Predicting Endometriosis Onset Using Machine Learning Algorithms

Ewa J Kleczyk (✉ ewa.kleczyk@symphonyhealth.com)

Symphony Health <https://orcid.org/0000-0001-8902-8349>

Apama Peri

Symphony Health

Tarachand Yadav

Symphony Health

Ramachandra Komera

Symphony Health Solutions: Symphony Health

Maruthi Peri

Symphony Health Solutions: Symphony Health

Vara Guduru

Symphony Health Solutions: Symphony Health

Stalin Amirtharaj

Symphony Health Solutions: Symphony Health

Ming Huang

Symphony Health Solutions: Symphony Health

Research article

Keywords: Endometriosis, infertility, likelihood, Logistic Regression, Machine Learning, eXtreme Gradient Boosting

Posted Date: December 31st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-135736/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Endometriosis is a common progressive female health disorder in which tissues similar to the lining of the uterus grow on other parts of the body like ovaries, fallopian tubes, bowel, and other parts of reproductive organs. In women, it is one of the most common causes of pelvic pain and infertility. In the US, one in every ten women of reproductive age group has endometriosis. The actual cause of endometriosis is still unknown, and it is quite difficult to diagnose. There are several theories regarding the cause; however, not a single theory has been scientifically proven.

Methods

In this paper, we try to identify the drivers of endometriosis' diagnoses via leveraging advanced Machine Learning (ML) algorithms. The primary risks of infertility and other health complications can be minimized to a great extent, if likelihood of endometriosis can be predicted well in advance. As a result, the proper medical care and treatment can be given to the impacted patients. To demonstrate the feasibility, Logistic Regression (LR) and eXtreme Gradient Boosting (XGB) algorithms were trained on 36 months of medical history data.

Results

The machine learning models were used to predict the likelihood of disease on qualified patients from the healthcare claims patient level database. Several directly and indirectly features were identified as important in accurate prediction of the condition onset, including selected diagnosis and procedure codes.

Conclusions

Leveraging the machine learning approaches can aid early prediction of the disease and offer an opportunity for patients to receive the needed medical treatment earlier in the patient journey. Creating a typing tool that can be integrated into the Electronic Health Records (EHR) systems and easily accessed by healthcare providers could further aid the objective of improving the diagnosis activities and inform the diagnostic processes that would result in timely and precise diagnosis, ultimately increasing patient care and quality of life.

1. Background

Recent advancement in Artificial Intelligence (AI) and Machine Learning (ML) has provided the opportunity for AI and ML application in the healthcare area, while also slowly improving on the performance benchmark set by the classical statistical techniques [1]. In recent years, healthcare service providers have also shown interest towards data science and machine learning in disease diagnosing. Disease prediction using data mining and machine learning algorithms with patient medical history such as diagnosis of disease, medical and surgical procedures, therapeutics, and treatments, etc., has been slowly introduced to aid decision making processes [2, 3, 4]. Many statistical and machine learning techniques have been applied to either pathological or clinical data to study the disease in detail and also predict its likelihood of occurrence. Deep learning algorithms such as Convolutional Neural Network (CNN) have been found to predict disease onset and progression with a greater precision compared to analyzing just medical image data [5].

Since healthcare is one of the leading industries with a large amount of structured and unstructured data, it is imperative to use the known advanced techniques to extract the hidden data patterns. Machine Learning algorithms with the help of big data technology has made it easier to mine the vast amount of unstructured data and aided in making important decisions related to patients' health [6]. Due to its high precision and robustness in comparison to conventional statistical methods, most medical scientists have been attracted towards these models to understand the key drivers of disease onset and progression prediction. Artificial Intelligence, Machine Learning, and big data have been playing a pivotal role in improving healthcare infrastructure, patient care, as well as disease diagnosing, prediction and forecasting, drug discovery, etc., and thereby, reducing medical costs, shortening the time to diagnoses and treatment, as well as enhancing patients' quality of life and access to healthcare [7].

With this motivation in mind, we selected endometriosis as the condition to study in this article. Endometriosis is one of the most common disorders seen in women of a menstruating age in which tissues like the endometrium lining grow on the outer part of the uterus and other organs of the pelvic region. The signs and symptoms vary from patient to patient with some patients having mild symptoms, while others display a moderate to severe level of condition occurrence. The most common symptoms of endometriosis are pelvic pain, dysmenorrhea,

and infertility. There is no guaranteed treatment for endometriosis at this time; however, with an early diagnosis and available medical and surgical options, healthcare providers can reduce the risks of potential complications and improve the quality of life for their patients. If we can identify or predict the probability of endometriosis onset by analyzing the medical history of diagnosed patients, the results might help benefit both the healthcare providers' diagnosis process and patients' well-being and quality of life. In this study, the Logistic Regression (LR) and eXtreme Gradient Boosting (XGB) algorithms were used to predict endometriosis occurrence when leveraging medical history of the diagnosed patients.

The remainder of the article is organized as follows: in Sect. 2, we briefly review the project objective; in Sect. 3, we describe different methods used in data preparation, feature engineering, feature selection and model training and validation; in Sect. 4, we present the model outputs and results; and in Sect. 5, we conclude the study with a summary of our findings.

2. Objectives

The following objectives will be addressed in this article:

- i. Train machine learning algorithms to predict the likelihood of endometriosis.
- ii. Identify the most significant medical events in the patient journey that lead to the diagnosis of endometriosis.
- iii. Score entire database using the best performing trained models.
- iv. Profile patients using the predicted scores.

3. Methods Overview

The data source for this project is the healthcare claims patient level database with the study time period from January 31, 2019 to December 31, 2019. Patient cohorts: study target and control were established using endometriosis ICD 10 diagnosis codes. As endometriosis is a female only condition, female patients 18 and older were part of the study target cohort. A control cohort is often used to create a patient sample to compare with the study target cohort and is selected using cohort matching algorithms. 36 months of patient medical history prior to the first disease event in 2019 were extracted for both the study target and control cohorts. The healthcare claims patient level data includes diagnosis codes, medical and surgical codes, therapeutics and treatments prescribed at the transactional level.

A number of analytical methods was leveraged for the analysis from the rules-based patient qualification criteria to Machine Learning algorithms to derive probability of endometriosis onset. The following sub-sections of the article present a detailed explanation for each of the selected methods. The healthcare claims patient level dataset considered in the analysis is specific to the US healthcare market.

3.1. Healthcare claims patient level database

The healthcare claims patient level database is an anonymous longitudinal patient data set that can be used by organizations that are directly or indirectly associated to healthcare [9, 41]. There has been an increasing interest in patient-level data, as researchers, healthcare providers, and pharmaceutical companies are realizing the potential of creating better comparisons of effective treatment outcomes by analyzing longitudinal data that represent individual patient-based experiences and interactions with the US healthcare system [42].

The healthcare claims patient level database leveraged for this study consists of medical, hospital, and prescriptions claims across all payment types [10, 44]. The database covers more than 317 million patients in the US, spans over more than 17 years of medical health history, and includes more than 1.9 million healthcare providers [43]. Figure 1 presents the summary of information in the database.

3.2. Cohort selection

For this study, we identified 314,101 confirmed endometriosis patients in 2019 in the healthcare claims patient level database, using predefined ICD 10 diagnosis codes (Table 1). Female patients age 18 and above were selected to the study target cohort. For the control cohort, a random sample of 3 million female patients with the same age criterion was extracted from the database.

Table 1
ICD 10 diagnosis codes of endometriosis

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

To select a control cohort of an equal size to the study target groups out of 3 million patients, a noble technique known as ‘propensity score match’ was used [18]. Propensity matching algorithm [19], a statistical technique, selects the control cohort based on similar characteristics or covariates observed in the study target cohort. Covariates considered for selection were patient age and medical history [20]. Table 2 presents the distribution comparison between the study target and control cohorts by age and Census geographies. The patient age variable was created via grouping age ranges and US states were grouped into regions.

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

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

3.3. Data extraction

The next step in the analysis process was to extract the entire medical history of the patients from the available information in the healthcare claims patient level database. In order to ensure extraction of healthcare history data prior to the first condition event, the event date for the target cohort was established for each patient. In the case of the control cohort, the first activity in 2019 was considered as the event date.

Using these event dates of respective patients, 36 months of medical history data was extracted. Historical data presented all the medical events in patient history, including diagnoses for comorbid conditions, medical and surgical procedures, therapeutics, and treatment prescribed to patients. Top 1000 diagnosis codes, top 800 medical and surgical procedures, and top 500 prescribed drugs were only considered for further analysis as these top codes constituted more than 80% of total data. A pivot table was created where data at the transaction level was aggregated by the anonymized patient ID. After historical medical claims data preprocessing for both cohorts independently, a dataset was integrated into a single data frame. The integrated data frame had more than 2,600 features. The dataset was further standardized and split into two groups, a training and test set, using 70:30 ratio respectively [21]. The training dataset is used to identify the key features of endometriosis onset, while the test group is used to validate if these features would predict the test group condition onset accurately [22]. Splitting the data into train and test sets helps to assess the model performance and its generalizing ability on unseen data [23].

3.4. Machine Learning algorithms’ overview

Machine Learning algorithms can be grouped into two categories: supervised and unsupervised learning.

3.4.a. Supervised learning algorithms

Supervised learning is the process of training or building the machine learning algorithms in which algorithms learn to map from input space (X) to output space (Y), i.e. $Y = f(X)$ [25]. The major objective is to approximate the mapping function (f) in order to ensure that when a new data point (x) is added we can predict (y) outcome [26]. Supervised learning algorithms are mainly used for classification and prediction problems [32]. Following are the most popular supervised algorithms: *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.* [31].

3.4.b. Unsupervised learning algorithms

Unsupervised learning algorithms, on the other hand, try to learn the hidden pattern within the input dataset (X) [28]. These models are called unsupervised because there is no supervision to guide the models as compared to the supervised learning [29]. Algorithms are left at their own abilities to learn, discover and showcase the patterns in the input data (X). These algorithms are highly popular in the tasks to discover the natural clusters, dimension reduction, anomaly detection, etc. *k-Means clustering, principal component analysis (PCA), factor analysis (FA), singular value decomposition (SVD), apriori algorithm (association rule)* are some popular examples of unsupervised learning algorithms [31].

Depending on the study objectives and the available data, algorithms are explored, tested for performance and data type fit, and selected accordingly. We framed the endometriosis onset prediction into a supervised classification problem and selected Logistic Regression and XGB models to develop a highly predictive algorithm of the disease onset. SVM, RF, AdaBoost, ANN, etc. are the other options that were explored in disease prediction; however, Logistic Regression and XGB were selected to predict the condition onset. Logistic Regression allows study of the odds of endometriosis occurrence for a given medical event [15], while XGB has more flexibility in fine tuning the hyper-parameters in comparison to other tree based algorithms [11].

Logistic Regression

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist [14, 15]. Mathematically, a binary logistic model has a dependent variable with two possible values, where the two values are labeled "0" and "1" [33]. Outputs with more than two values are modeled by multinomial logistic regression. Logistic Regression is used in various fields, including healthcare and social sciences [34].

xExtreme Gradient Boosting

Gradient boosting algorithm is a machine learning algorithm which is an ensemble of weak prediction models, mostly decision trees [11]. An individual tree is a simple, often unreliable, model but when multiple trees are grouped together, they can create a robust algorithm [12]. XGB starts by creating a first simple tree [35], which then progresses sequentially and builds upon the weaker learners, with each iteration revising the previous tree until an optimal point is reached, such as the number of trees (estimators) to build the solution [36].

Chi-Square Test

The Chi-square test is one of the most widely used non-parametric tests [37], often utilized to test the independence between observed and expected frequencies of one or more attributes in a contingency table, popularly known as 'test goodness of fit' [38]. In this work, the Chi-square test is used to identify top significant features given the dependent variable (Y) [40].

Logistic Regression, being the simplest of the machine learning algorithms, was selected as the base model for the analysis and used to compare other models' performance. Both Logistic Regression and XGB models were trained, and top 1,000 features from each algorithm were selected out of more than 2,600 features used in the model runs. To decrease the number of data elements and to select only the most important variables to predicting the condition onset, we also used a Chi-Square test to identify the top 1,000 features. As a next step, the unique features from each model were utilized to train the final machine learning model to predict the endometriosis occurrence probability. Algorithms were trained on Python 3.5 using 'scikit-learn' and 'xgboost' libraries.

4. Results

4.1. Significant features selection

Table 3
presents the machine learning model performance metrics, which indicate that both the Logistic Regression and XGB models performed relatively well in predicting the condition onset. The models' accuracy ranged between 88% – 96%.

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

Figure 2 presents the Receiver Operating Characteristic (ROC) curves on the test set for Logistic Regression and XGB models. The Area under the ROC Curve (AUC) had values between 0.88–0.96. Chi-square test was also applied on data before standardization.

The top 1,000 features were selected from the Logistic Regression, XGB and Chi-square algorithms to train the final machine learning model. Most of the top features identified by the selected models were related to medical and surgical procedures as well as diagnosis codes. Patients diagnosed with endometriosis underwent a series of medical and surgical procedures and had various diagnostic symptoms and comorbid conditions. The Chi-square significance test was run at the 95% significance confidence interval to aid in identification of the topmost significant features.

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

Feature	Feature Description
D_N85_8	Other specified non inflammatory disorder of uterus
D_N94_6	Dysmenorrhea, unspecified
D_N94_9	Unspecified condition associated with female genital organs and menstrual cycle
D_R10_2	Pelvic and Perineal Pain
D_Z01_419	Encounter for gynecological examination (general) (routine) without abnormal findings
P_00840	Anesthesia Intraperitoneal Lower Abd W/Laps Nos
P_00944	Anesthesia vaginal hysterectomy incl biopsy
P_52000	Cystourethroscopy
P_58571	Laps total hysterect 250 GM/< w/rmvl tube/ovary
P_58573	Laparoscopy tot hysterectomy > 250 g w/tube/ovar
P_58662	Laps Fulg/Exc Ovary Viscera/ Peritoneal Surface
P_76830	Us Transvaginal
P_J1950	Injection. Leuprolide acetate (for depot suspens)
R_Norethindrone_Acetate	Norethindrone Acetate
SPCLT_EM	Emergency Medicine
SPCLT_FM	Family medicine
SPCLT_HO	Hematology/Oncology
SPCLT_OBG	Obstetrics and gynecology

Section 1 of this work describes endometriosis and its associated signs and symptoms such as ‘*painful periods*’, ‘*lower abdominal and pelvic pain*’, ‘*heavy bleeding during periods*’, ‘*pain during urination and bowel movement*’, ‘*constipation and diarrhea*’, ‘*infertility*’, ‘*painful sexual intercourse*’, etc. [16, 17]. Identifying these prominent medical events from patients’ medical history by the models is the objective of this work. Hence, it is desirable to validate the model performance by analyzing the top features, whether they would help predict endometriosis’ onset.

Table 4 presents the top features identified by the machine learning models, which are directly or indirectly associated with endometriosis. Features such as ‘*non inflammatory disorder of uterus (D_N85_8)*’, ‘*pelvic and perineal pain (D_R10_2)*’ are the diagnosis codes, presenting the association with the risks and symptoms of endometriosis [45]. Procedure codes such ‘*anesthesia of lower abdomen for laparoscopy (P_00840)*’, ‘*vaginal hysterectomy including biopsy (P_00944)*’ are the top procedures often associated with the diagnosis as well treatment of endometriosis [45]. Furthermore, the machine learning models suggest that patients often consult with specialists including ‘*emergency medicine (SPCLT_EM)*’, ‘*family medicine (SPCLT_FM)*’, ‘*obstetrics and gynecology (SPCLT_OBG)*’ when experiencing related symptoms and gynecological issues. Overall, the machine learning models selected top features closely related to the onset of endometriosis, which implies that when tracking any of the features the condition onset could be diagnosed sooner.

4.2. Feature selection for market definition

Top features from all three algorithms, which were specific to target cohort were identified. These features presented to be important in diagnosing the endometriosis condition and were selected for patient scoring criteria. The therapeutics as well as medical and surgical procedure codes specific to endometriosis treatment such as Orilissa, Marilissa, and Lupron Depot were excluded. Around 9.5 million female patients age 18 and above were qualified for scoring.

4.3. Propensity model training and validation

Using the top features selected, Logistic Regression and XGB models were re-trained. As the number of features was reduced, in the beginning we observed a drop in model performance. After several iterations and hyper-parameter tuning, the predictive power of XGB

significantly improved compared to the previous iterations; however, we did not see any improvement in the Logistic Regression model results. Interestingly, both models were able to identify additional new features aligned with endometriosis.

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

Features	Feature Description
D_D25_0	Submucous leiomyoma of uterus
D_F43_0	Acute stress reaction
D_N83_291	Other ovarian cyst, right side
D_N85_2	Hypertrophy of uterus
D_N92_4	Excessive bleeding in the premenopausal period
D_N94_12	Deep dyspareunia
D_N94_3	Premenstrual tension syndrome
D_N94_5	Secondary dysmenorrhea
D_N97_0	Female infertility associated with anovulation
D_Z79_890	Hormone replacement therapy
D_Z80_41	Family history of malignant neoplasm of ovary
P_58661	Laparoscopy w/rmvl adnexal structures
R_ACETAMINOPHEN	Acetaminophen
R_MEGESTROL_ACETATE	Megestrol acetate
R_LIDOCAINE_HCL	Lidocaine hcl

The re-trained machine learning models identified all the top features discussed in Sect. 4.1. In Table 5, we present the additional features recognized by XGB and Logistic Regression models, which are highly significant in predicting the likelihood of endometriosis. The models suggest that features like ‘*submucous leiomyoma of uterus (D_D25_0)*’, ‘*ovarian cyst (D_N83_291)*’, ‘*deep dyspareunia (D_N94_5)*’, ‘*female infertility associated with anovulation (D_N97_0)*’ are important in predicting the likelihood of endometriosis. The models have also flagged *Acetaminophen (R_ACETAMINOPHEN)*, *Megestrol acetate (R_MEGESTROL_ACETATE)* & *Lidocaine hcl (R_LIDOCAINE_HCL)* drugs as the strong predictors of endometriosis.

Table 6

shows that the XGB model performed better compared to the Logistic Regression model. Figure 3 shows the Receiver Operating Characteristic (ROC) curves on test sets for both re-trained Logistic Regression and XGB models. The Area under the ROC Curve (AUC) values of LR and XGB models on test were 0.87 and 0.96 respectively. Figure 4 suggests that the XGB model was able to more accurately differentiate target from control than LR model. Hence, we used XGB model to score the qualified patients.

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

4.4. Scoring qualified patients

The last step of the model evaluation is to score qualified patients to assess the model's predictability of condition onset. A complete medical history of 9.5 million qualified patients was extracted for 36 months, which included diagnosis codes, medical and surgical procedure codes, medications and treatments prescribed as well as practitioners' therapy expertise and Board-Certified Specialty. After data pre-processing, the likelihood of endometriosis was predicted using the trained XGB model.

A probability distribution of 9.5 million scored patients is shown in Fig. 5. We observed that most of the predicted probability values are concentrated either towards 0 or 1. Considering 0.5 as the threshold, the XGB model suggests that around 36% of the scored patients are likely to get diagnosed with endometriosis sometime in the future. Assuming an ability to leverage the significant variables in diagnosing the condition onset, practitioners can give special medical care and advice in time to these patients, thereby, reducing the risks of endometriosis and its related complications.

5. Discussion

Overall, the machine learning models have identified top features that can explain endometriosis onset in advance. As noted, Tables 4 and 5 in the 4. Results Section, these features include diagnosis codes, medical and surgical procedure codes, as well as physician specialties that often support patients through their healthcare journey.

For the preliminary Logistic Regression, XGB, and Chi-Square runs as noted in Table 4, the following top variables were identified as important in predicting the condition onset: 1) diagnoses codes: '*non inflammatory disorder of uterus (D_N85_8)*', '*dysmenorrhea (D_N94_6)*', '*pelvic and perineal pain (D_R10_2)*', '*unspecified condition associated with female genital organs and menstrual cycle (D_N94_9)*' clearly show association with the risks and symptoms of endometriosis [45]; 2) medical and surgical procedure codes such '*anesthesia of lower abdomen for laparoscopy (P_00840)*', '*vaginal hysterectomy including biopsy (P_00944)*', '*cystourethroscopy (P_52000)*', '*laparoscopy, surgical with fulguration or excision of lesions of the ovary, peritoneal surface (P_58662)*' are associated with the diagnosis as well treatment of endometriosis [45].

From the patient medical journey and healthcare access side, the machine learning models suggest that patients often consult with specialists, including '*emergency medicine (SPCLT_EM)*'; '*family medicine (SPCLT_FM)*'; '*obstetrics and gynecology (SPCLT_OBG)*' when experiencing endometriosis related symptoms and gynecological issues. Patients with the history of endometriosis or untreated endometriosis are at a higher risk of developing either an ovarian cancer or '*endometriosis associated adenocarcinoma*', which can also serve as an indicator of potential occurrence of the condition [52, 53, 54]. The machine learning models selected as one of the top healthcare provider specialties '*hematology/oncology (SPCLT_HO)*'. This finding suggests that if a patient has any signs and symptoms as noted above, a consultation with an oncologist is recommended [55, 56]. Overall, the machine learning models selected top features directly related to the onset of endometriosis, which implies that when tracking any of the features the condition onset could be diagnosed sooner.

As noted in Table 5 above, Logistic Regression and XGB models identified additional features, which are important in predicting the likelihood of endometriosis. The models suggest that features like '*submucous leiomyoma of uterus (D_D25_0)*'; '*ovarian cyst (D_N83_291)*'; '*hypertrophy of uterus (D_N85_2)*'; '*excessive bleeding in the premenopausal period (D_N92_4)*'; '*deep dyspareunia (D_N94_5)*'; '*female infertility associated with anovulation (D_N97_0)*'; '*premenstrual tension syndrome (D_94_3)*'; '*hormone replacement therapy (D_Z79_890)*'; '*family history of malignant neoplasm of ovary*' are highly significant in predicting the likelihood of endometriosis. There are also several articles, which support the models' claims that *fibroids, ovarian cysts, infertility, menstrual period complications, family history of neoplasm of ovary, hormone therapy etc.* have strong association with endometriosis [48]. Recent clinical research also supports that women of reproductive age with '*chronic stress*' are at a higher risk of developing endometriosis [47].

The machine learning models have also identified *Acetaminophen (R_ACETAMINOPHEN)*, *Megestrol acetate (R_MEGESTROL_ACETATE)* & *Lidocaine hcl (R_LIDOCAINE_HCL)* drugs as the strong predictors of endometriosis, as these drugs are often prescribed as analgesics, birth control & treatment of endometrial cancer and to numb the skin/muscles respectively. Furthermore, features such as '*submucous leiomyoma of uterus (D_D25_0)*' and '*hypertrophy of uterus (D_N85_2)*' are significant predictors [49, 50] in the disease onset; however, more clinical research is needed to support this statement, as these conditions have similar symptoms, but patients are less likely to develop endometriosis [51].

Overall, the top data elements present the key features that should be considered when diagnosing endometriosis in adult women in order to decrease the time to diagnosis. As noted in the 4.4 Section of the article, when using these variables in the diagnostic processes, we can with a high accuracy predict the condition onset and differentiate accurately between patients with and without the disease.

6. Conclusions

In this article, we validated the crucial role of AI and ML in the disease diagnosis, prediction, and forecasting. We analyzed medical history of patients with endometriosis using machine learning algorithms and re-trained XGB model on selected important features, which were applied to predict the likelihood of endometriosis occurrence in the adult female population. Early prediction of the disease can offer an opportunity for patients to receive needed medical treatment earlier in the patient journey. Creating a typing tool that can be integrated into the Electronic Health Records (EHR) systems and easily accessed by healthcare providers could further aid the objective of improving the diagnosis activities and inform the diagnostic processes that would result in timely and precise diagnosis, ultimately increasing patient care and quality of life. In our future work, we plan to explore advanced deep learning algorithms to further enhance the model performance and increase the accuracy of the machine learning models in predicting the likelihood of the disease onset.

Abbreviations

- Artificial Intelligence (AI)
- Machine Learning (ML)
- Logistic Regression (LR)
- eXtreme Gradient Boosting (XGB)
- Principal Component Analysis (PCA)
- Factor Analysis (FA)
- Singular Value Decomposition (SVD)
- Receiver Operating Characteristic Curve (ROC)
- Area under the ROC Curve (AUC)
- Electronic Health Records (EHR)

Declarations

- Ethics approval and consent to participate: Symphony Health, PRA Health Sciences Privacy Risk Review Group reviewed the article. The de-identified dataset is used for analytics only. No direct patient identifiers are noted. The analysis presents only a negligible risk of re-identification of an individual, which is consistent with HIPAA Privacy Rules. No additional administrative permissions or ethics approvals were required to access and use the medical claims data described in this study.
- Consent for publication: Not applicable.
- Availability of data and materials: The data that support the findings of this study are available from Symphony Health, PRA Health Sciences, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.
- Competing Interest: The authors declare that they have no competing interests.
- Funding: Authors work for Symphony Health, PRA Health Sciences. The data used in the article is the property of Symphony Health, PRA Health Sciences. Authors used the data for publication of this article.
- Authors' contributions: all authors have read and approved the manuscript.
 - EJK: Research Principal, Data Scientist – corresponding author, responsible for the overall study design, analytics plan, and documentation
 - AP: Project Manager - responsible for day to day activities of the research
 - TY: Lead Data Scientist: leading and conducting the analytics for the research project
 - RK, M, VG, S, and MH: Data Scientists on the research project, performing the analytical analysis
- Acknowledgements: Authors would like to recognize Heather Valera and Koichi Iwata for their review of document drafts, and their valuable feedback in improving the article content.

References

1. Doupe P, Faghmous J, Basu S., Machine Learning for Health Services Researchers. *Value Health*. 22(7): 808-815, 2019.
2. William H. Crown, PhD. Potential application of machine learning in health outcomes research and some statistical cautions. *International Society for Pharmacoeconomics and Outcomes Research (ISPOR)*, 2015. 1098-3015\$36.00, DOI: <https://doi.org/10.1016/j.jval.2014.12.005>
3. Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *arXivLabs*. 2019 v4, <https://arxiv.org/abs/1806.00388>
4. Varun H Buch, Irfan Ahmed, Mahiben Maruthappu. Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice* 2018; 68 (668): 143-144. DOI: <https://doi.org/10.3399/bjgp18X695213>
5. Alvin Rajkomar, Sneha Lingam, Andrew G. Taylor, Michael Blum, John Mongan. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of Digital Imaging* 30, 95–101(2016). DOI: <https://doi.org/10.1007/s10278-016-9914-9>
6. Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE*, 2169-3536 (2017), DOI: <https://doi.org/10.1109/ACCESS.2017.2694446>
7. Adriana Gabriela Alexandru, Irina-Miruna Radu, Madalina - Lavinia Bizon. Big data in healthcare - opportunities and challenges. *Informatica Economică vol.22, no. 2/2018*. DOI: <https://doi.org/10.12948/issn14531305/22.2.2018.05>
8. Iroju Olaronke, Ojerinde Oluwaseun. Big data in healthcare: Prospects, challenges and resolutions. *IEEE*, 16602629, 2016. DOI: <https://doi.org/10.1109/FTC.2016.7821747>
9. Getting the Most Out of Longitudinal Patient Data. *Anonymous patient-level data (APLD)* [Online] <https://www.rxdatascience.com/blog/getting-most-out-of-longitudinal-patient-data>
10. Integrated Dataverse (IDV®). [Online] <https://symphonyhealth.prahs.com/what-we-do/view-health-data>
11. Jerome H. Friedman. Greedy function approximation : A gradient boosting machine. *The Annals of Statistics Volume 29*, (2001), 1189-1232 DOI: <https://doi.org/10.1214/aos/1013203451>
12. Extreme Gradient Boosting. [Online] <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>,
13. <https://info.cambridgespark.com/latest/getting-started-with-xgboost>
14. S. Cramer. The origins of logistic regression. *Tinbergen Institute discussion paper, TI 2002-119/4*

15. Logistic Regression. [Online] https://en.wikipedia.org/wiki/Logistic_regression
16. Endometriosis signs and symptoms. [Online] <https://www.hopkinsmedicine.org/health/conditions-and-diseases/endometriosis>
17. Endometriosis signs and symptoms. [Online] <https://www.health.qld.gov.au/news-events/news/signs-symptoms-endometriosis>
18. M Sanni Ali, Daniel Prieto-Alhambra, Luciane Cruz Lopes, Dandara Ramos, Nivea Bispo, Maria Y. Ichihara, Julia M. Pescarini, Elizabeth Williamson, Rosemeire L. Fiaccone, Mauricio L. Barreto, and Liam Smeeth. Propensity score methods in health technology assessment: principles, extended applications, and recent advances. *Front Pharmacol* 10: 973 (2019). DOI: <https://dx.doi.org/10.3389/fphar.2019.00973>
19. Rosenbaum P. R., Rubin D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55 (1983). DOI: <https://doi.org/10.1093/biomet/70.1.41>
20. Rosenbaum P. R., Rubin D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39:1, 33–38 (1985). DOI: <https://doi.org/10.1080/00031305.1985.10479383>
21. Yun Xu, Royston Goodacre. 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* 2(3) (2017). DOI: <https://doi.org/10.1007/s41664-018-0068-2>
22. Rachel Lea Ballantyne Draelos. Best Use of Train/Val/Test Splits, with Tips for Medical Data. *Glass Box Machine Learning and Medicine*. [Online] <https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/>
23. Kevin Dobbin, Richard Simon. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4:31 (2011). DOI: <https://doi.org/10.1186/1755-8794-4-31>
24. Andrius Vabalas, Emma Gowen, Ellen Poliakoff, Alexander J Casson. Machine learning algorithm validation with a limited sample size. *Plos One* (2019). DOI: <https://doi.org/10.1371/journal.pone.0224365>
25. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," The elements of statistical learning. *Springer*, 2009, pp. 9–39.
26. Alpaydm, E. (2014). Introduction to machine learning. Cambridge, MA: MIT Press.
27. Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.
28. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," The elements of statistical learning. *Springer*, 2009, pp. 485–585.
29. Agnieszka Wosiak, Agata Zamecznik, Katarzyna Niewiadomska-Jarosik. Supervised and unsupervised machine learning for improved identification of intrauterine growth restriction types. *Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE (2016)
30. Hinton, Geoffrey; Sejnowski, Terrence. Unsupervised Learning: Foundations of Neural Computation. *MIT Press* (1999). ISBN 978-0262581684.
31. Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Ahmed J. Aljaaf, Abir Hussain. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and Unsupervised Learning for Data Science* (pp.3-21). DOI: https://doi.org/10.1007/978-3-030-22475-2_1
32. Osvaldo Simeone. A very brief introduction to machine learning with applications to communication systems. *arXiv preprint arXiv:1808.02342v4* (2018)
33. Hosmer, David W.; Lemeshow, Stanley (2013). Applied Logistic Regression. *New York: Wiley*. ISBN 978-0-470-58247-3.
34. Alan Agresti (2012). Categorical Data Analysis. Hoboken. *John Wiley and Sons*. ISBN 978-0-470-46363-5.
35. Chen, Tianqi; Guestrin, Carlos, "XGBoost: A Scalable Tree Boosting System". In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco*, 2016. pp. 785–794. arXiv:1603.02754. DOI: <https://doi.org/10.1145/2939672.2939785>
36. Hastie, T., Tibshirani, R., Friedman, J. H., "10. Boosting and Additive Trees". *The Elements of Statistical Learning (2nd ed.)*. *New York: Springer*. pp. 337–384 (2009)
37. Cochran, William G. (1952). The Chi-square test of goodness of fit. *The Annals of Mathematical Statistics*. 23 (3): 315–345. DOI: <https://doi.org/10.1214/aoms/1177729380>
38. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society. Vol. 85, No. 1* (1922), pp. 87-94. DOI: <https://doi.org/10.2307/2340521>
39. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval. Feature selection, Chi-Square feature selection *Cambridge University Press*. 2008
40. Chi-Square feature selection. "Scikit-learn" python library. [Online] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

41. Marketing, patient data, and privacy concerns. <https://www.reutersevents.com/pharma/commercial/marketing-patient-data-and-privacy-concerns>
42. Data insights. <https://prahs.com/healthcare-intelligence/data-insights>
43. Symphony Health Solutions. <https://symphonyhealth.prahs.com/>
44. Symphony Health Solutions, What we do. <https://symphonyhealth.prahs.com/what-we-do>
45. OBG Manag. Endometriosis and infertility: Expert answers to 6 questions to help pinpoint the best route to pregnancy. *Mdedge ObGyn* 27(6):30-35 (2015). <https://www.mdedge.com/obgyn/article/99912/surgery/endometriosis-and-infertility-expert-answers-6-questions-help-pinpoint/>
46. Jon k. Hathaway, MD, PhD, FACS. Decoding Coding. What is the Best Way to Code for Endometriosis? *NewsScope, volume 33, issue -2* (2019). <https://newsscope.aagl.org/volume-33-issue-2/decoding-coding-what-is-the-best-way-to-code-for-endometriosis/>
47. Fernando M. Reis, Larissa M. Coutinho, Silvia Vannuccini, Stefano Luisi & Felice Petraglia, Is Stress a Cause or a Consequence of Endometriosis? *Reproductive Sciences* volume 27, pages39–45(2020). DOI <https://doi.org/10.1007/s43032-019-00053-0>
48. Endometriosis – Risks, Signs, Symptoms, Diagnosis and Treatment
 - <https://www.mayoclinic.org/diseases-conditions/endometriosis/symptoms-causes/syc-20354656>
 - <https://www.webmd.com/women/endometriosis/endometriosis-causes-symptoms-treatment>
49. Bo Liang, Yang-Gui Xie, Xiao Ping Xu, and Chun-Hong Hu1. Diagnosis and treatment of submucous myoma of the uterus with interventional ultrasound. *NCBI, PMC Oncol Lett* (2018). DOI: <https://doi.org/10.3892/ol.2018.8122>
50. Endometriosis vs. Adenomyosis: Similarities and Differences <https://www.healthline.com/health/womens-health/adenomyosis-vs-endometriosis>
51. Endometrial Hyperplasia. <https://my.clevelandclinic.org/health/diseases/16569-atypical-endometrial-hyperplasia>.
52. Marina Kvaskoff, Andrew W Horne, Stacey A Missmer. Informing women with endometriosis about ovarian cancer risk. *The Lancet Journal, volume 390, issue 10111, P2433-2434, (2017)*. DOI: [https://doi.org/10.1016/S0140-6736\(17\)33049-0](https://doi.org/10.1016/S0140-6736(17)33049-0)
53. Aline Veras Moraes Brilhante, Kathiane Lustosa Augusto, Manuela Cavalcante Portela, Luiz Carlos Gabriele Sucupira, Luiz Adriano Freitas Oliveira, Ana Juariana Magalhães Veríssimo Pouchaim, Livia Rocha Mesquita Nóbrega, Thaís Fontes de Magalhães, and Leonardo Robson Pinheiro Sobreira. Endometriosis and Ovarian Cancer: an Integrative Review (Endometriosis and Ovarian Cancer). *Asian Pac J Cancer Prev. (2017) 18(1): 11–16*. DOI: <https://doi.org/10.22034/APJCP.2017.18.1.11>
54. John P. Cunha, DO, FACOEP. What Will Happen if Endometriosis Is not Treated? *emedicinehealth (2019) [online]* https://www.emedicinehealth.com/ask_what_will_happen_if_endometriosis_not_treated/article_em.htm#doctor%E2%80%99s_response
55. A. Michael Coppa, MD. What Happens if Endometriosis is Left Untreated? <https://www.drcoppaobgyn.com/blog/what-happens-if-endometriosis-is-left-untreated>
56. Endometriosis and ovarian cancer risk. [Online] <https://ovarian.org.uk/news-and-blog/blog/endometriosis-and-ovarian-cancer-risk/>

Figures

Figure 1

Healthcare Claims Patient Level Database Summary

Figure 2

presents the Receiver Operating Characteristic (ROC) curves on the test set for Logistic Regression and XGB models. The Area under the ROC Curve (AUC) had values between 0.88 -0.96. Chi-square test was also applied on data before standardization.

Figure 3

ROC curves of LR and XG models on test set

Figure 4

Distribution of probability on test data set for both LR and XGB model. Figure on right hand is of XGB and most of values are grouped at extreme values.

Figure 5

Distribution of patients by predicted probability score