A. D. Algarni, R. Alkanhel, and F. E. Abd El-Samie, "A stacked learning framework for accurate classification of polycystic ovary syndrome with advanced data balancing and feature selection techniques," *Frontiers in Physiology*, vol. 16, Art. no. 1435036, 2025, doi: 10.3389/fphys.2025.1435036.

[27] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.

[28] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.

[29] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.

[30] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration," *Annals of Internal Medicine*, vol. 162, no. 1, pp. W1–W73, 2015, doi: 10.7326/M14-0698.

[31] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.

[32] K. G. M. Moons, D. G. Altman, Y. Vergouwe, and P. Royston, "Prognosis and prognostic research: Application and impact of prognostic models in clinical practice," *BMJ*, vol. 338, Art. no. b606, 2009, doi: 10.1136/bmj.b606.

[33] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E. A. Janssen, A. Blilie, L. Egevad, O. Spjuth, and M. Eklund, "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nature Communications*, vol. 13, no. 1, Art. no. 7761, 2022, doi: 10.1038/s41467-022-34945-8.

[34] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 1, Art. no. 9, 2016, doi: 10.3978/j.issn.2305-5839.2015.12.38.

[35] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[36] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[37] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[38] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX, USA, 2015, pp. 2901–2907, doi: 10.1609/aaai.v29i1.9602.

[39] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.

[40] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, Art. no. 15, 2012, doi: 10.1145/2382577.2382579.

[41] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-06649-8.

[42] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008, doi: 10.5555/1390681.1390693.

[43] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *The Annals of Statistics*, vol. 49, no. 1, pp. 486–507, 2021, doi: 10.1214/20-AOS1965.

[44] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934, doi: 10.1093/biomet/26.4.404.

[45] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.

[46] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a world beyond 'p < 0.05'," *The American Statistician*, vol. 73, suppl. 1, pp. 1–19, 2019, doi: 10.1080/00031305.2019.1583913.

[47] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *BMJ*, vol. 350, Art. no. g7594, 2015, doi: 10.1136/bmj.g7594.

[48] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[49] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, Art. no. 22588, 2023, doi: 10.1038/s41598-023-49962-w.

[50] D. R. Cutler, T. C. Edwards, Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, doi: 10.1890/07-0539.1.

[51] R. García-Carretero, R. Holgado-Cuadrado, and Ó. Barquero-Pérez, "Assessment of classification models and relevant features on nonalcoholic steatohepatitis using random forest," *Entropy*, vol. 23, no. 6, Art. no. 763, 2021, doi: 10.3390/e23060763.

[52] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, Bonn, Germany, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[53] A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "A simple, step-by-step guide to interpreting decision curve analysis," *Diagnostic and Prognostic Research*, vol. 3, Art. no. 18, 2019, doi: 10.1186/s41512-019-0064-7.

[54] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, pp. 1605–1641, 2010, doi: 10.5555/1756006.1859904.

About the Authors

Adewale Alex Adegoke currently works as a Senior Data Analyst at Office for National Statistics supporting economics statistics production. He previously worked as a Data Systems Manager at the Westminster Foundation for Democracy (WFD) in the UK, where applied his extensive knowledge of Data science and analytics to implement robust data solutions for insights and strategic decision-making that are pivotal to delivering global programmes. His research focuses on exploring and examining the reliability and dependability of machine learning systems used in the healthcare and wellbeing domains, with a keen interest in measuring prediction uncertainty, distribution-free confidence methods, AI security/guardrails and responsible use of AI. He holds a MSc in Applied Artificial Intelligence and Data Science from Southampton Solent University UK, as well as a B.Eng. in Civil Engineering from the Federal University of Technology, Minna Nigeria. During his MSc, he made significant contributions to several innovative research initiatives, notably in applying natural language processing for emotion detection and sentiment analysis of social media discussions.

Idris Babalola is a Senior Data Scientist with the Department of Health and Social Care United Kingdom. He holds a B.Eng. in Chemical Engineering from Federal University of Technology Minna Nigeria and an MSc in AI and Data Science from Southampton Solent University UK. He has contributed to EU-funded applied machine learning research with Istya Air, a French-based company focused on indoor air quality (IAQ) prediction. The prototypes developed during this phase were later operationalised and served as a foundation for large-scale deployment during the Olympic Games Paris 2024. He has also previously held part-time academic positions at Southampton Solent University, including MSc Research Supervisor in data science and artificial intelligence, and Associate Lecturer in Computing. His research interests focus on the reliability and trustworthiness of machine learning systems in health, including model uncertainty quantification, conformal prediction, AI security/guardrails and responsible use of AI. He also works extensively in applied NLP, retrieval-augmented generation and API-driven LLM experimentation for real-world healthcare applications. He currently volunteers his service as a peer reviewer for Information Research in Sweden, reviewing AI and data science manuscripts within the information science domain.

Peter Adebayo Odesola is a data analyst and applied AI researcher whose work focuses on designing trustworthy, statistically robust machine learning systems for healthcare and other data domains. He holds a BTech in Human Anatomy from Ladoke Akintola University of Technology, Nigeria and a Postgraduate Diploma in Education completed prior to his MSc. in Artificial Intelligence and Data Science from Solent University, Southampton UK. His professional experience spans roles as an Independent Researcher in Health, AI and Data Science in collaboration with data scientists consultant and industry experts.