**Endometriosis early detection**

Daniel Moshe, Adi Haber

A blue glowing uterus

Description automatically generated

# Table of Contents

[Table of Contents 2](#_Toc168925157)

[Background 3](#_Toc168925158)

[Problem Statement 4](#_Toc168925159)

[Proposed Solution 4](#_Toc168925160)

[Introduction 4](#_Toc168925161)

[Machine Learning 4](#_Toc168925162)

[Endometriosis 5](#_Toc168925163)

[UK BioBank [6] 5](#_Toc168925164)

[Solution Description 5](#_Toc168925165)

[Stage one: Feature Selection 5](#_Toc168925166)

[Stage Two: Planning Feature Extraction 11](#_Toc168925167)

[Stage Three: Analyzing the Data 12](#_Toc168925168)

[Stage Four: Feature Engineering 15](#_Toc168925169)

[Next Stages 17](#_Toc168925170)

[Data Imputations and Cleansing 17](#_Toc168925171)

[Further Feature Engineering 17](#_Toc168925172)

[Choosing a Machine Learning Model 18](#_Toc168925173)

[Model Evaluation 18](#_Toc168925174)

[Anticipated Challenges 19](#_Toc168925175)

[Acknowledgments 19](#_Toc168925176)

[Bibliography 20](#_Toc168925177)

# Background

Endometriosis, a chronic inflammatory condition, primarily manifests through symptoms such as pain and infertility [1]. It occurs when tissue resembling the uterine lining grows outside the uterus, adhering to pelvic organs, and occasionally other areas of the body. This abnormal adhesion triggers inflammation and the formation of scar tissue, resulting in debilitating pain and, in some instances, infertility. Endometriosis predominantly affects women of reproductive age, with research suggesting that approximately 5-10% of this demographic, totalling around 180 million individuals globally, are affected [2].

Main known indicators of the endometriosis include:

* Pelvic pain and/or lower abdominal pain
* Painful menstrual cramps
* Abnormal menstrual bleeding pattern (either by amount or irregularity)
* Family history [3]
* Infertility [4]

Endometriosis can be classified to 4 stages, based on the size and depth of the adhesions:

* Stage I: minimal
* Stage II: mild
* Stage III: moderate
* Stage IV: severe

The symptoms of endometriosis do not correlate with the stages, meaning Stage IV patients can have no symptoms and stage I patients can have severe symptoms.

Diagnosing endometriosis presents a challenge since adhesions are not always detectable through imaging techniques like ultrasound or MRI. Typically, a definitive diagnosis necessitates undergoing laparoscopic surgery [3] and a subsequent biopsy.

Problem Statement  
Our primary objective revolves around the prompt identification of endometriosis. Remarkably, 60% of women dealing with endometriosis navigate consultations with three or more clinicians before receiving a diagnosis, leading to an average delay of seven years before definitive diagnosis [5]. This prolonged delay intensifies symptoms, lowers overall quality of life, and contributes to enduring reproductive health challenges. Conventional diagnostic methods, predominantly reliant on invasive procedures and subjective assessments, further complicate the diagnostic process.

# Proposed Solution

This project endeavors to aid in diagnosing endometriosis by analyzing patient data. We will collect data on endometriosis and healthy patients from the UK Biobank and select a group of features (symptoms and risk factors) from which we will try to detect the existence of endometriosis. Our primary goal is to build the optimal machine-learning model for accurate endometriosis detection based on the features we found.

# Introduction

## Machine Learning

Machine learning, a subset of artificial intelligence, revolutionizes medical research by extracting insights from vast datasets to enhance diagnostic accuracy, treatment efficacy, patient outcomes, and identifying risk factors. There are two primary subcategories of machine learning - supervised and unsupervised learning.

Supervised learning algorithms use labeled data to train models to predict outcomes or classify instances, offering valuable insights into disease detection and prognosis.

Unsupervised learning techniques uncover hidden patterns within unlabeled data, enabling researchers to identify unexplored disease subtypes or biomarkers.

Deep learning, a subset of machine learning, utilizes neural networks with multiple layers to automatically extract complex features from raw data, paving the way for advanced image analysis, genomic sequencing, and complex medical issues.

With the integration of these machine learning paradigms, medical researchers unlock unprecedented opportunities to unravel the complexities of diseases, revolutionizing healthcare delivery.

## Endometriosis

Endometriosis, a prevalent chronic gynecological condition reliant on estrogen, concerns the presence of uterine endometrial tissue outside its normal cavity. This disorder is characterized by the presence of endometrial tissue outside the uterus, leading to pelvic pain and fertility issues.

## UK BioBank [6]

UK Biobank is a large-scale biomedical database and research resource, containing in-depth, de-identified genetic and health information from half a million UK participants. The database, which is regularly augmented with additional data, is globally accessible to approved researchers and scientists undertaking vital research into the most common and life-threatening diseases. UK Biobank’s research resource is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

# Solution Description

## Stage one: Feature Selection

The first step of our research was to find the features we would like to try and detect endometriosis by. We have sorted the features into groups:

**General Features**

The general features include:

* Sex/Genetic sex
* Age at recruitment
* Year of birth
* Weight/BMI

As endometriosis mostly effects women, we will use the sex data field to filter female patients.

We chose to include weight\BMI as features because we have found several research papers stating there is an inverse relationship between BMI and the risk of endometriosis, meaning endometriosis is more commonly associated with lower BMI [7].

The year of birth and age at recruitment features will help us infer which of the other features are relevant to our goal. Endometriosis mainly effects women of reproductive age, meaning data of patients over this age group may not be relevant.

**Pain Indicators**

One of the most recognizable symptoms of endometriosis is reported to be unusually painful periods, and pelvic pain in general. Research shows that 45% of patients with chronic pelvic pain also have endometriosis [8]. Other pain related symptoms that are usually associated with endometriosis include:

* Lower abdomen pain
* Back pain
* Hip pain
* Recurring Headaches and migraines
* Genital organ pain

We have added these symptoms, including pelvic pain and menstruation pain as features.

**Infertility**

Research shows that 30% of patients with infertility have endometriosis. Researchers believe the two are connected, meaning endometriosis might cause fertility issues in some of the patients, depending on the severity (stage) of the endometriosis [8].

Therefore, we expect that a diagnosis of infertility might also indicate a risk of endometriosis.

**Hormonal Contraception**

Hormonal birth control is one of the most common treatments for endometriosis. The hormonal contraceptive regulates estrogen levels. Endometriosis is estrogen-dependent therefore endometriosis patients tend to have excess estrogen. To stop estrogen production, meaning decrease the growth of endometrial-like tissue, Hormonal birth control is usually suggested as a treatment.

With this information in mind, we expect an inverse relationship between risk of endometriosis and using hormonal contraceptive. We also expect that a large portion of the patients diagnosed with endometriosis might also be taking a hormonal contraceptive.

**Menstruation**

In most patients with endometriosis, we can see an abnormal menstrual bleeding pattern, whether it is in amount (excessive bleeding for a longer duration than usual, meaning the menstrual cycle is shorter than 27 days) or in irregularity [1].

Moreover, the age of a patient at menarche seems to have a connection to endometriosis. Patients that started menstruating before the age of 12 have a higher risk of endometriosis.

The features we have decided to select in this category include:

* Age at menarche
* length of menstrual cycle
* menstrual pain/cramps

**Diet**

Some research papers show a connection between endometriosis and nutrition. It is shown that increased dietary fat consumption and low fiber consumption might increase estrogen concentrations, therefore increasing the risk of endometriosis. The same goes for red meat and dairy consumption [9].

On the contrary, vitamin D, C and E, and antioxidants in general are associated with lower risk of endometriosis, as they are anti-inflammatory [9].

The features we have selected to select for the stated dietary preferences for endometriosis:

* For dietary fat intake, we have selected fat consumption.
* For fiber consumption, we chose to look at fresh fruit and vegetables intake.
* For red meat consumption, we chose beef an processed meet intake.
* For dairy consumption, we chose lactose intake and cheese intake.
* For vitamins C, D, E we chose the corresponding vitamin intake features.

**Pregnancy**

Research shows a relationship between endometriosis and pregnancy complications. It has been suggested that endometriosis might change the uterine environment by causing progesterone resistance in the endometrium. This difficulties in embryo development and implantation [10].

We have decided to select as features some of the most common pregnancy complications:

* General complications
* Preterm delivery
* Spontaneous miscarriage/abortion
* Stillbirth
* placenta praevia (vaginal bleeding during pregnancy)

Moreover, many research papers indicate a relationship between delivery type and endometriosis: the rate of delivery by caesarean sections is higher in endometriosis patients. We have added delivery by caesarean section as one of our features as well.

**Mental state**

As stated above, pain is one of the most recognizable symptoms of endometriosis. The presence of pain significantly influences the psychological and social well-being of individuals with endometriosis, affecting their overall quality of life [12].

Mental symptoms seem to not get medical attention regarding endometriosis detection. We wanted to see if it is prevalent in endometriosis patients, so we added features regarding patient anxiety and depression.

**Related diseases**

Many research papers show a variety of diseases related to endometriosis. We have decided to choose some of the recurring ones, as well as some we were curious about.

The diseases we chose as features:

* Irritable bowel syndrome (IBS)
* Fatigue
* Anemia
* Lupus
* Fibromyalgia
* Hypothyroidism
* Allergies
* Blood clots
* Asthma
* Overian dysfunction
* Ovarian cancer
* uterine cancer
* cervical cancer
* Breast cancer
* Melanoma

After researching every feature group, we created a CSV file with all the features we selected from the UK Biobank showcase and their corresponding codes in the dataset. The goal of creating this CSV was to use it as input for the next step, planning the feature extraction.

Stage Two: Planning Feature Extraction

As a part of our research for this project, we decided to search for a python library used to extract features from the UK Biobank. We were surprised we did not find a generic library that takes a csv of the features and generated a pandas data frame. So, we have decided to create one.

The UK Biobank data is organized in CSV files, in which the columns correspond to the different data fields available. The data (at least the part of the data we have access to) is organized in two main CSV files, and each file has a “fields” file that describes which fields are present in it.

The library currently consists of two main classes: one for feature extractions, and one for basic imputations (as the UKB data is very sparse, we thought this might be a common use for this generic library).

The first class is UKBDatasetCreator. This class creates a pandas dataset from a csv file that contains the feature codes, as detailed in the UK Biobank showcase.

## Stage Three: Analyzing the Data

We Started by analyzing the gender distribution in our dataset, to ensure its relevance to the condition under study. Given that endometriosis mostly affects female patients, we identified and subsequently removed all male entries from our dataset. This left us with about 275,000 records, containing only female patient data.

A graph of patients by gender

Description automatically generatedFigure 1: female to male patient ratio in the dataset

At this stage, we excluded male patients from our dataset.

Next, we checked the prevalence of Endometriosis-affected women in the dataset. There are 10,171 patients that were diagnosed with endometriosis.

This revealed that the prevalence of Endometriosis-affected women in our dataset is approximately 4%. This distribution differs from existing research on the prevalence of endometriosis in the general population, at around 10%. This discrepancy might be because all women recruited were aged 40–69, meaning they were in the reproductive age range (15-49, as defined by the World Health Organization) in the late 90’s – early 2000’s, when there was less awareness of Endometriosis. Moreover the 10% distribution in the general population is an estimate based on sample tests, meaning even if 10% of the population has endometriosis, not all Endometriosis patients know they have it or are diagnosed with it. Despite this, we took several measures to minimize the effect of this discrepancy on our analysis, as described below.

A pink circle with yellow text

Description automatically generatedFigure 2: prevalence of Endometriosis in initial dataset.

To try and balance the rate of Endometriosis patients, we have decided that our train set will have a 50% ratio of Endometriosis patients, so that our model will not consider the prevalence of Endometriosis in the evaluation process. In the test set, we decided to set the prevalence of Endometriosis at 20%, slightly more than the prevalence in the general population, to account for the 10% being calculated from a sample.

Figure 3: prevalence of Endometriosis in test and train sets

Next, we checked the average diagnosis age. To calculate the average diagnosis, we used the features: date of first diagnosis, date of birth. We discovered that on average, Biobank patients were diagnosed at the age of 42. In the general population, an average woman will be diagnosed in her 30s. This poses a limitation of our data and consequently a future limitation of our model, as the model’s train set includes patients mostly diagnosed after the current average age of diagnosis.

A graph of age at diagnosis

Description automatically generatedFigure 4: diagnosis age counts

## Stage Four: Feature Engineering

Feature engineering is used in a machine learning project to improve model accuracy and efficiency. By transforming existing features, we can create new features we think will better the predictive power of our model.

**Endometriosis diagnosis feature**

The first feature transformation we did was to create a feature that will represent endometriosis diagnosis, and the source of the diagnosis. The UK Biobank has two separate features for this purpose, and they use different data encodings. The feature we determine endometriosis patients by is “[Date N80 first reported (endometriosis)](https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=132122)”. The data for this feature is dates (with a few special date encodings). We wanted to create a nominal feature that will include not only the diagnosis (has endometriosis or not), but also the source of diagnosis (medical diagnosis or self-diagnosis). We created a feature called “endo diagnosis”, a nominal feature with the following encoding:

* 0 – not diagnosed with endometriosis.
* 1 – endometriosis medical diagnosis.
* 2 – endometriosis self-diagnosis.

We decided to save the ‘date first reported’ feature to use it to gauge information about the patients age and the relevancy of other features to the diagnosis.

**Number of ICD-10 diagnoses**

Endometriosis is very hard to diagnose, and its symptoms are commonly mistaken for those of other medical conditions. Moreover, the Endometriosis diagnosis process is usually very long, taking 7-8 years on average, meaning many other illnesses can be diagnosed in this time. Knowing this, we assume that Endometriosis patients will have more IDC-10 diagnoses than non-Endometriosis patients. For this reason, we decided to add the number of ICD-10 diagnosis each patient has as a feature. We inferred this feature from the ‘hesin\_diag’ file, which contains icd-10 codes of different diagnosis each patient has. We grouped the data by patient id and added the number of diagnoses each patient has to our main dataset.

**Estrogen exposure**

One of the most common risk factors of Endometriosis is prolonged estrogen exposure [3]. Thus, we wanted to include estrogen exposure as a feature, with the assumption that longer estrogen exposure will correlate with an Endometriosis diagnosis.

We calculated Estrogen exposure (in years) by reducing the age of menarche (age at first menstrual cycle) from age of menopause (age after last menstrual cycle).

# Next Stages

## Data Imputations and Cleansing

The UK Biobank data is very sparse, containing numerous NaN values. For instance, one feature we wanted to analyse was whether the patient has anaemia, to explore a potential correlation between anaemia and endometriosis. However, upon extracting this feature, we found that only 256 women had documented data on anaemia diagnosis. During the data imputation and cleansing stage, we will decide whether to retain such features and impute the missing values or exclude these features. For the features we choose to retain, we will determine the appropriate imputation technique, whether it be median or mean imputation, or using a regression or clustering machine learning model. We plan to keep a copy of the un-imputed data to apply machine learning models capable of handling missing values.

Additionally, the Biobank data contains numerous errors and inconsistencies, such as patients with future first diagnosis dates or diagnosis dates before their birth. During the cleansing phase, we will identify and clean all such inconsistencies.

## Further Feature Engineering

After completing the data cleansing process, further feature engineering is essential to enhance the overall performance of our models. This step involves creating new features or modifying existing ones to better capture the underlying patterns and relationships within the data. For example, we might want to combine sparse features into a less sparse feature that will provide more information than each one individually. Moreover, after the data cleaning phase we might be left with a smaller number of features than anticipated, so we can use this stage to find and create new features. By performing thorough feature engineering, we aim to improve the model's ability to generalize to unseen patient data, ultimately leading to more accurate and reliable predictions.

## Choosing a Machine Learning Model

We are planning to try 3 main groups of models:

**Boosting Models:** Like XGBoost and CatBoost, which are proven to be good candidates for processing tabular data and can work with sparse data (data that contains many NaN values). As The UKB data is very sparse, we hope that these models will be able to generalize the data and give a good prediction.

**Classic Machine Learning Models:** Like logistic regression and SVM, on the imputed data.

**Deep Learning Models:** We wanted to try and see if a Neural network can identify complex patterns in the data, as Endometriosis is a condition with multiple different causes and symptoms, which might be hard for a traditional ML model to generalize.

After trying all 3 types, we will get to the evaluation stage, where we will choose the best model, and repeat earlier stages to improve our final predictions.

## Model Evaluation

After selecting the best model for predicting endometriosis, the evaluation stage becomes a cyclical and iterative process to refine and enhance the model's performance. This involves repeatedly assessing the model's accuracy, precision, recall, and other relevant metrics using a validation dataset. Based on the evaluation results, we might go back to modifying the feature set by adding new features, imputing missing values differently, or removing less impactful features. This iterative process helps in uncovering the most informative features and the best imputation strategies, thereby gradually improving the model's predictive power.

# Anticipated Challenges

In our research, we confront several notable constraints that shape the scope and reliability of our findings. Firstly, while utilizing the UK Biobank dataset provides valuable insights, its demographic skew towards women averaging 50 years old presents a limitation. Given that our objective is to assist in diagnosing endometriosis in younger women, the dataset may not fully represent the nuances of the condition in this demographic.

This fact introduces a layer of uncertainty regarding the accuracy of the diagnoses, as self-diagnoses may lack the certainty and precision of clinical assessments. Furthermore, the dynamic nature of endometriosis progression and treatment outcomes necessitates longitudinal data, which may be limited in our dataset. Despite these constraints, our research strives to navigate these complexities and contribute towards advancing the understanding and diagnosis of endometriosis.

# Acknowledgments

We thank Adi Shraibman, Dorit Shweiki, and Yonatan Bilu for their mentorship and support throughout the project.  
We thank The Academic College of Tel Aviv-Yaffo for providing us access to the UK BioBank data.  
We thank Grammarly and ChatGPT for grammar and tone suggestions.

# Bibliography

[1] Wang, P.-H., Yang, S.-T., Chang, W.-H., Liu, C.-H., Lee, F.-K., & Lee, W.-L. (2022). Endometriosis: Part I. Basic concept. Taiwanese Journal of Obstetrics and Gynecology, 61(6), 927–934. <https://doi.org/10.1016/j.tjog.2022.08.002>

[2] Zondervan, K. T., Becker, C. M., Koga, K., Missmer, S. A., Taylor, R. N., & Viganò, P. (2018). Endometriosis. Nature reviews. Disease primers, 4(1), 9. <https://doi.org/10.1038/s41572-018-0008-5>

[3] Blass, I., Sahar, T., Shraibman, A., Ofer, D., Rappoport, N., & Linial, M. (2022). Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach. Journal of Personalized Medicine, 12(7), 1114–1114. <https://doi.org/10.3390/jpm12071114>

[4] Tom Gunnar Tanbo, & Péter Fedorcsák. (2017). Endometriosis‐associated infertility: aspects of pathophysiological mechanisms and treatment options. Acta Obstetricia et Gynecologica Scandinavica, 96(6), 659–667. <https://doi.org/10.1111/aogs.13082>

[5] Horne, A. W., & Missmer, S. A. (2022). Pathophysiology, diagnosis, and management of endometriosis. BMJ (Clinical research ed.), 379, e070750. <https://doi.org/10.1136/bmj-2022-070750>

[6] UK BioBank (2015). Ukbiobank.ac.uk. <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us>

[7] Hong, J., & Yi, K. W. (2022). What is the link between endometriosis and adiposity?. Obstetrics & gynecology science, 65(3), 227–233. <https://doi.org/10.5468/ogs.21343>

[8] Mehedintu, C., Plotogea, M. N., Ionescu, S., & Antonovici, M. (2014). Endometriosis still a challenge. Journal of medicine and life, 7(3), 349–357. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4233437>

[9] Barnard, N. D., Holtz, D. N., Schmidt, N., Kolipaka, S., Hata, E., Sutton, M., Znayenko-Miller, T., Hazen, N. D., Cobb, C., & Kahleova, H. (2023). Nutrition in the prevention and treatment of endometriosis: A review. Frontiers in nutrition, 10, 1089891. <https://doi.org/10.3389/fnut.2023.1089891>

[10] Farland, L. V., Prescott, J., Sasamoto, N., Tobias, D. K., Gaskins, A. J., Stuart, J. J., Carusi, D. A., Chavarro, J. E., Horne, A. W., Rich-Edwards, J. W., & Missmer, S. A. (2019). Endometriosis and Risk of Adverse Pregnancy Outcomes. Obstetrics and gynecology, 134(3), 527–536. <https://doi.org/10.1097/AOG.0000000000003410>

[11] Afshari-Stasiak, S., Andrysiewicz, S., Andrysiewicz, A., Bamberska, J., Szubert, M. (2023). Rate of caesarean sections is higher in endometriosis patients – experience of single tertiary obstetric clinic in the light of epidemiologic data. J Pre Clin Clin Res., 17(2), 56-61. <https://doi.org/10.26444/jpccr/163481>

[12] Szypłowska, M., Tarkowski, R., & Kułak, K. (2023). The impact of endometriosis on depressive and anxiety symptoms and quality of life: a systematic review. Frontiers in public health, 11, 1230303. <https://doi.org/10.3389/fpubh.2023.1230303>