

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Applying Machine Learning Algorithms to Predict Endometriosis Onset

Ewa J. Kleczyk, Tarachand Yadav and Stalin Amirtharaj

Abstract

Endometriosis is a commonly occurring progressive gynecological disorder, in which tissues similar to the lining of the uterus grow on other parts of the female body, including ovaries, fallopian tubes, and bowel. It is one of the primary causes of pelvic discomfort and fertility challenges in women. The actual cause of the endometriosis is still undetermined. As a result, the objective of the chapter is to identify the drivers of endometriosis' diagnoses via leveraging selected advanced machine learning (ML) algorithms. The primary risks of infertility and other health complications can be minimized to a greater extent if a likelihood of endometriosis could be predicted well in advance. Logistic regression (LR) and eXtreme Gradient Boosting (XGB) algorithms leveraged 36 months of medical history data to demonstrate the feasibility. Several direct and indirect features were identified as important to an accurate prediction of the condition onset, including selected diagnosis and procedure codes. Creating analytical tools based on the model results that could be integrated into the Electronic Health Records (EHR) systems and easily accessed by healthcare providers might aid the objective of improving the diagnostic processes and result in a timely and precise diagnosis, ultimately increasing patient care and quality of life.

Keywords: endometriosis, infertility, likelihood, logistic regression, machine learning, eXtreme gradient boosting, nomogram, odds ratio

1. Introduction

Recent advancements in artificial intelligence (AI) and machine learning (ML) have offered an opportunity for utilization of these advanced methodologies in the healthcare industry, while also at the same time improving upon the performance and accuracy benchmarks established by the classical statistical techniques [1]. A variety of ML techniques have been already applied to clinical data to examine a number of conditions and therapeutic areas, their onset, progression, and treatment options. In addition, deep learning algorithms such as convolutional neural network (CNN) have been employed in medical image data to predict disease onset and progression with even greater precision [2–5].

ML algorithms applied to a large amount of structured and unstructured data and combined with available data processing technology have already improved researchers' ability to mine the vast amount of data and assisted in making the patient healthcare decisions [6]. As a result of the high precision and robustness of

ML algorithms compared to the classical statistical methods, the insights derived from the application of these methods became important in driving the strategies and processes related to healthcare access, patient care, as well as disease diagnostics, healthcare trend forecasting, drug discovery, etc., thereby, further impacting the ability to reducing medical costs, shortening the time to diagnoses and treatment, and enhancing patients' quality of life and outcomes [7].

Endometriosis is one of the most commonly occurring disorders in women of menstruating age. Tissues, resembling the endometrium lining, grow on the outer part of the uterus and other organs of the pelvic area. The signs and symptoms differ across patients with some individuals experiencing mild symptoms, while others displaying moderate to severe signs. The most common symptoms of endometriosis include pain in the pelvic area, dysmenorrhea, and the inability to have children. Most commonly laparoscopy, surgery under general anesthesia, is performed to confirm the diagnosis of endometriosis [8]. Since it is an invasive procedure, it may not be suitable for all women. Laparoscopy is also quite expensive and women require a confirmation of a variety of indicatives of endometriosis before undergoing this procedure [9]. There are also a number of studies researching biomarkers of endometriosis via assessing endometrial tissue, uterine or menstrual fluids, immunological markers in blood or urine, gene expressions, etc. [10].

The availability of noninvasive methods to predict the likelihood of endometriosis could reduce the diagnostic delays and the number of women undergoing surgery unnecessarily, and thus avoiding unwanted complications and potential trauma [11]. In other research studies, researchers developed a new ensemble technique called GenomeForest that analyzed the gene expression data. The method systematically examined capabilities in classifying endometriosis and control samples, using both transcriptomics and methylomics data [12, 13].

Another research study developed symptom-based models that predicted the likelihood of endometriosis using logistic regression (LR). Symptomatic data including patient demographics, women's past medical history, obstetrics, family history, etc. were collected through a 25-item self-administered questionnaire [14]. Researchers also systematically applied selected ultrasound techniques in the diagnosis of endometriosis and concluded that these methods should remain the first-line procedures in the evaluation of patients with endometriosis [15].

In recent years, researchers aimed at developing CNN-based CAD systems that could classify endometrial lesions images obtained from hysteroscopy and evaluate the diagnostic performance of the model [16]. Their system slightly outperformed gynecologists in classifying endometrial lesion images. With a large number of diagnostic procedures, there is, however, no guaranteed treatment for endometriosis at this time. With an early diagnosis and available medical and surgical options; however, healthcare providers might be able to reduce the risks of potential complications and improve the quality of life for their patients [17, 18].

In the above research studies, researchers used either relatively small samples, or a limited number of variables to develop models or systems to predict the likelihood of endometriosis. The source of data represented mostly clinics and care providers in a controlled environment. There have been a limited amount of research studies performed thus far leveraging US-based patient-level claims data in predicting endometriosis. Claims data consist of the entire patient medical journey, such as diagnosis, procedures, prescriptions, physician, and patient demographics [19, 20]. In this chapter, US patient-level claims datasets at a transactional level were leveraged to develop accurate ML algorithms to predict the likelihood of endometriosis onset. Predicting the probability of endometriosis occurrence via leveraging the diagnosed patients' medical history might benefit both the diagnostics process as well as improved patients' quality of life. The LR and eXtreme Gradient Boosting

(XGB) algorithms were employed to identify the key drivers of endometriosis onset. An earlier version of this chapter is available on the Research Square website. The posting allowed for the dissemination of these important insights with the research community in advance, while at the same time, leveraging the received feedback to enhance the research design in this chapter.

2. Methodology overview

As mentioned earlier, the analysis design was described in the earlier version of the chapter available on the Research Square website. It leveraged the US healthcare claims patient-level database with the period from January 31, 2019 to December 31, 2019 [21]. Patients with a history of medical diagnosis ICD 10 codes for endometriosis were labeled as targets and the remaining patients were assigned as controls. As endometriosis is a women-only condition, female patients 18 and older were selected for the study target cohort. A control cohort, using a propensity matching algorithm, was built as a comparison group to the study targets. Thirty six (36) months of patients' medical history before the first condition event in 2019 were extracted for both cohorts. The US healthcare claims data included diagnosis, medical, procedural, surgical, and hospital codes, as well as medical treatments and therapies prescribed to patients. The dataset was presented at the transactional level to ensure proper capture of medical events longitudinally [21]. Several analytical approaches were employed for the analysis from the rules-based patient qualification criteria to ML algorithms to derive the probability of endometriosis onset. The healthcare claims patient-level dataset considered in the analysis represented healthcare claims sourced for the United States regions only.

2.1 Healthcare claims patient-level database

The US healthcare claims patient-level database is an anonymous longitudinal patient dataset often applied by healthcare organizations to derive insights [22, 23], while at the same time informing the effective treatment outcome options, patient access strategies, and areas for improvement in the diagnostic process [19]. The US healthcare claims patient-level database employed for this chapter consisted of medical, procedural, surgical, hospital, and prescriptions claims across all types of insurance payments and all geographic areas in the United States [24, 25]. The healthcare claims database overall covered more than 317 million active patients with over more than 17 years of medical health history and involved more than 1.9 million healthcare providers [25]. **Figure 1** presents the summary of information in the database.

2.2 Cohort selection

For this chapter, a sample of 314,101 confirmed endometriosis patients in 2019 in the US healthcare claims patient-level database was leveraged for the analysis. The patients were identified using predefined ICD 10 diagnosis codes (**Table 1**). Female patients of age 18 and older were identified for the target cohort. For the control cohort, a random sample of 3 million female patients with the same age specifications was selected from the database [21].

To define a control cohort of an equal size to the study target group, a 'propensity score matching' methodology was employed [18]. The algorithm selected the controls based on several similar characteristics or covariates. Covariates included patient age and medical history [26, 27]. **Table 2** presents the summary of the distribution comparison between the study target and control cohorts by age and

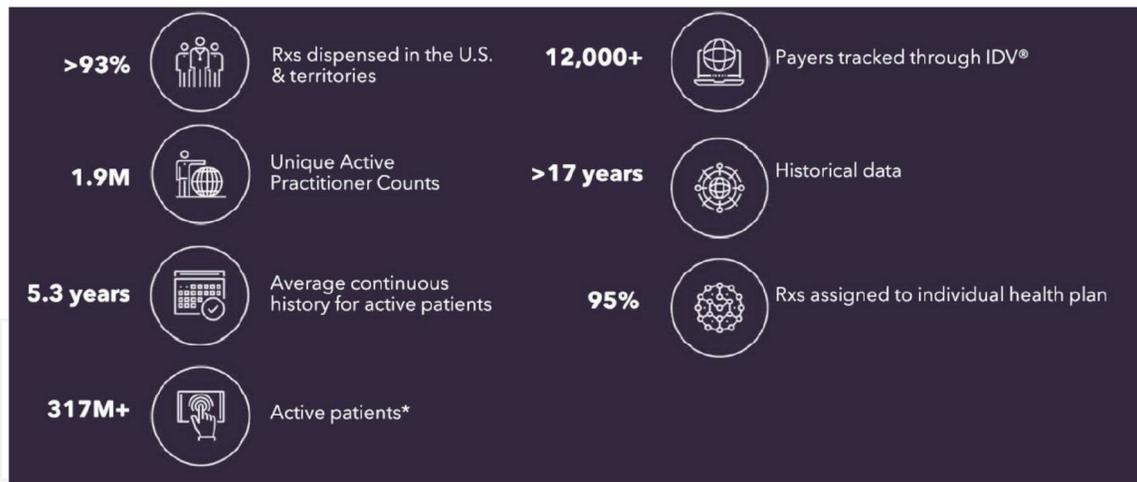


Figure 1.
Healthcare claims patient level database summary.

Diagnosis	Codes diagnosis long description
N80.0	Endometriosis of uterus
N80.1	Endometriosis of ovary
N80.2	Endometriosis of fallopian tube
N80.3	Endometriosis of pelvic peritoneum
N80.4	Endometriosis of rectovaginal septum and vagina
N80.5	Endometriosis of intestine
N80.6	Endometriosis in cutaneous scar
N80.8	Other endometriosis
N80.9	Endometriosis, unspecified

Table 1.
ICD 10 diagnosis codes of endometriosis.

Age group	Target (%)	Control (%)
18-24	6.45	6.55
25-34	25.01	25.24
35-44	37.57	37.08
45-54	23.13	23.18
55-64	6.22	6.31
65+	1.62	1.64
Region	Target (%)	Control (%)
South	39.90	39.90
Midwest	22.78	22.76
Northeast	18.82	18.84
West	17.02	17.02
Other	1.48	1.48

Table 2.
Comparison between target and control cohort by age and region respectively.

Census geographies. The patient age variable was created via grouping age ranges, while states were grouped into the US regions [21].

2.3 Data extraction

The next step in the analysis process was to pull the patients' medical history from the available information in the US healthcare claims patient-level database [21]. The event date for the target cohort was established for each individual in the study to ensure the extraction of the healthcare information before the first condition event. For the control cohort, the first activity in 2019 was leveraged as the event date [21].

The approach for the data extraction and the study target and control setup was the same as presented in the earlier version of the chapter available on Research Square. Using the medical event dates, representing the first date of endometriosis diagnosis, as the index date, 36 months of medical history was extracted for each patient. Historical data presented all available medical events in the patients' healthcare history before the condition diagnosis, including diagnoses for comorbid conditions, medical and surgical procedures, therapeutics, healthcare provider's specialty, and treatments prescribed to patients. A transactional level dataset, representing the top 1000 diagnosis codes, top 800 medical and surgical procedures, and top 500 prescribed drugs, was utilized to enable additional insights since these top codes constituted more than 80% of the dataset [21].

A pivot table was built at the transaction level and aggregated at the patient-level. Each row of the dataset represented an individual patient and the values within the row represented the counts of transactions that were generated during the patient's journey for the respective medical events. The columns of the table were the medical events, such as diagnosis and procedure codes, drugs prescribed, and physician specialties. The aggregated data table had more than 6 million rows and 2600 columns. The aggregated data table had missing values for selected patients and data elements, as not all records had complete medical information captured in the study period. Any medical events absent in the patient's history were represented with the value of *zero* (0), which implied that no such event was observed in the individual's medical history. The final aggregated dataset was leveraged as an analytical dataset for the remaining parts of the chapter [21].

The analytical dataset was further normalized and divided into two groups: a training and test set. A ratio of 70:30 was applied to the dataset [28]. The training dataset was employed to identify the key data elements driving endometriosis diagnoses, while the test group was used to confirm whether these elements would predict the condition occurrence accurately [29]. Splitting the data into training and test sets aided the assessment of the model performance and its ability to generalize the hidden data trends [21, 30].

2.4 Overview of machine learning algorithms

In this section of the chapter, a summary of the classical statistical modeling and ML approaches is presented to review the available methods for healthcare research, and also to summarize the selected methodology applied in this study. Statistical modeling has evolved in the last few decades and shaped the future of business analytics and data science, including the current use and applications of ML algorithms [31]. It represents a branch of applied mathematics, in which statistical methods are leveraged to analyze a dataset. Statistical models are the mathematical representation of real-world scenarios with certain assumptions undertaken. They play a fundamental role in making statistical inferences while studying the characteristics of a population, upon which hypotheses were framed [8]. These models

are not only useful in finding relationships between variables and the significance of those relationships, but they are also useful in the prediction and forecasting of future events.

ML is a subfield of the AI area, which includes statistics, mathematics, computer algorithms, etc., focused on building applications that learn and improve their predictive capabilities automatically over time without being specifically programmed to do so. ML models are built upon a statistical framework since they involve a large amount of data elements often described using statistical distributions. In the last two decades, ML algorithms have received a significant amount of attention in the fields of computer vision, natural language processing, autonomous driving vehicles, healthcare and drug development, e-commerce, to list a few due to the increased amounts of data availability and significant advancements in the computing power. ML algorithms can be broadly categorized as supervised, unsupervised, and semi-supervised algorithms [5, 7, 32, 33].

2.4.1 Supervised learning algorithms

Supervised learning is a set of algorithms that learn from the input space (X) to the output space (Y), i.e. $Y = f(X)$ [34]. The major objective is to estimate the mapping function (f) to ensure that with an addition of a new data point (x), the outcome, (y), could be predicted [35]. Supervised learning algorithms are often applied to classification and prediction problems [32]. The following are the selected examples of supervised algorithms often employed in research studies: *logistic regression, decision trees (DTs), random forest (RF), extreme gradient boosting, support vector machines (SVMs), Naïve Bayes, adaptive boosting (AdaBoost), artificial neural network (ANN), etc.* [36].

2.4.2 Unsupervised learning algorithms

Different from the supervised learning algorithms, the unsupervised learning algorithms try to understand the hidden patterns within the input dataset (X) [37]. The algorithms learn and uncover the patterns without the researcher's assistance [38]. These algorithms are often leveraged to find the naturally occurring clusters, reduce data dimensions, detect anomalies, etc. *k-means clustering, principal component analysis (PCA), factor analysis (FA), singular value decomposition (SVD), apriori algorithm (association rule)* represent a few examples of these types of algorithms [36]. In some cases, a semi-supervised approach is used to enhance the model performance with the help of a small amount of labeled data [36].

Depending on the study objectives and the availability and granularity of data, algorithms are reviewed for analytical relevance, tested for performance, data type fit, and selected as optimal algorithms accordingly. For this chapter, LR and XGB models were chosen to develop a predictive algorithm for the endometriosis onset. LR estimated the odds of the condition occurrence for a given medical event [39], while XGB provided more flexibility in fine-tuning the hyper-parameters when compared to other tree-based algorithms [40].

2.4.3 Logistic regression

An LR is a statistical model as well as the simplest version of ML algorithms that uses a logistic function to model a binary dependent variable with two possible outcomes: '0' and '1' [39, 41, 42]. A *multinomial logistic regression* is also often considered for research studies with multiple outcomes. LR is applied in a variety of fields, including healthcare research and social sciences [43].

In regression modeling, analysis often involves interpreting the independent variables' coefficients. Regression coefficients describe the size and direction of the relationship between regressors (x) and the outcome variable (y). They explain the behavior of the dependent variable given a unit change in an independent variable while holding all other data elements constant. The magnitude and sign of the coefficients signify the resulting relationship with the dependent variable. Interpreting the LR's coefficients also includes their interpretation, as well as the odds and odds ratios [41].

Odds exemplify the ratio of probabilities of two mutually exclusive events [41], at the same time the *odds ratio* represents the ratio of two different odds. The simplest way to calculate the *odds ratio* in the LR is to exponentiate the coefficient of a predictor [39]. As a result, if the *odds ratio* for the age variable in years is 1.25, then for each additional year, the probability of event/success increases by 25%. For categorical features, the interpretation of the *odds ratio* can be more meaningful than the interpretation of *odds* [41].

2.4.4 *x*Extreme gradient boosting

A gradient boosting is another ML algorithm, which is an ensemble of simple, weak, and unreliable predictors, mainly decision trees [40]. When multiple trees are grouped, they create a robust and reliable algorithm [44]. XGB starts by creating a first simple tree [45] and builds upon the weaker learners. Each iteration revises the previous tree until an optimal point is reached [46].

Feature importance is the value generated by tree-based models, including *decision trees*, *random forest*, *XGB*, etc. [40]. The measure signifies the importance of features in the model as well as how good the feature is at reducing the node impurity. Feature importance is also known as '*gini importance*' or '*mean decrease impurity*,' and is defined as the total decrease in node impurity averaged over trees in the ensemble [44]. It is calculated as: *weight*, *gain*, and *cover*, where '*weight*' represents the number of times a feature is observed in a tree, '*gain*' denotes the average gain of splits, and '*cover*' is defined as the average coverage of splits. Finally, *coverage* represents the number of samples impacted by the split [46].

2.4.5 Chi-Square test

The Chi-Square test is nonparametric [33], often employed to test the independence between the observed and expected frequencies of one or more data elements. It is known as the '*goodness of fit test*' [47]. In this chapter, the Chi-Square test was utilized to select the top significant features [48].

2.4.6 *p*-value

The *p*-value is the probability of an observed result, assuming that the null hypothesis is correct. The *p*-value is used to test if the null hypothesis can be rejected in favor of the alternative hypothesis. A lower *p*-value implies a stronger indication in support of the alternative hypothesis [23]. In this analysis, the significance level was set at 5% to aid the feature importance evaluation and statistical results' identification.

2.4.7 Classification metrics

The following classification metrics are often leveraged to validate the ML models' performance. A confusion matrix is generated from the predicted probability

values with 0.5 as the classification threshold. Patients with probability values greater than or equal to 0.5 are classified as 1 and below 0.5 are classified as 0. Below is the list of metrics used in evaluating models performance [32, 43, 46, 49]:

Confusion matrix:

- *True positive (TP)*—Target patient correctly identified by the model as target patient
- *False positive (FP)*—Control patient misclassified by the model as target patient
- *True negative (TN)*—Control patient correctly classified by the model as a control patient
- *False negative (FN)*—Target patient misclassified by the model as a control patient

Model performance metrics:

- *Accuracy*: % of total patients correctly identified among total patients
- *Positive predictive value (PPV, Precision)*: % of true target patients among total predicted target patients
- *True positive rate (TPR, Sensitivity, Recall, Hit Rate)*: % of true target patients who were correctly identified among total target patients
- *False positive rate (FPR)*: % of true control patients incorrectly identified among total control patients
- *Specificity*: % of those control who will have a negative target result
- *F1 score*: is the harmonic mean of precision and recall
- *AUC*: Area under the receiver operating characteristic (ROC) curve. To validate the trade-off between true positive rate and false-positive rate

In this chapter, the LR, being the simplest of all ML algorithms, was chosen as the base model. Both the LR and XGB models were trained on the analytical dataset defined in the earlier section of this chapter. The top 1000 features from each algorithm were selected to reduce the dataset dimension. As the next step, the Chi-Square test from the *scikit-learn* Python package was utilized to identify the top most significant features from the list of data elements employed in both models. Finally, algorithms were re-trained on the top significant features to identify the key data elements in predicting the endometriosis onset. All ML algorithms were trained on Python 3.5 using '*scikit-learn*' and '*xgboost*' libraries.

3. Results

3.1 Important features selection

Table 3 presents the ML model performance metrics of the initial run, where the objective was to select the top features and study whether the data captured was

Algorithms	Statistic	Train set	Test set
LR	Accuracy	96%	96%
	Sensitivity/TPR/recall	95%	95%
	Specificity/TNR	98%	97%
	Precision/PPV	98%	97%
	f1-Score	0.96	0.96
	AUC	0.96	0.96
XGB	Accuracy	90%	88%
	Sensitivity/TPR/recall	86%	84%
	Specificity/TNR	95%	93%
	Precision/PPV	95%	92%
	f1-Score	0.9	0.88
	AUC	0.9	0.88

Table 3.
 Classification metrics of train and test sets for LR and XGB model.

reasonably proven in disease prediction. Algorithms were trained on 70% of the analytical dataset and were tested on the remaining 30%. Metrics captured indicated that both the LR and XGB models performed relatively well in predicting the condition onset. The models' accuracy ranged between 88% and 96%. **Figure 2** presents the ROC curves on the test set for LR and XGB models respectively. The area under the ROC curve (AUC) values were 0.88 and 0.96, respectively for both models.

From the outputs of the initial model run, the top 1000 features with absolute regressor coefficients in descending order greater than zero (0) were selected from the LR. Similarly, another set of top 1000 features with feature importance greater than zero (0) were identified from XGB. Both sets were combined to establish a unique list of top features. As the next step, the Chi-Square test for feature selection from Python *scikit-learn* package was applied to select the top 1000 most significant features for the final model run. The top features were selected at a standard significance level of 5% ($\alpha = 0.05$). Most of the top significant features were associated with a series of medical and surgical procedures, as well as various diagnostic and comorbid conditions.

As noted above, **Table 4** presents the list of most significant features identified by the Chi-Square test, which were associated with the endometriosis diagnosis.

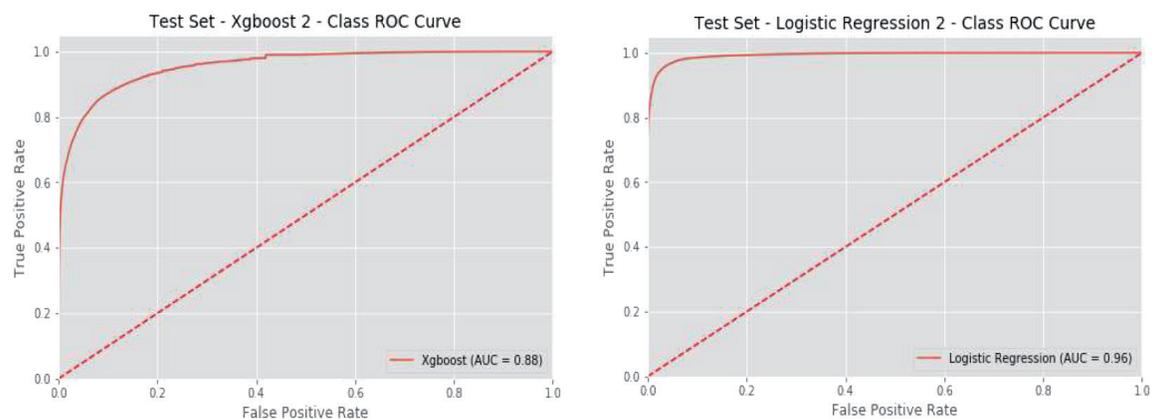


Figure 2.
 XGB & LR ROC curves on test set.

Feature	Feature description	Chi-square	LR: feature coefficients
D_N85_8	Other specified non-inflammatory disorder of uterus	0	3.48
D_N94_6	Dysmenorrhea, unspecifie	0	0.17
D_N94_9	Unspecified condition associated with female genital organs and menstrual cycle	0	6.9
D_R10_2	Pelvic and perineal pain	0	-0.04
D_Z01_419	Encounter for gynecological examination (general) (routine) without abnormal findings	0	-1.95
P_00840	Anesthesia intraperitoneal lower abd w/laps nos	0	1.54
P_00944	Anesthesia vaginal hysterectomy incl biopsy	0	1.55
P_52000	Cystourethroscopy	0	5.78
P_58571	Laps total hysterect 250 gm/<w/rmvl tube/ovary	0	3.25
P_58573	Laparoscopy tot hysterectomy >250 g w/tube/ovar	0	5.31
P_58662	Laps fulg/exc ovary viscera/ peritoneal surface	0	4.17
P_76830	Us transvaginal	0	1.93
P_J1950	Injection. Leuprolide acetate (for depot suspens)	0	3.74
R_Norethindrone_Acetate	Norethindrone acetate	0	0.26
SPCLT_EM	Emergency medicine	0	-9.47
SPCLT_FM	Family medicine	0	-3.63
SPCLT_HO	Hematology/oncology	0	-4.6
SPCLT_OBG	Obstetrics and gynecology	0	-2.43

Table 4.
Most significant features from LR, XGB, and Chi-Square test.

The table also presents the LR coefficients to provide relative direction between the endometriosis onset and the selected top regressors. As noted in the earlier version of the chapter available on Research Square, data elements including ‘non-inflammatory disorder of uterus,’ ‘pelvic and perineal pain’ presented examples of the diagnosis codes, indicated a positive relationship with symptoms of endometriosis [21, 50]. Procedure codes such as ‘anesthesia of lower abdomen for laparoscopy,’ ‘vaginal hysterectomy including biopsy’ were also identified as the procedures often correlated with the diagnosis as well treatment of endometriosis [50]. Furthermore, the Chi-Square test suggested that patients often consulted with a variety of healthcare specialists, including ‘emergency medicine (SPCLT_EM),’ ‘family medicine (SPCLT_FM),’ ‘obstetrics and gynecology (SPCLT_OBG)’ when experiencing gynecological symptoms and concerns; however, a larger number of office visits might negatively impact the likelihood for the condition diagnosis, as noted by the negative regressor coefficients.

3.2 Feature selection for the cohort selection

The significant features from Section 3.1, which were specific to the target cohort, seemed promising in defining the drivers of the endometriosis condition onset, and hence, were selected to identify the patient base list for scoring. Therapeutics as well as medical and surgical procedure codes specific to endometriosis treatment such as Orilissa, Marilissa, and Lupron Depot, were excluded from the analysis to avoid introducing any biases into the next phase of the study. Around 9.5 million female patients age 18 and above qualified for the scoring process.

3.3 Machine learning model training and outcome validation

The LR and XGB models were re-trained, using the top significant features. A drop in the model performance at the beginning of the re-training process was observed. After several iterations and hyper-parameter tuning, the predictive power of the XGB model significantly improved compared to the previous iterations; however, no improvement in the LR model performance metrics was observed. Interestingly, both models were able to identify additional new features aligned with endometriosis.

Table 5 presents the top features identified by the XGB and LR models to be important in predicting the likelihood of endometriosis along with the statistical measures and metrics to assess the importance and significance of the features. The Chi-Square test (p -value) signified the importance of data elements in differentiating the target and control patients. The XGB feature importance weighed the value of features in the model in predicting the outcome. Similarly, the LR odds ratios helped to understand the odds of being diagnosed with endometriosis, given a particular medical event.

Overall, results suggest that features including ‘other ovarian cyst, right side,’ ‘hypertrophy of uterus,’ ‘submucous leiomyoma of uterus,’ ‘excessive bleeding in the premenopausal period,’ ‘unspecified condition associated with female genital organs,’ and ‘menstrual cycle’ were important in predicting the likelihood of endometriosis. The models had also flagged ‘acetaminophen’ and ‘megestrol acetate’ drugs as strong predictors of the condition.

Table 6 shows that the XGB model performed better overall compared to the LR model. **Figure 3** shows the receiver operating characteristic (ROC) curves on the test sets for both re-trained models. The area under the ROC curve (AUC) values of the LR and XGB models were 0.87 and 0.96, respectively. Furthermore, **Figure 4** suggests that the XGB model was able to differentiate more accurately the targets from the controls than the LR model; hence, based on the final model results, the XGB model was utilized to score the qualified patients.

3.4 Scoring qualified patients

The last step of the model evaluation was to score the qualified patients to assess the model’s accuracy in predicting the endometriosis onset. A sample of 9.5 million patients was identified and complete medical history was extracted for 36 months. After dataset preparation, the probability of endometriosis was estimated, leveraging the re-trained XGB model.

Probability distribution of 9.5 million scored patients is shown in **Figure 5**. Most of the predicted probability values were concentrated either toward ‘0’ or ‘1’. When considering 0.5 as a threshold, the XGB model identified around 36% of the scored patients as being likely to receive an endometriosis diagnosis within the next

Feature	Long description	Chi-square (p)	XGB_feature_importance	LR_beta_coeff	Odds_ratio
P_58662	Laps fulg/exc ovary viscera/peritoneal surface	0	0.0318	4.70	109.73
P_58571	Laps total hysterect 250 gm/< w/rmvl tube/ovary	0	0.0212	4.17	64.53
D_N85_8	Other specified noninflammatory disorders of uterus	0	0.0094	2.56	12.88
D_N83_291	Other ovarian cyst, right side	0	0.0092	2.84	17.06
P_58661	Laparoscopy w/rmvl adnexal structures	0	0.0089	2.43	11.32
D_N85_2	Hypertrophy of uterus	0	0.0088	2.67	14.42
P_00944	Anesthesia vaginal hysterectomy incl biopsy	0	0.0076	1.77	5.86
P_52000	Cystourethroscopy	0	0.0075	1.62	5.04
D_D25_2	Subserosal leiomyoma of uterus	0	0.0069	2.25	9.53
P_72197	mri pelvis w/o & w/contrast material	0	0.0067	2.72	15.17
R_ACETAMINOPHEN	Acetaminophen	0	0.0066	2.01	7.46
D_N81_4	Uterovaginal prolapse, unspecified	0	0.0063	1.86	6.40
D_N94_9	Unspecified condition associated with female genital organs and menstrual cycle	0	0.0063	2.57	13.10
D_N92_4	Excessive bleeding in the premenopausal period	0	0.0061	2.30	9.99
D_D25_0	Submucous leiomyoma of uterus	0	0.0059	2.46	11.76
D_R10_2	Pelvic and perineal pain	0	0.0056	0.60	1.83
D_N94_5	Secondary dysmenorrheal	0	0.0056	2.81	16.64
D_Z79_890	Hormone replacement therapy	0	0.0047	2.23	9.34
D_Z80_41	Family history of malignant neoplasm of ovary	0	0.0045	2.12	8.37
D_N94_3	Premenstrual tension syndrome	0	0.0042	2.43	11.37
R_LIDOCAINE_HCL	Lidocaine hcl	0	0.0041	2.12	8.30

Feature	Long description	Chi-square (p)	XGB_feature_importance	LR_beta_coeff	Odds_ratio
R_MEGESTROL_ACETATE	Megestrol acetate	0	0.0039	2.19	8.94
D_F43_0	Acute stress reaction	0	0.0032	2.36	10.61
D_N94_12	Deep dyspareunia	0	0.0023	2.35	10.51
D_N97_0	Female infertility associated with anovulation	0	0.0022	2.19	8.89
SPCLT_AN	Anesthesiology	0	0.0012	(0.55)	0.58
SPCLT_DR	Diagnostic radiology	0	0.0009	(0.87)	0.42
SPCLT_OBG	Obstetrics and gynecology	0	0.0008	(0.64)	0.53
SPCLT_EM	Emergency medicine	0	0.0006	(1.92)	0.15
SPCLT_FM	Family medicine	0	0.0004	(1.05)	0.35
SPCLT_IM	Internal medicine	0	0.0004	(0.92)	0.40
SPCLT_HO	Hematology/oncology	0	0.0003	(0.79)	0.45

Table 5.
List of top features identified by the re-trained models.

Algorithms	Statistic	Train set	Test set
LR	Accuracy	87%	87%
	Sensitivity/TPR/recall	75%	75%
	Specificity/TNR	98%	98%
	Precision/PPV	98%	98%
	f1-score	0.85	0.85
	AUC	0.87	0.87
XGB	Accuracy	96%	94%
	Sensitivity/TPR/recall	93%	90%
	Specificity/TNR	99%	98%
	Precision/PPV	99%	97%
	f1-score	0.96	0.93
	AUC	0.96	0.94

Table 6.
Classification metric of LR and XGB model on train and test set.

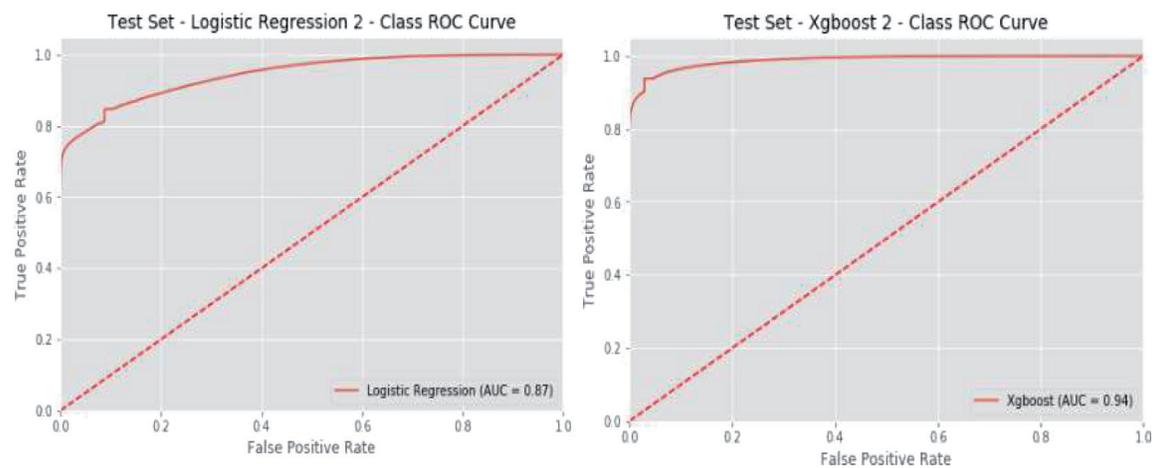


Figure 3.
ROC curves of LR and XG models on test set.

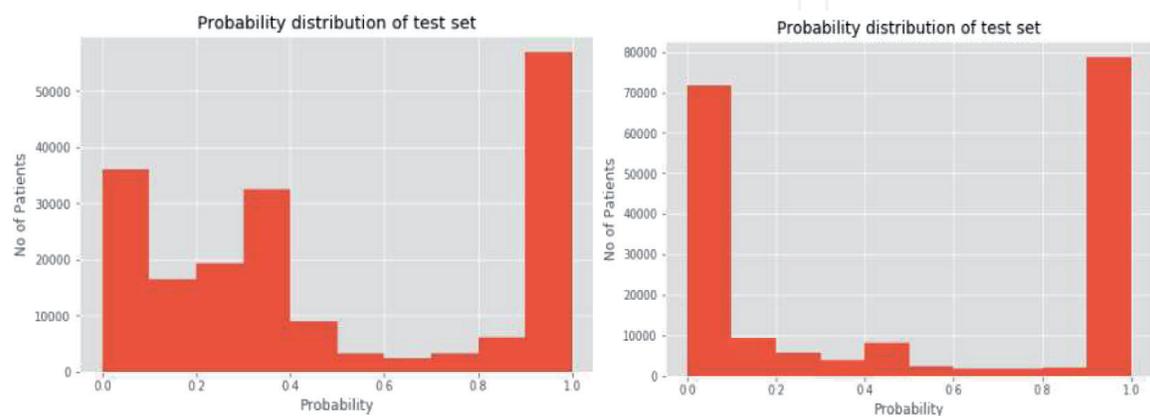


Figure 4.
Distribution of probability on test data set for both the LR and XGB models. Figure on right side is of XGB and most of scores are grouped at extreme values.

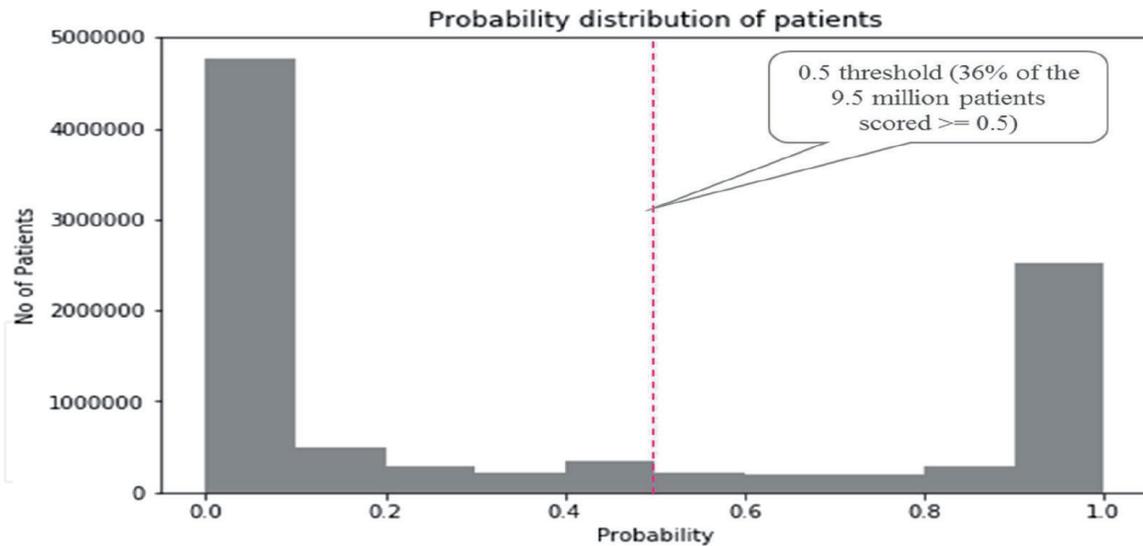


Figure 5.
 Distribution of patients by predicted probability score.

12 months. Assuming an ability to leverage the significant variables in diagnosing the condition onset, practitioners could provide focused and specialized medical care in time to their patients, thereby, reducing the risks of endometriosis and its related complications.

There is also a different way to present the data elements driving the prediction of disease onset and the scoring of patients for the likelihood of the disease. A nomogram (otherwise known as nomograph) is defined as an alignment chart or a two-dimensional diagram applied to estimate the graphical computation of a mathematical function [51]. A nomogram comprises a set of scales, where each scale denotes a selected feature of the studied population.

The nomogram tool is often employed in clinical medicine to predict patients' outcomes when considering their clinical features [52]. It is also used in clinical oncology to aid healthcare providers in their treatment decisions. It leverages regression models

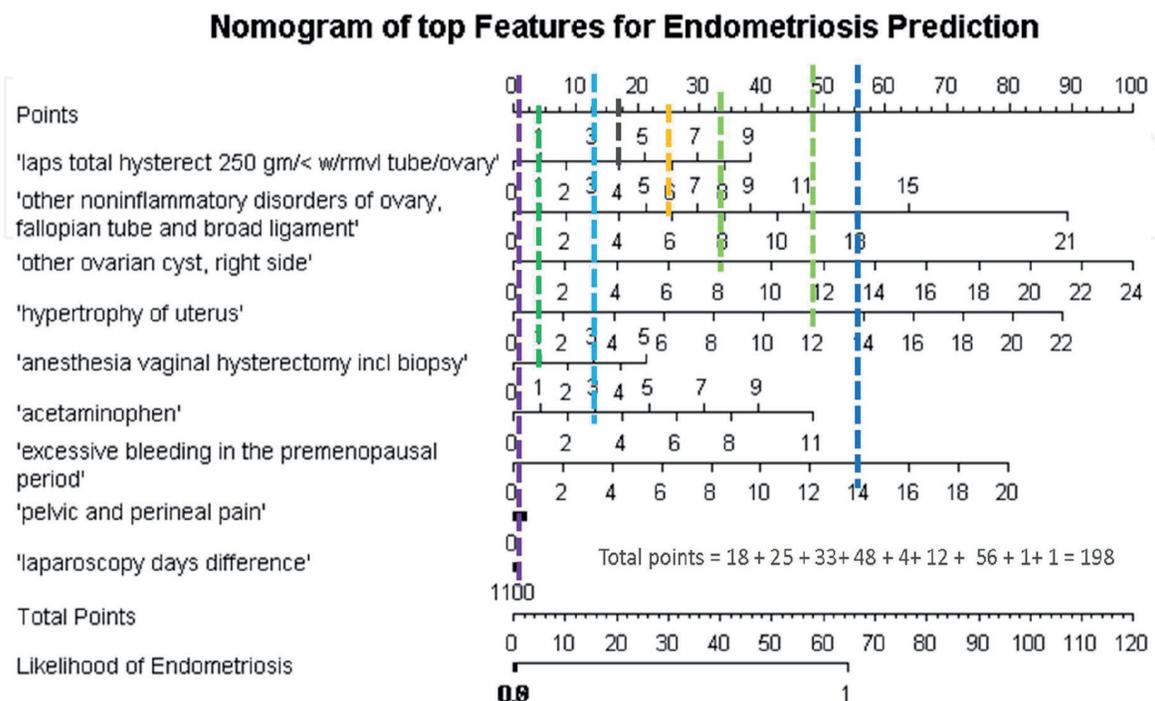


Figure 6.
 Nomogram of top features to predict likelihood of endometriosis.

such as the LR and parametric survival model as the basis for its framework [53]. For this chapter, a nomogram was selected to present a selected group of top features important to predicting the likelihood of endometriosis, as shown in **Figure 6**. The following attributes were noted on the chart as important in driving the diagnosis: 'laps total hysterect 250 gm/< w/rmvl tube/ovary,' 'other noninflammatory disorders of ovary, fallopian tube, and broad ligament,' 'other ovarian cyst, right side,' 'hypertrophy of uterus,' 'acetaminophen,' and 'pelvic and perineal pain.'

To predict the disease onset, the contribution of each feature was measured as a point score (topmost axis in the nomogram) based on the values that each feature could take with individual point scores being added to determine the likelihood of endometriosis onset. When the value of the feature was '0', its contribution was '0' points. The dotted line depicted the point score for an individual value of each respective feature with the total point being 198, which implied a very high probability of the disease onset. Nomogram was found to be a helpful tool to graphically study the outcomes given a group of few features; however, it was also challenging to leverage it, knowing a large number of studied features [52, 53].

4. Discussion

As mentioned in Section 3, the LR and XGB ML models were able to identify the top features that could help to explain endometriosis onset in advance. **Tables 4** and **5** present the important features to predict the condition onset. These features included diagnosis codes, medical and surgical procedure codes, as well as physician specialties that often support patients through their healthcare journey.

Furthermore, **Table 5** also presents the *LR odds ratio* and XGB feature importance index to aid the understanding and interpretation of the results. As noted in the above section, *odds ratios* defined the odds of being diagnosed with endometriosis when the feature changes by a unit, holding other features constant. For example, the *odds ratio* of 'uterovaginal prolapse, unspecified' was 6.40, which implied that for every additional diagnosis of 'uterovaginal prolapse, unspecified', the odds of endometriosis went up by 540%. Similarly, if a patient had an additional appointment with an 'obstetrics and gynecology' specialist then the odds decreased by 47%.

As a reminder, the first part of the ML analysis was to identify the top features from an extensive list of data elements (**Table 4**). LR, XGB, and Chi-Square tests were employed to derive the final list of features to re-train the model. **Table 5** presents the most promising features with their respective significance and importance values. A number of the variables from the model were also cited in other medical and scientific journal publications, including articles from Johns Hopkins Medicine [17] and Queensland Health [18] on endometriosis signs, symptoms, and diagnosis, which confirmed the model's validity from the medical and clinical side.

In the next part of this section, the selected most important features by their respective groups were reviewed and evaluated for their relevance to the endometriosis diagnostic process. The preliminary insights for this research are available on the Research Square website. The advanced preview allowed for valuable feedback that helped to enhance the research design for this chapter.

1. Diagnoses codes: 'other ovarian cyst, right side,' 'unspecified condition associated with female genital organs and menstrual cycle,' 'other specified noninflammatory disorders of the uterus,' 'excessive bleeding in the premenopausal period,' 'female pelvic peritoneal adhesions (post-infective),' 'uterovaginal prolapse, unspecified,' etc. clearly showed association with the risks and symptoms

of endometriosis [54]. Feature importance from XGB suggested that these features drove the model, whereas *odds ratio* from LR also indicated the direction of increase or decrease in odds of getting diagnosed with the condition. To further define the magnitude of importance, **Table 5** presents that if a patient was diagnosed with ‘excessive bleeding in the premenopausal period’ then the odds of receiving endometriosis diagnosis in the near future increased by 899%. Similar to these findings, Mayo Clinic articles also stated that patients might experience occasional heavy bleeding before being diagnosed with the condition [55].

2. Medical and surgical procedures: ‘laps fulg/exc ovary viscera/peritoneal surface’, ‘laps total hysterect 250 gm/< w/rmvl tube/ovary’, ‘anesthesia vaginal hysterectomy incl biopsy’, ‘laparoscopy w/rmvl adnexal structures’, ‘MRI pelvis w/o & w/contrast material’, ‘cystourethroscopy’, etc. were also associated with the diagnosis as well treatment of endometriosis. The finding showed that for every additional procedure on ‘mri of pelvis,’ the odds of endometriosis increased by 1471%. Recent research from Abdominal Radiology, published by Springer Nature, also supported this claim that MRI could be more precise in the diagnosis of endometriosis compared to other diagnostic techniques [56].

As presented in **Table 5**, the procedure ‘laps total hysterect 250 gm/< w/rmvl tube/ovary’ had the odds ratio of 64.53, which implied that if a patient had a ‘laparoscopy with hysterectomy’ then the odds of endometriosis onset increased significantly. Previous studies on endometriosis also cited ‘laparoscopy procedure as the gold standard’ in the diagnosis process [8]. However, while the nomogram graph (**Figure 6**) also suggested that a patient was likely to get diagnosed with endometriosis post this procedure, the data element was further analyzed to understand how it might have correlated to the actual diagnoses, knowing that many laparoscopic procedures were performed to treat other female gynecological conditions. **Figure 6** shows that the feature ‘laparoscope days difference’ presented little importance in predicting the likelihood of the disease onset. The data element measured the significance of laparoscopic procedures in predicting the likelihood of endometriosis via calculating the days’ difference between the laparoscopic procedure and the event date for both target and control cohorts.

Furthermore, the additional analysis revealed that around 60% of the target patients compared to only about 5% of the control group were diagnosed with endometriosis after a laparoscopic procedure performed on the same day of diagnosis. This finding implies that laparoscopy might not actually be a significant driver of the endometriosis diagnosis as presented in the XGB model when accounting for the time component before the diagnosis, although there were statistical significant differences between the two groups.

3. From the patient medical journey and healthcare access side, the ML models suggested that patients often consult with multiple healthcare specialists, including ‘emergency medicine,’ ‘family medicine,’ ‘hematology/oncology,’ ‘internal medicine,’ ‘obstetrics and gynecology’ when experiencing endometriosis-related symptoms and gynecological issues. Since, endometriosis tends to be difficult to diagnose, patients often had a number of unrelated office visits with symptoms associated later with endometriosis. This finding presented that many female patients faced substantial challenges in receiving proper care and treatment. Consequently, patients visited multiple specialists in search of answers for their signs and symptoms [57]. In agreement with these statements, both LR and XGB models presented negative weights and low importance to these healthcare providers’ features, which suggested that if

a patient visited these specialists more frequently, the longer it took to receive a confirmatory endometriosis diagnosis.

Furthermore, women with a history of endometriosis were found more likely to be diagnosed with either an 'ovarian cancer' or 'endometriosis-associated adenocarcinoma' in the future [21, 58–60]. With this in mind, having the ML models identify 'hematology/oncology (SPCLT_HO),' as one of the top Board Certified specialties, further suggested that an office visit with an oncologist should be recommended for any patients presenting signs and symptoms as noted above to rule out any potential cancer risk [21, 61, 62].

4. LR and XGB models also identified additional data elements, which were important in predicting the likelihood of endometriosis onset. The models suggested, as noted in the earlier version of the chapter posted on the Research Square website that data elements like 'deep dyspareunia,' 'female infertility associated with anovulation,' 'premenstrual tension syndrome,' 'hormone replacement therapy,' 'family history of malignant neoplasm of ovary' were identified as highly significant to the prediction endometriosis. Past medical articles supported these claims of fibroids, ovarian cysts, infertility, menstrual period complications, family history of neoplasm of the ovary, hormone therapy, etc. having a strong association with the condition [21, 54]. Furthermore, the finding that women of reproductive age who experience chronic stress were also at a higher risk of developing endometriosis was noted in other medical articles, implying that healthcare providers should consider this symptom in their diagnostic process [21, 63].

5. As mentioned in the preliminary version of the chapter on the Research Square website, 'acetaminophen,' 'megestrol acetate,' 'lidocaine hcl,' etc. were found to be strong predictors of endometriosis occurrence, as these drugs were often prescribed as analgesics to help control pelvic pain. Data elements, including 'submucous leiomyoma of the uterus' and 'hypertrophy of uterus,' were identified as the significant predictors as well [55, 64]; however, more clinical research is required in support of this claim, as these diseases presented similar symptoms, which might impact the ability for healthcare providers to diagnose endometriosis [21, 65].

Overall, the analysis results presented the important data elements to be considered when diagnosing endometriosis in women of reproductive age, to time more accurately disease onset and aid the diagnostic process. As noted in Section 3, when leveraging these features in the diagnostic process, a high accuracy prediction of the disease occurrence was identified, with the model differentiating with high precision between patients with and without the condition. Furthermore, a nomogram graphical representation could be leveraged as one of the tools to graphically predict the outcome given a set of features. Top features were utilized to showcase the practicality of the tool; however, the tool has limitations on the number of data elements that could be applied in the analysis.

5. Conclusions

In this chapter, the crucial role of AI and ML algorithms in disease diagnosis prediction and forecasting was presented, studied, and validated. Patient medical history was leveraged for the ML analysis. LR and XGB models identified important

medical attributes, which were then leveraged to predict the likelihood of endometriosis onset. Early diagnosis can offer an opportunity for women to receive required medical care much earlier in the patient journey.

Leveraging the findings of this study and other related studies can help inform the development of analytical tools and algorithms to be integrated into the Electronic Health Records (EHR) systems to simplify and enhance the diagnosing activities performed by healthcare providers. The enhancements could further inform the diagnostic processes to aid in a timely and precise diagnostic process, ultimately increasing the quality of patient care and life.

Future research should focus on enhancing the ML analysis and exploring advanced deep learning methodologies to improve the accuracy and precision of the current results. Furthermore, imputing the missing data elements with mean and mode values, or even predictive models, can further augment the model performance and increase the accuracy of the ML models in predicting the likelihood of the disease onset. Creating time-based variables (30, 60, 120 days before diagnosis) to account for the time to endometriosis diagnosis would add a significant improvement in the feature engineering step to help with establishing a timeline of events important in the endometriosis diagnostic process.

Acknowledgements

Authors would like to recognize Heather Valera, Suzanne Rosado, and Koichi Iwata for their review of document drafts, and their valuable feedback in improving the article content.

Funding

Authors work for Symphony Health, ICON plc Organization. The data used in the article is the property of Symphony Health, ICON plc Organization. Authors used the healthcare claims data for the sole purpose of publication of this article.

Competing interest

The authors declare that they have no competing interests.

Availability of data and materials

The dataset leveraged for this chapter is a property of Symphony Health, ICON, plc. Data sharing restrictions apply to the availability of these data, and therefore, the dataset is not available for public use.

IntechOpen

Author details

Ewa J. Kleczyk^{1*}, Tarachand Yadav² and Stalin Amirtharaj²

1 Symphony Health, ICON, plc, Blue Bell, PA, USA

2 ICON, plc, Blue Bell, PA, USA

*Address all correspondence to: ewa.kleczyk@symphonyhealth.com

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health*. 2019;22(7):808-815. Available from: <https://pubmed.ncbi.nlm.nih.gov/31277828/> [Accessed: October 1, 2020]
- [2] Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *International Society for Pharmacoconomics and Outcomes Research (ISPOR)*. 2015. DOI: 10.1016/j.jval.2014.12.005 [Accessed: October 1, 2020]
- [3] Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *arXiv Labs*. 2019. Available from: <https://arxiv.org/abs/1806.00388> [Accessed: October 1, 2020]
- [4] Buch VH, Ahmed I, Maruthappu M. Artificial intelligence in medicine: Current trends and future possibilities. *British Journal of General Practice*. 2018;68(668):143-144. DOI: 10.3399/bjgp18X695213 [Accessed: October 1, 2020]
- [5] Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of Digital Imaging*. 2016;30:95-101. DOI: 10.1007/s10278-016-9914-9
- [6] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. *IEEE*. 2017;5:8869-8879. DOI: 10.1109/ACCESS.2017.2694446 [Accessed: October 1, 2020]
- [7] Alexandru AG, Radu IM, Bizon ML. Big data in healthcare—Opportunities and challenges. *Informatica Economică*. 2018;22(2):43-54. DOI: 10.12948/issn14531305/22.2.2018.05
- [8] Rolla E. Endometriosis: Advances and controversies in classification, pathogenesis, diagnosis, and treatment. Version 1. *F1000Research*. 2019;8:F1000. DOI: 10.12688/f1000research.14817.1. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6480968/> [Accessed: March 30, 2021]
- [9] Chapron C, Fauconnier A, Goffinet F, Breart G, Dubuisson JB. Laparoscopic surgery is not inherently dangerous for patients presenting with benign gynaecologic pathology. Results of a meta-analysis. *Human Reproduction*. 2002;17:1334-1342
- [10] Parasar P, Ozcan P, Terry KL. Endometriosis: Epidemiology, diagnosis and clinical management. *Current Obstetrics and Gynecology Reports*. 2017;6(1):34-41. DOI: 10.1007/s13669-017-0187-1
- [11] Hoogeveen M, Dorr PJ, Puylaert JBCM. Endometriosis of the rectovaginal septum: Endovaginal US and MRI findings in two cases. *Abdominal Imaging*. 2003;28:897-901
- [12] Akter S, Xu D, Nagel SC, Bromfield JJ, Pelch KE, Wilshire GB, et al. GenomeForest: An ensemble machine learning classifier for endometriosis. *AMIA Joint Summits on Translational Science proceedings*. 2020;2020:33-42
- [13] Sadia A, Dong X, Nagel Susan C, Bromfield John J, Katherine P, Wilshire Gilbert B, et al. Machine learning classifiers for endometriosis using transcriptomics and methylomics data. *Frontiers in Genetics*. 2019;10:766. DOI: 10.3389/fgene.2019.00766
- [14] Nnoaham KE, Hummelshoj L, Kennedy SH, Jenkinson C, Zondervan KT, World Endometriosis Research Foundation Women's Health

- Symptom Survey Consortium. Developing symptom-based predictive models of endometriosis as a clinical screening tool: Results from a multicenter study. *Fertility and Sterility*. 2012;**98**(3):692-701.e5. DOI: 10.1016/j.fertnstert.2012.04.022. Epub 2012 May 30
- [15] Noventa M, Saccardi C, Litta P, Vitagliano A, D'Antona D, Abdulrahim B, et al. Ultrasound techniques in the diagnosis of deep pelvic endometriosis: Algorithm based on a systematic review and meta-analysis. *Fertility and Sterility*. 2015;**104**(2):366-383.e2. DOI: 10.1016/j.fertnstert.2015.05.002
- [16] Zhang Y, Wang Z, Zhang J, et al. Deep learning model for classifying endometrial lesions. *Journal of Translational Medicine*. 2021;**19**:10. DOI: 10.1186/s12967-020-02660-x
- [17] Endometriosis signs and symptoms. Available from: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/endometriosis> [Accessed: October 1, 2020]
- [18] Endometriosis signs and symptoms. Available from: <https://www.health.qld.gov.au/news-events/news/signs-symptoms-endometriosis> [Accessed: October 1, 2020]
- [19] PRA Health Sciences. Data Insights. Available from: <https://prahs.com/healthcare-intelligence/data-insights>
- [20] Symphony Health Solutions. Available from: <https://symphonyhealth.prahs.com/>
- [21] Kleczyk EJ, Peri A, Yadav T, Komera R, Peri M, Guduru V, et al. Prsedicting endometriosis onset using machine learning algorithms. ResearchSquare. Available from: <https://www.researchsquare.com/article/rs-135736/v1>. DOI: 10.21203/rs.3.rs-135736/v1 [Accessed: October 4, 2021]
- [22] Getting the Most Out of Longitudinal Patient Data. Anonymous patient-level data (APLD). Available from: <https://www.rxdatascience.com/blog/getting-most-out-of-longitudinal-patient-data> [Accessed: October 1, 2020]
- [23] Marketing, Patient Data, and Privacy Concerns. Available from: <https://www.reutersevents.com/pharma/commercial/marketing-patient-data-and-privacy-concerns> [Accessed: October 5, 2020]
- [24] Integrated Dataverse (IDV®). <https://symphonyhealth.prahs.com/what-we-do/view-health-data> [Accessed: October 1, 2020]
- [25] Symphony Health Solutions, What We Do. Available from: <https://symphonyhealth.prahs.com/what-we-do>
- [26] Ali MS, Prieto-Alhambra D, Lopes C, Ramos D, Bispo N, Ichihara MY, et al. Propensity score methods in health technology assessment: Principles, extended applications, and recent advances. *Frontiers in Pharmacology*. 2019;**10**:973. DOI: 10.3389/fphar.2019.00973 [Accessed: October 1, 2020]
- [27] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;**70**:41-55. DOI: 10.1093/biomet/70.1.41 [Accessed: October 1, 2020]
- [28] Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985;**39**(1):33-38. DOI: 10.1080/00031305.1985.10479383 [Accessed: October 1, 2020]
- [29] Xu Y, Goodacre R. On splitting training and validation set: A comparative study of cross-validation,

bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*. 2017;2(3):249-262. DOI: 10.1007/s41664-018-0068-2 [Accessed: October 1, 2020]

[30] Ballantyne Draelos RL. Best Use of Train/Val/Test Splits, with Tips for Medical Data. *Glass Box Machine Learning and Medicine*. Available from: <https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/>. [Accessed: October 5, 2020]

[31] Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nature Methods*. 2018;15:233-234. DOI: 10.1038/nmeth.4642 [Accessed: March 30, 2021]

[32] Simeone O. A very brief introduction to machine learning with applications to communication systems. arXiv preprint arXiv:1808.02342v4. 2018

[33] Cochran WG. The Chi-square test of goodness of fit. *The Annals of Mathematical Statistics*. 1952;23(3):315-345. DOI: 10.1214/aoms/1177729380 [Accessed: October 5, 2020]

[34] Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *Plos One*. 2019;14(11):20. DOI: 10.1371/journal.pone.0224365 [Accessed: October 5, 2020]

[35] Kotsiantis SB. Supervised machine learning: A review of classification techniques. *Informatica*. 2007;31:249-268

[36] Hinton G, Sejnowski T. *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA: MIT Press; 1999. pp. vii-xv. ISBN: 978-0262581684

[37] Wosiak A, Zamecznik A, Niewiadomska-Jarosik K. Supervised

and unsupervised machine learning for improved identification of intrauterine growth restriction types. In: *Federated Conference on Computer Science and Information Systems (FedCSIS)*. Gdańsk, Poland: IEEE; 2016

[38] Hastie T, Tibshirani R, Friedman J. "Unsupervised Learning," *The Elements of Statistical Learning*. New York, NY: Springer Series in Statistics, Springer; 2009. pp. 485-585

[39] Logistic Regression. Available from: https://en.wikipedia.org/wiki/Logistic_regression

[40] Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 2001;29:1189-1232. DOI: 10.1214/aos/1013203451 [Accessed: October 1, 2020]

[41] Cramer JS. The origins of logistic regression. *Tinbergen Institute Discussion Paper*. TI 2002-119/4. Available from: <https://papers.tinbergen.nl/02119.pdf> [Accessed: October 1, 2020]

[42] Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: Wiley; 2013. ISBN: 978-0-470-58247-3

[43] Agresti A. *Categorical Data Analysis*. Hoboken: John Wiley and Sons; 2012. ISBN: 978-0-470-46363-5

[44] Extreme Gradient Boosting. Available from: <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>

[45] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R, editors. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; August 13-17, 2016; San Francisco, CA, USA. ACM; 2016. pp. 785-794.

DOI: 10.1145/2939672.2939785
[Accessed: October 5, 2020]

[46] Hastie T, Tibshirani R, Friedman JH. '10. Boosting and Additive Trees'. *The Elements of Statistical Learning*. 2nd ed. New York: Springer; 2009. pp. 337-384

[47] On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*. 1922;**85**(1):87-94. DOI: 10.2307/2340521

[48] Chi-Square feature selection. "Scikit-learn" Python Library. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html [Accessed: October 5, 2020]

[49] Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval. Feature Selection, Chi-Square Feature Selection*. Cambridge, UK: Cambridge University Press; 2008

[50] OBG Management. Endometriosis and infertility: Expert answers to 6 questions to help pinpoint the best route to pregnancy. *Mdedge ObGyn*. 2015;**27**(6):30-35. Available from: <https://www.mdedge.com/obgyn/article/99912/surgery/endometriosis-and-infertility-expert-answers-6-questions-help-pinpoint/> [Accessed: October 5, 2020]

[51] Kattan MW, Marasco J. What is a real nomogram? *Seminars in Oncology*. 2010;**37**:23-26

[52] Su D, Zhou X, Chen Q, et al. Prognostic nomogram for thoracic esophageal squamous cell carcinoma after radical esophagectomy. *PLoS One*. 2015;**10**:e0124437

[53] Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*. 2001;**16**(3):199-231.

DOI: 10.1214/ss/1009213726 [Accessed: March 30, 2021]

[54] Endometriosis—Risks, Signs, Symptoms, Diagnosis and Treatment. Available from: <https://www.mayoclinic.org/diseases-conditions/endometriosis/symptoms-causes/syc-20354656> [Accessed: October 5, 2020]

[55] Liang B, Xie YG, Xu XP, Hu CH. Diagnosis and treatment of submucous myoma of the uterus with interventional ultrasound. *Oncology Letters*. 2018;**15**(5):6189-6194. DOI: 10.3892/ol.2018.8122 [Accessed: October 5, 2020]

[56] Tong A, VanBuren WM, Chamié L, Feldman M, Hindman N, Huang C, et al. Recommendations for MRI Technique in the evaluation of pelvic endometriosis: Consensus statement from the Society of Abdominal Radiology Endometriosis Disease-Focused Panel. *Abdominal Radiology*. 2020;**45**(6):1569-1586. DOI: 10.1007/s00261-020-02483-w [Accessed: March 30, 2021]

[57] Agarwal SK, Antunez-Flores O, Foster WG, et al. Real-world characteristics of women with endometriosis-related pain entering a multidisciplinary endometriosis program. *BMC Women's Health*. 2021;**21**:19. DOI: 10.1186/s12905-020-01139-7 [Accessed: March 30, 2021]

[58] Kvaskoff M, Horne AW, Missmer SA. Informing women with endometriosis about ovarian cancer risk. *The Lancet Journal*. 2017;**390**(10111):2433-2434. DOI: 10.1016/S0140-6736(17)33049-0 [Accessed: October 5, 2020]

[59] Brilhante A, Augusto KL, Cavalcante Portela M, Sucupira L, Oliveira F, Pouchaim A, et al. Endometriosis and ovarian cancer: An integrative review (endometriosis and ovarian cancer). *Asian Pacific Journal of*

Cancer Prevention. 2017;**18**(1):11-16.
DOI: 10.22034/APJCP.2017.18.1.11
[Accessed: October 5, 2020]

[60] Cunha JP. What Will Happen If Endometriosis Is Not Treated? Emedicinehealth. 2019. Available from: https://www.emedicinehealth.com/ask_what_will_happen_if_endometriosis_not_treated/article_em.htm#doctor%E2%80%99s_response [Accessed: October 5, 2020]

[61] Coppa AM. What Happens if Endometriosis is Left Untreated? Available from: <https://www.drcoppaobgyn.com/blog/what-happens-if-endometriosis-is-left-untreated>

[62] Endometriosis and Ovarian Cancer Risk. Available from: <https://ovarian.org.uk/news-and-blog/blog/endometriosis-and-ovarian-cancer-risk/> [Accessed: October 5, 2020]

[63] Reis FM, Coutinho LM, Vannuccini S, Luisi S, Petraglia F. Is stress a cause or a consequence of endometriosis? *Reproductive Sciences*. 2020;**27**:39-45. DOI: 10.1007/s43032-019-00053-0 [Accessed on October 5, 2020]

[64] Endometriosis vs. Adenomyosis: Similarities and Differences. Available from: <https://www.healthline.com/health/womens-health/adenomyosis-vs-endometriosis> [Accessed: October 5, 2020]

[65] Endometrial Hyperplasia. Available from: <https://my.clevelandclinic.org/health/diseases/16569-atypical-endometrial-hyperplasia> [Accessed: October 5, 2020]