A Machine Learning Approach to Predict Endometriosis

Parand Shams, MS[1,2], Mayur Sarangdhar, PhD[2,3,4], Judith W. Dexheimer, PhD[2,3]

[1]Department of Biostatistics, Health Informatics and Data Science, University of Cincinnati, [2]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center; [3]Department of Pediatrics, College of Medicine, University of Cincinnati, [4]Division of Oncology, Cincinnati Children's Hospital Medical Center

Parand.Shams@cchmc.org

Cincinnati Children's changing the outcome together | University of CINCINNATI

This study was approved by the CCHMC IRB (IRB #2019-1288).

Background

- Endometriosis affects 190 million women during their reproductive years
- Non-specific symptoms
- Lesions trigger chronic inflammation, pelvic pain, and infertility
- Average diagnostic delay is 6 to 12 years (Figure 1)

Objective

- Automated identification could decrease diagnostic delay.
- To develop a non-invasive, machine learning-based screening model capable of distinguishing patients with endometriosis.

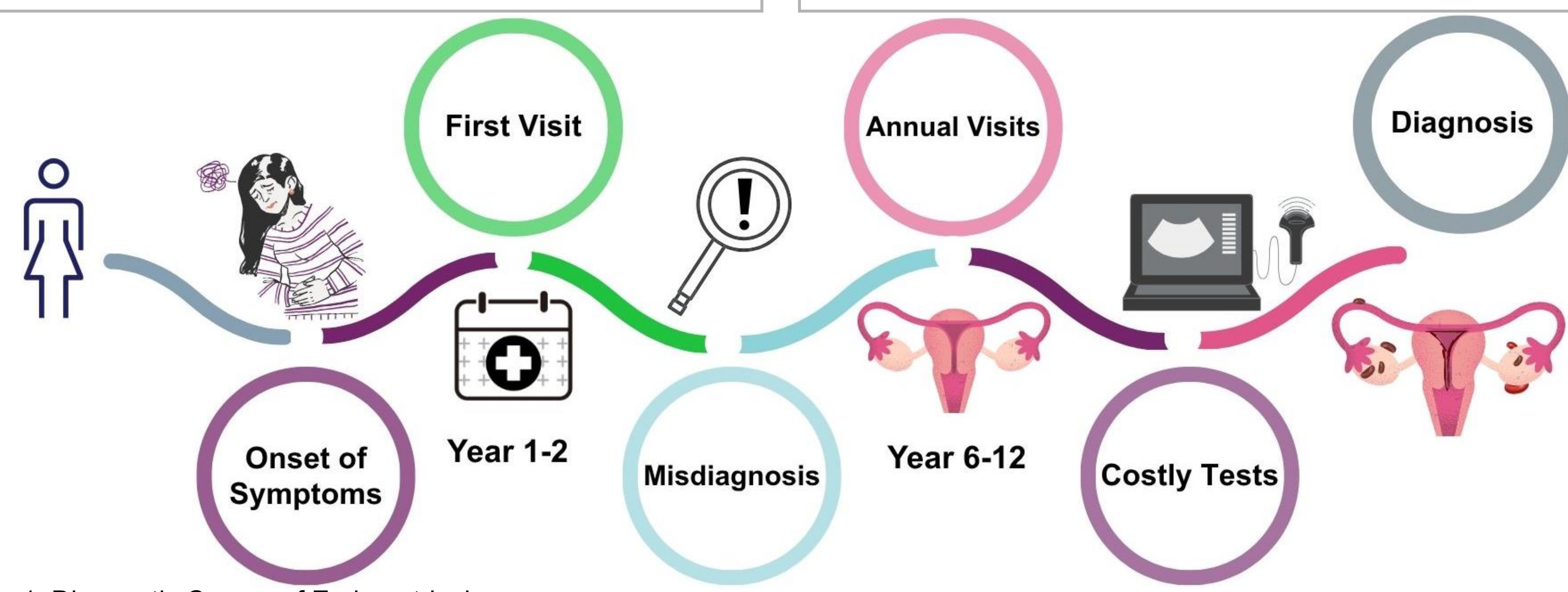

Figure 1: Diagnostic Course of Endometriosis

Data

- Adults 18 years and older
- Visit to OB/GYN clinic at the University of Cincinnati Medical Center (UCMC) between 2011 and 2019.
- De-identified electronic health records (EHR) collected
  - 380 unique number of patients
  - 155 patients have Endometriosis diagnosis
  - 13,266 total number of visits (office, telephone, hospital encounter)
- Patients were excluded if an erroneous encounter.

Methods

- Endometriosis identified by ICD9/ICD10 codes (N80)
- Demographics, visit data, diagnoses, comorbidities, and medications extracted.
- Features were one-hot encoded.
- Univariate feature selection applied to identify correlated features.
- Multiple supervised machine learning algorithms evaluated.
- Five-fold cross validation applied to assess performance.
- Populations compared using Chi-squared.
- Model performance evaluated using area under receiver operating curve (AUC) score.

Results

ICD10	ICD10 labels	P-value
R10	Abdominal pain	<0.01
N83	Noninflammatory disorders of ovary, fallopian tube and broad ligament	<0.01
G89	Pain not elsewhere classified	<0.01
Z30	Contraceptive management	<0.05
N85	*Other noninflammatory disorders of uterus, except cervix	<0.05
R19	Other symptoms and signs involving the digestive system and abdomen	<0.05
E66	*Overweight and obesity	<0.05
M79	*Other and unspecified soft tissue disorders, not elsewhere classified	<0.05
N94	Pain associated with female reproductive organs	<0.05
O09	*Supervision of high-risk pregnancy	<0.05

* Comorbidities were predominantly significant in non-endometriosis patients.

Results

Demographics		All Patients (n=380)	Patients with Endometriosis (n=155)	Patients w/o Endometriosis (n= 225)	P-value
Age (mean)		41.46	39.8	42.6	<0.05
Race					
	Caucasian	60%	60%	60%	
	African American	33%	34%	32%	
Ethnicity					
	Hispanic or Latino	4.5%	4%	2.1%	
	Not Hispanic or Latino	95.3%	94.8%	96%	

- Patients had an average age of 41.46 (min 20, max 80, 95% CI(40.38-42.53).
- Top performing models were naïve bayes (AUC=0.79), logistic regression (AUC=0.75), and random forest (AUC=0.71) shown in Figure 2.

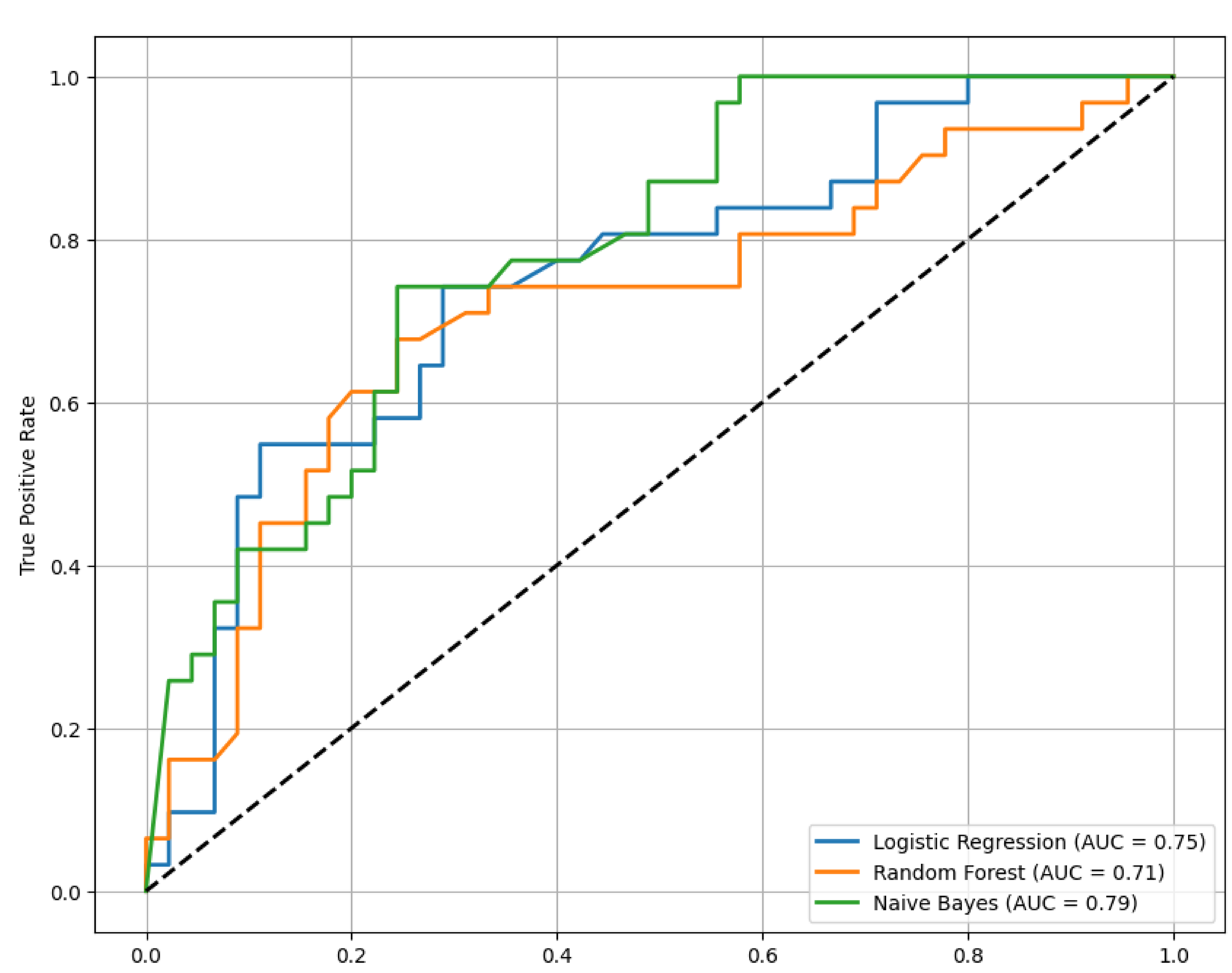

Figure 2: Area Under Receiver Operating Curve (AUC)

Limitations and Discussion

- Single site study; Limited cohort size
- Identified structured clinical features correlated with Endometriosis
- Incorporating longitudinal data to refine the model