

Symptom-Based Risk Prediction for Ovarian Cancer Across Menopausal Status

Nduta Mugo¹, Xin Guan¹, Amy Newnham², Georgios Gkoutos¹, Sudha Sundar¹, Andreas Karwath¹, on behalf of ROCkeTS collaborators

¹Department of Cancer & Genomic Sciences, University of Birmingham, UK, ²Warwick Medical School, University of Warwick, UK

Introduction

- Ovarian cancer (OC): late diagnosis and high mortality (~50%)
- Presents with non-specific symptoms
- Existing tools show variable performance by menopausal status
- Need for clinically interpretable risk stratification tools using routinely collected clinical and diagnostic data

Objective

- Develop and validate an interpretable symptom-based machine learning (ML) framework for ovarian cancer risk prediction, stratified by menopausal status

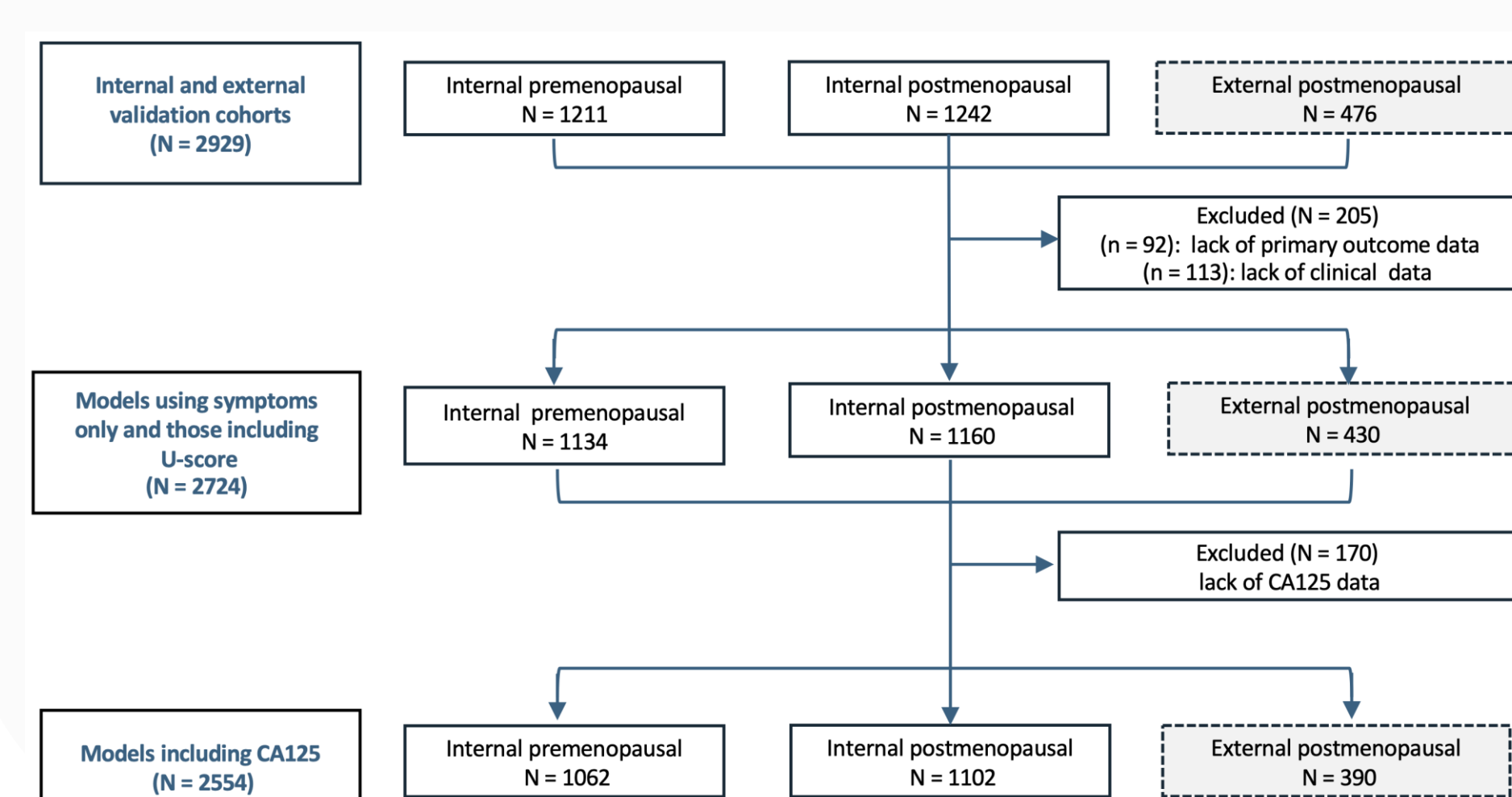
Datasets and Participants

- Training dataset:
 - Prospective multicentre cohort of women referred with suspected ovarian cancer from **ROCkeTS** study (2015-2023)
 - 67% referred from GP-initiated fast-track suspected cancer pathway
 - Participants stratified into **premenopausal** and **postmenopausal** groups
 - Internal cross-validation and external validation:
 - Total N = 2724 participants
 - Training postmenopausal Train-post: N=1160
 - Training premenopausal Train-pre: N = 1134
 - External postmenopausal Ext-post: N = 430
- ROCkeTS study detail previously published ^{1,3}

Outcome and Predictor Variables

- Outcome variable:**
 - Ovarian cancer diagnosis*
- OC = primary ovarian *or* secondary malignancy *or* borderline neoplasms *or* uncertain malignancy *or* unclassified malignancy
- Predictor variables:**
 - 15 symptoms; CA125; ultrasound score (U-score), BMI
- Structured symptom data collected at trial registration
- Ultrasound score collected within trial
- CA125 mainly from GP prior to referral (75% Train-pre, 76% Train-post & 90% Ext-post)

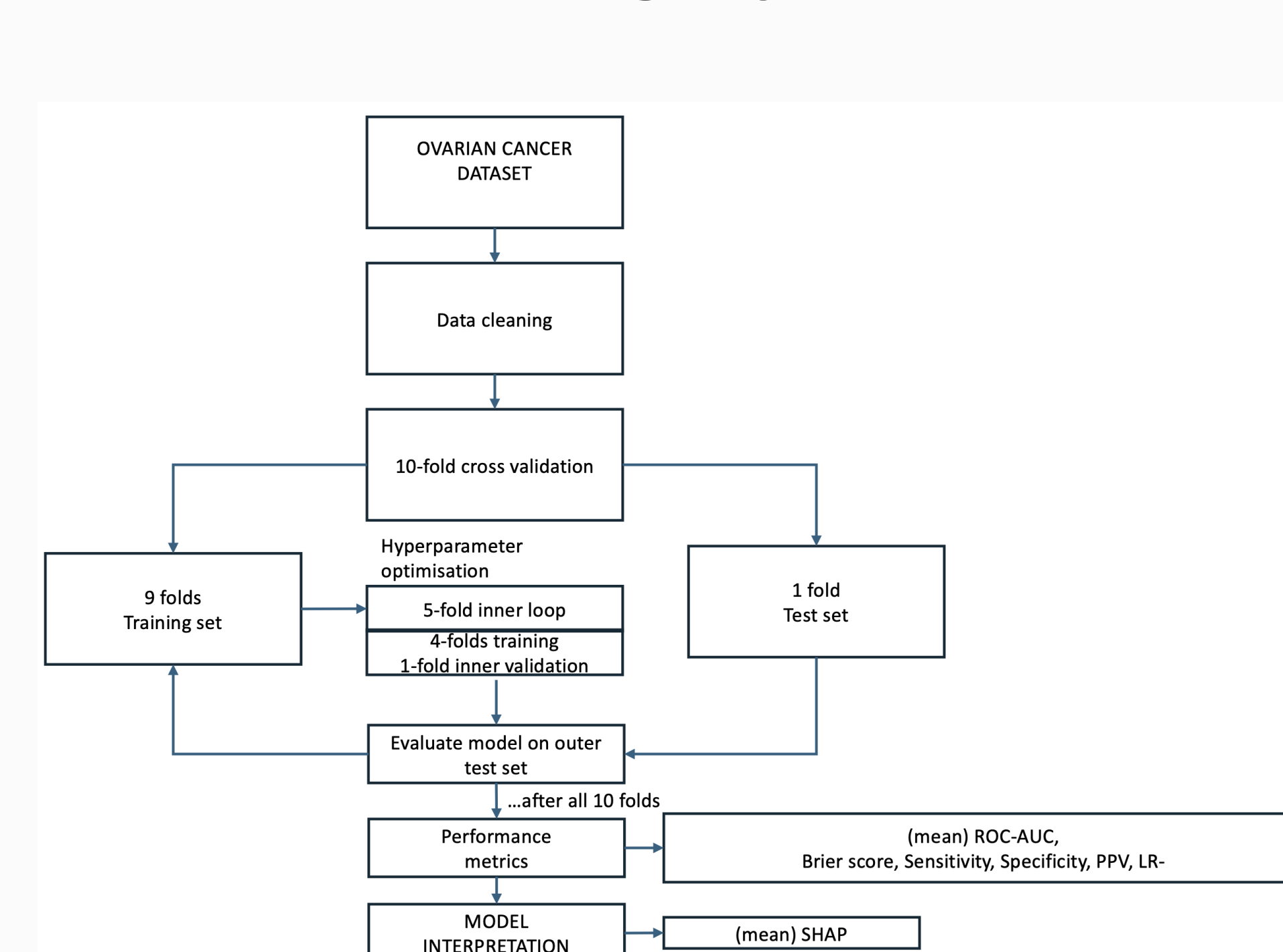
Data Flow



Study Design and Analysis

- Univariate symptom association with OC
- Internal nested cross-validation ML framework using RF, LR, XGB and SVM
- Performance evaluation: mean AUC, Brier score, sensitivity, specificity and PPV
- Threshold selection: 95% sensitivity
- Model explainability: SHAP

Machine Learning Pipeline

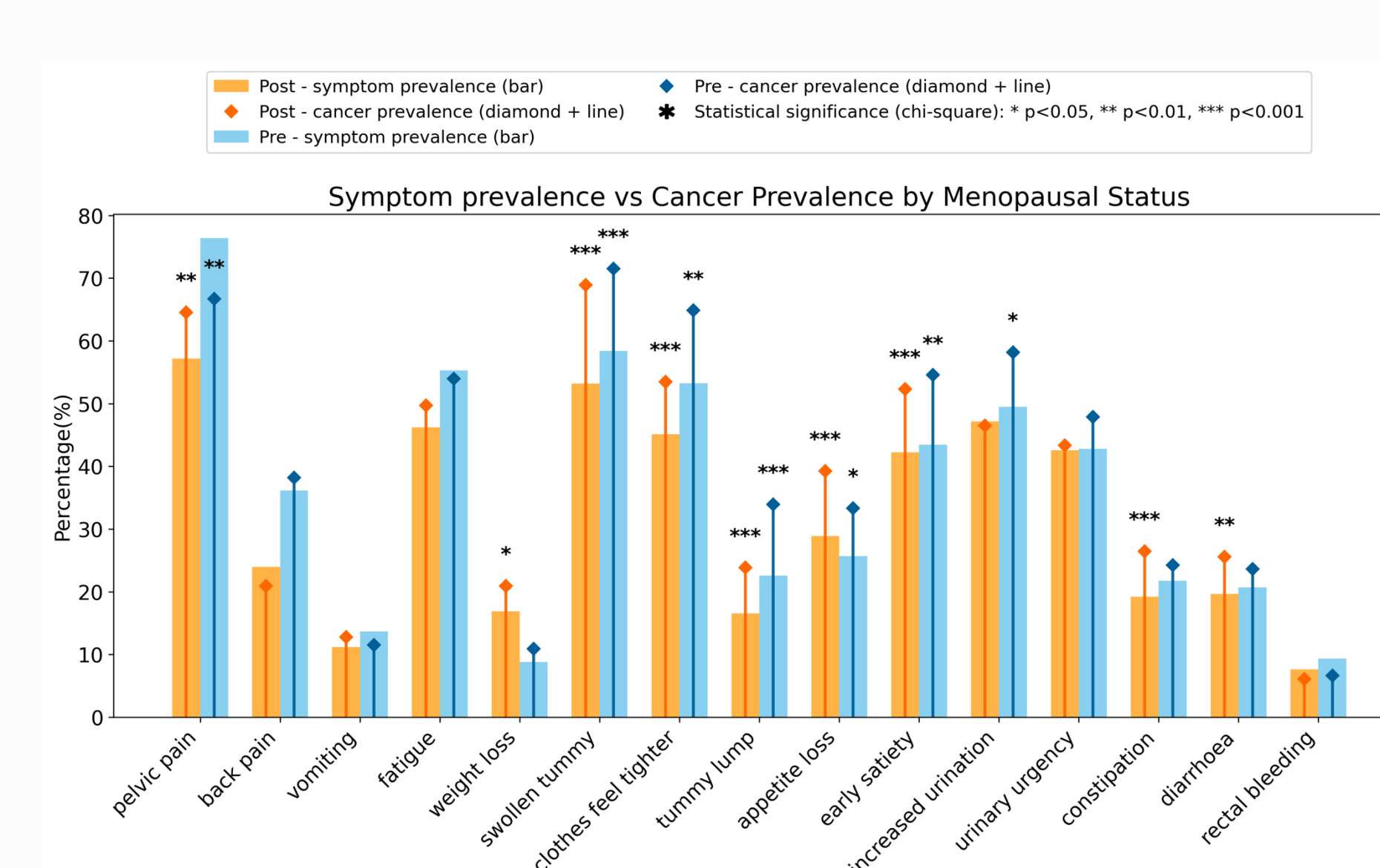


Key Findings

Study Population

	Train-pre	Train-post
Age(median years)	43 (34 – 48)	65 (58 – 73)
All malignancy (%)	14.6	29.6
Primary OC (%)	52.1	61.0
Early stage (FIGO I-II) (%)	60.5	42.9

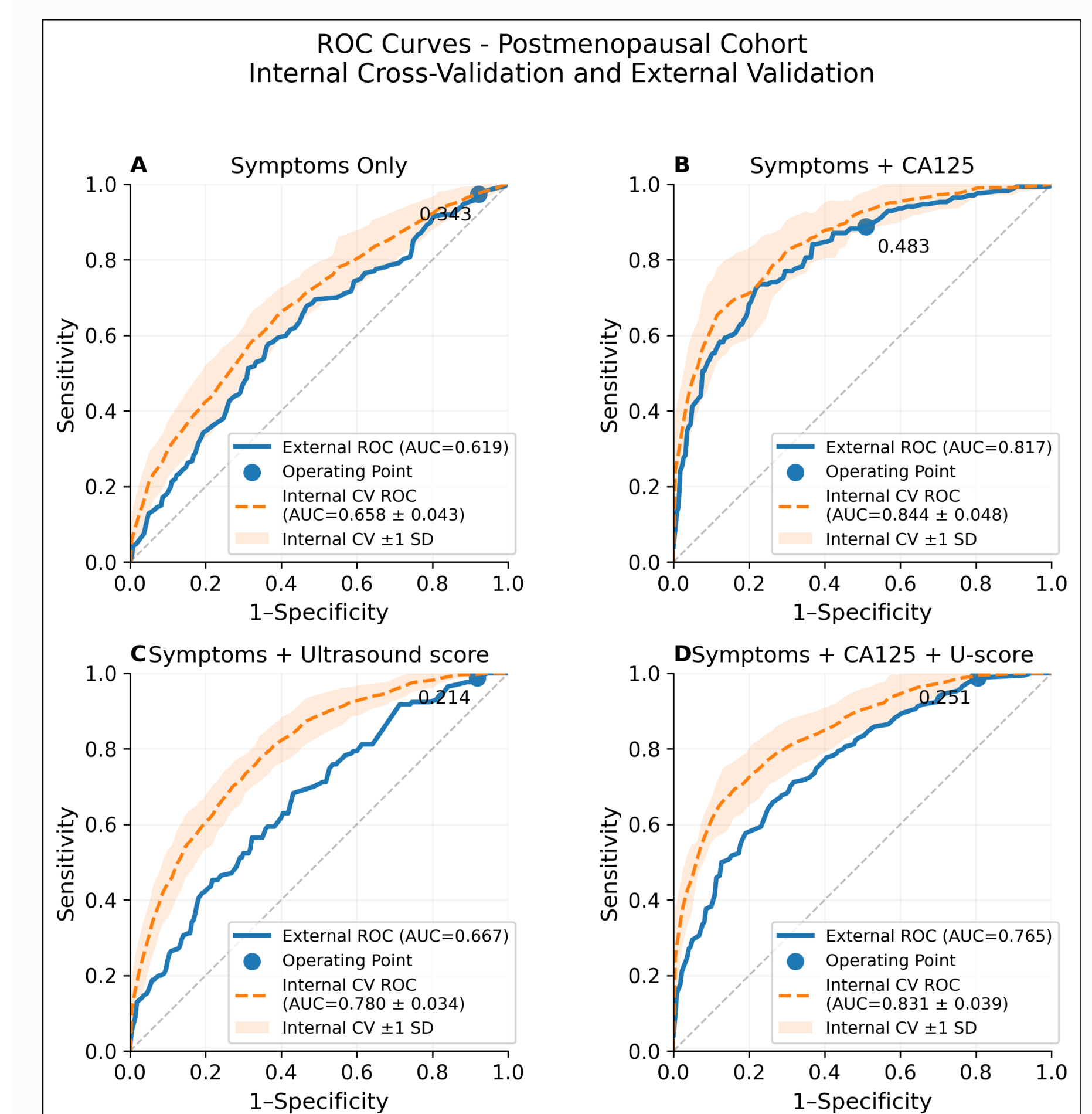
Symptom and Cancer Prevalence



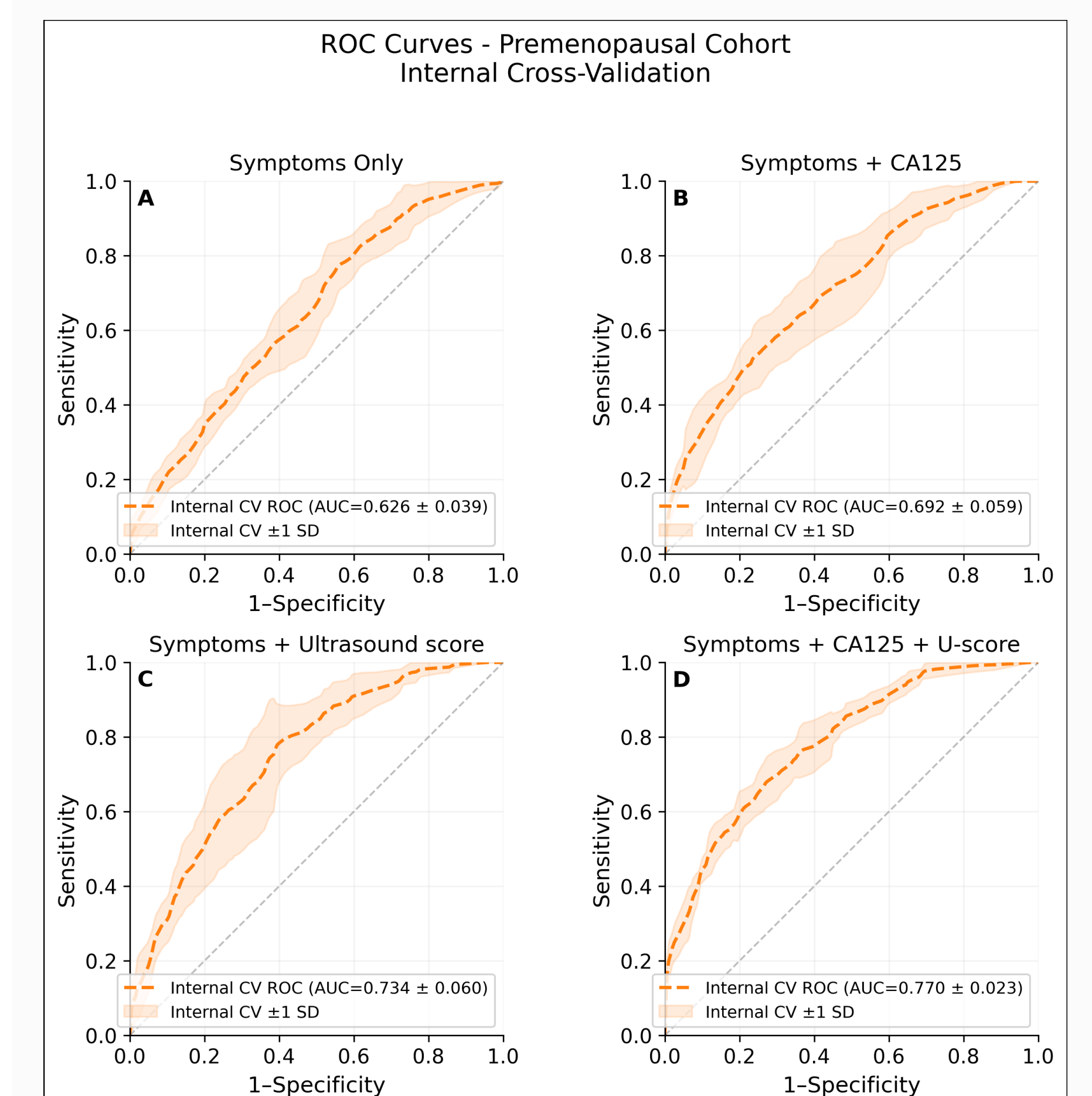
- Frequency of reported symptoms broadly similar. Pelvic pain and back pain more frequent in training premenopausal than postmenopausal patients
- Across both menopausal groups, pelvic pain and abdominal swelling showed high statistical significance to OC diagnosis
- Appetite loss, early satiety, constipation and diarrhoea had stronger association in training postmenopausal than premenopausal

Model Performance

Postmenopausal – Logistic Regression



Premenopausal – Logistic Regression



Explainability Analyses - SHAP

- Low feature importance for *weight loss* in both menopausal groups
- Abdominal swelling (swollen tummy)** most influential symptom in **postmenopausal** patients
- Pelvic pain** and **abdominal swelling** most significant in **premenopausal** cohort

Interpretation

- Better model performance than symptoms-only when diagnostic tests are included

Clinical Implications

- Symptoms alone are insufficient
- At the primary-secondary care referral interface, findings support diagnostic strategies that combine symptom assessment with CA125 in postmenopausal patients
- Symptom assessment with ultrasound tests in premenopausal patients.

¹ Davenport et al, Cochrane Database, 2022

² Sundar et al, Lancet Oncology, 2024

³ Sundar et al, BMJ, 2026