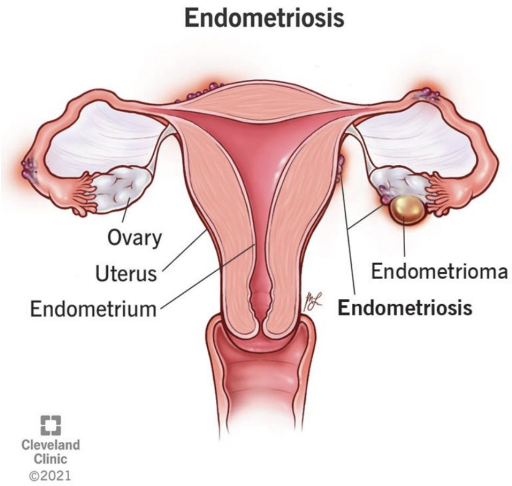
Leveraging Electronic Health Records -Derived Phenotypic Features for Early Diagnosis of Endometriosis Using Machine Learning

What is Endometriosis?

- ★ Chronic, autoimmune, inflammatory, estrogen-dependent, enigmatic, gynecological condition
- ★ Endometrium-like tissue growing outside of the lining of the uterus
- ★ Common symptoms: pelvic pain, painful periods, etc.
- ★ Common types of endometriosis
  - Peritoneal endometriosis (outer pelvic side walls)
  - Deep infiltrating endometriosis (inside the organs)
  - Ovarian endometrioma (on the walls of the ovaries)



Bulun SE, Yilmaz BD, Sison C, et al. Endometriosis. Endocrine Reviews. 2019;40(4):1048-1079. doi:https://doi.org/10.1210/er.2018-00242

Why is it considered an enigmatic condition?

- ★ Cause is unknown
  - Hypothesis: retrograde menstruation with endometrial cells
- ★ No ‘clinically reliable’ biomarkers to track its progression
- ★ No known cure
- ★ Noticeably ignored or misdiagnosed until later
- ★ Heterogeneous
  - Variety of symptoms and severity

Bulun SE, Yilmaz BD, Sison C, et al. Endometriosis. Endocrine Reviews. 2019;40(4):1048-1079. doi:https://doi.org/10.1210/er.2018-00242

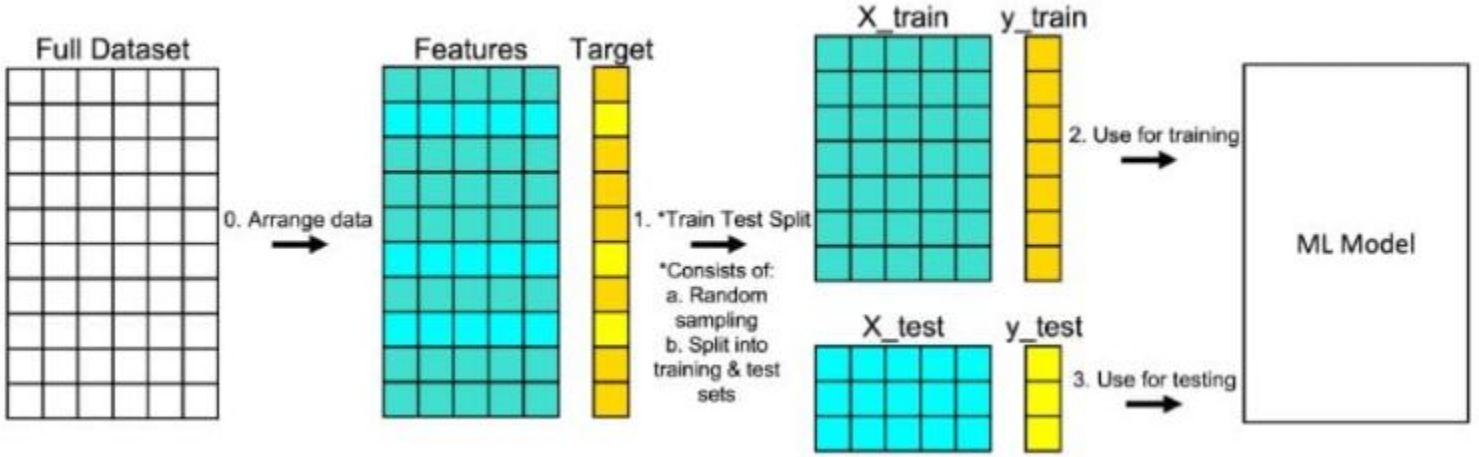
9.
Tomassetti C, Johnson NP, Petrozza J, et al. An international terminology for endometriosis, 2021,. *Human Reproduction Open*. 2021;2021(4). doi:https://doi.org/10.1093/hropen/hoab029

How can this be addressed?

1. Machine Learning
    a. Using classification machine learning algorithms to discover new ways to diagnose endometriosis without an invasive procedure
2. Computational Phenotyping
    a. Sorting patients into cohorts (with or without endometriosis) to try and understand the heterogeneity of the condition
        i. **Phenotype = period of time prior to diagnosis**
3. Feature Extraction
    a. Identifying and extracting relevant features from data within phenotype groups
        i. **Features = indicators of endometriosis**

Callahan TJ, Stefanksi AL, Ostendorf DM, et al. Characterizing Patient Representations for Computational Phenotyping. *AMIA Annual Symposium Proceedings*. 2023;2022:319-328. Accessed July 9, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10148332/
Bendifallah S, Puchar A, Suisse S, et al. Machine learning algorithms as new screening approach for patients with endometriosis. *Scientific Reports*. 2022;12(1). doi:https://doi.org/10.1038/s41598-021-04637-2

Process of Machine Learning



Pre-processed data set

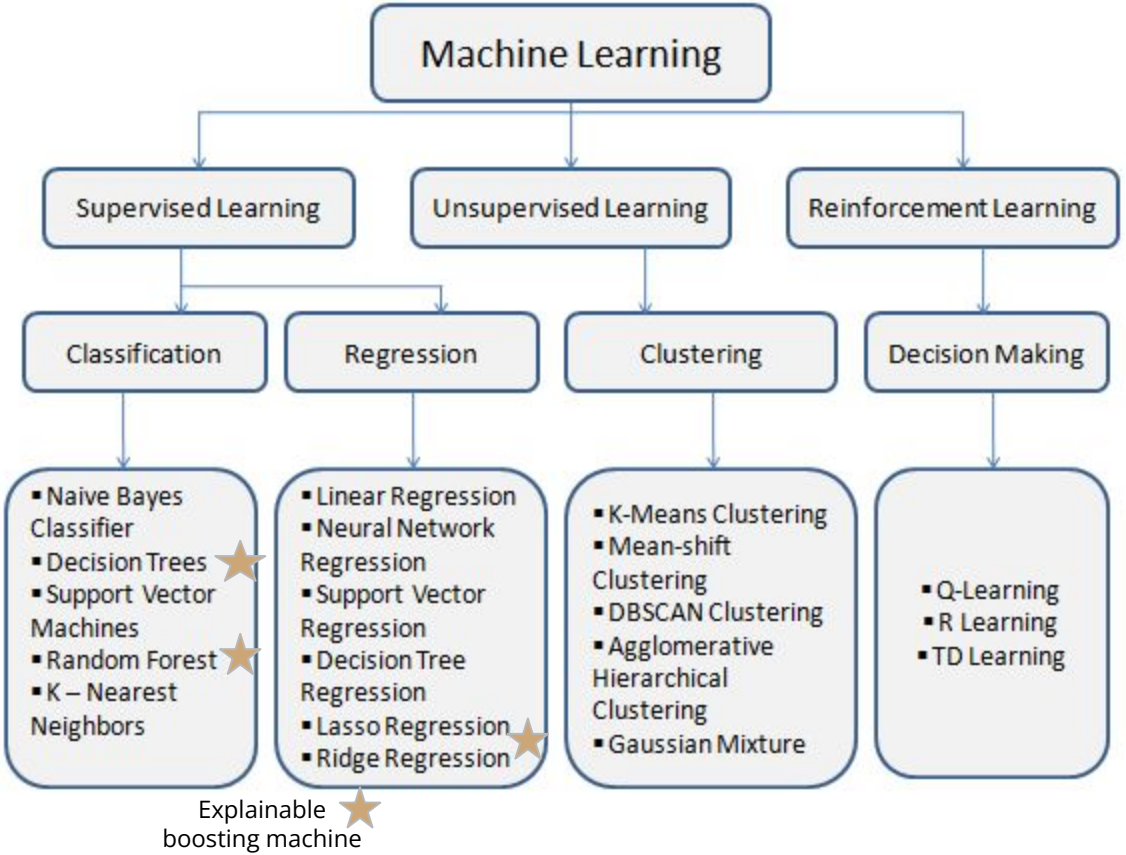
Highlighting what the machine is trying to learn (endometriosis vs. none)

Splitting into training and testing data sets

The objective: using machine learning models to extract key features that impact the phenotypic groups

Galarnyk M. Train Test Split: What it Means and How to Use It | Built In. builtin.com. Published July 28, 2022. https://builtin.com/data-science/train-test-split

Types of Machine Learning Models



Rajbanshi S. Machine Learning Algorithms | Introduction to Machine Learning. Analytics Vidhya. Published March 25, 2021. https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/

Choosing the best model: Classification vs. Regression

Classification:

- Mapping function from input to get output
- Discrete class labels
- Data needs to have labels first
- Can have both discrete and real-valued variables

Regression:

- Estimating the mapping function
- Quantitative value
- Prediction based on “features”

Gupta S. Regression vs. Classification in Machine Learning: What’s the Difference? Springboard Blog. Published October 6, 2021. https://www.springboard.com/blog/data-science/regression-vs-classification/

Machine Learning Models within this Project

- ★ Logistic Regression
- ★ Random Forest
- ★ XGBoost
- ★ Explainable Boosting Machine

What shows a model's effectiveness: Sensitivity, specificity, precision, negative predictive value, accuracy, ROC curve, etc.

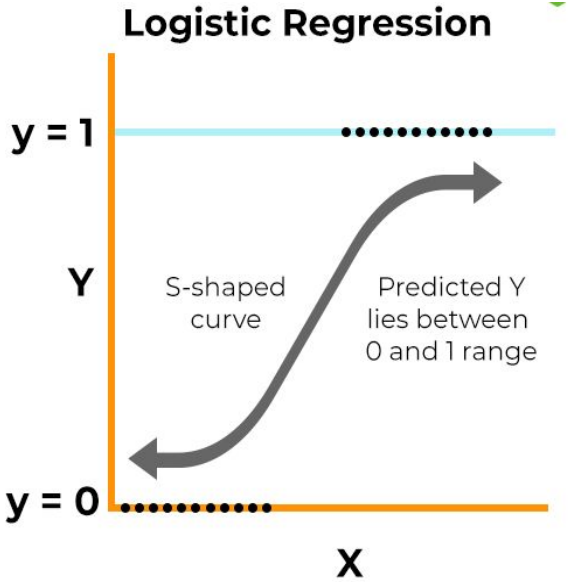
Logistic Regression

Using predictor variables (features) for binary prediction (endometriosis vs. no endometriosis)

Utilizes maximum likelihood function

Multivariable logistic regression
- Tests to see if the confounding variables' effect on the prediction is significant
- Model coefficient = holds weight to prediction
    - More weight = better predictor

Limitation(s): better for basic relationships



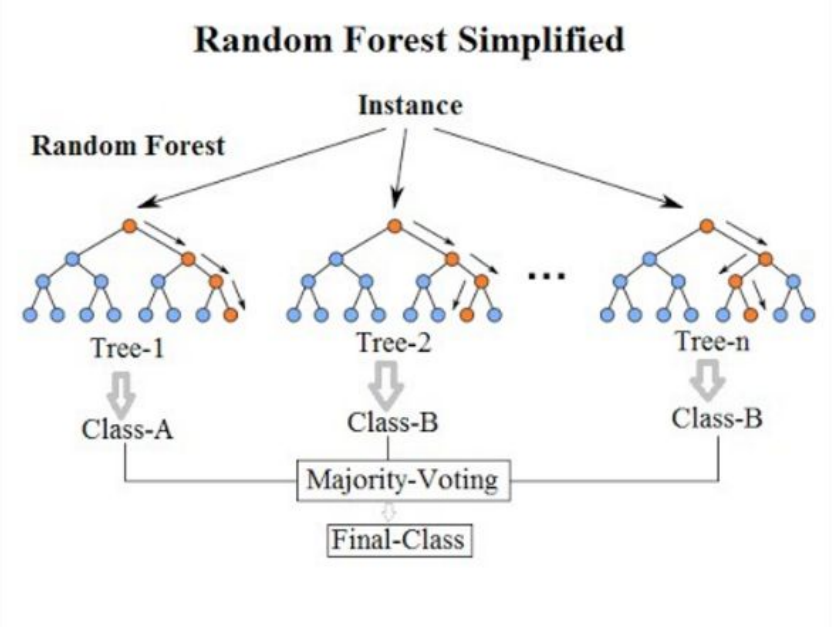
Meysami M, Kumar V, Pugh M, et al. Utilizing logistic regression to compare risk factors in disease modeling with imbalanced data: a case study in vitamin D and cancer incidence. *Frontiers in Oncology*. 2023;13. doi:https://doi.org/10.3389/fonc.2023.1227842

Rout AR. Advantages and Disadvantages of Logistic Regression. GeeksforGeeks. Published January 10, 2023. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

15.

Kanade V. What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices. Spiceworks. Published April 18, 2022. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

Random Forest



Model that maximizes the prediction of multiple decision trees

Factors

- Node size
  - Increasing node size makes simpler trees and tones out irrelevant features
- Number of trees
  - More trees = more accuracy
- Number of features

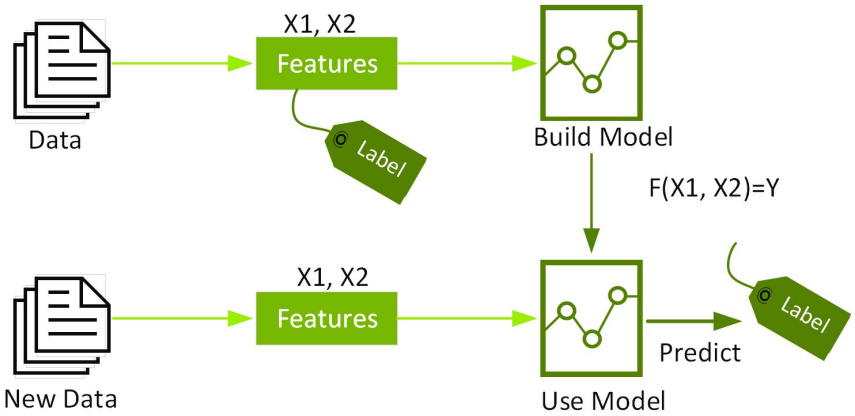
Bootstrapping and Aggregating

Limitation(s): more predictive trees results in slower run time

What does “node size” refer to in the Random Forest?. Cross Validated. Accessed July 10, 2024. https://stats.stackexchange.com/questions/158583/what-does-node-size-refer-to-in-the-random-forest
Random Forest Algorithms: A Complete Guide | Built In. builtin.com. https://builtin.com/data-science/random-forest-algorithm#procon
Bootstrap Aggregation, Random Forests and Boosted Trees | QuantStart. Quantstart.com. Published 2013. https://www.quantstart.com/articles/bootstrap-aggregation-random-forests-and-boosted-trees/

XGBoost

A collection of gradient boosted decision trees



Each tree is improved from the next by using the residual errors from the previous

- Lower the loss function of each tree

SHAP values = evaluating the features that promote the prediction

- Greater the value, the better the feature

Better for ‘non-linear’ data

Limitation(s): prone to overfitting, needs structured data

Nvidia. What is XGBoost? NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/xgboost/

Explainable Boosting Machine

Glassbox tree-based model best for human interpretation

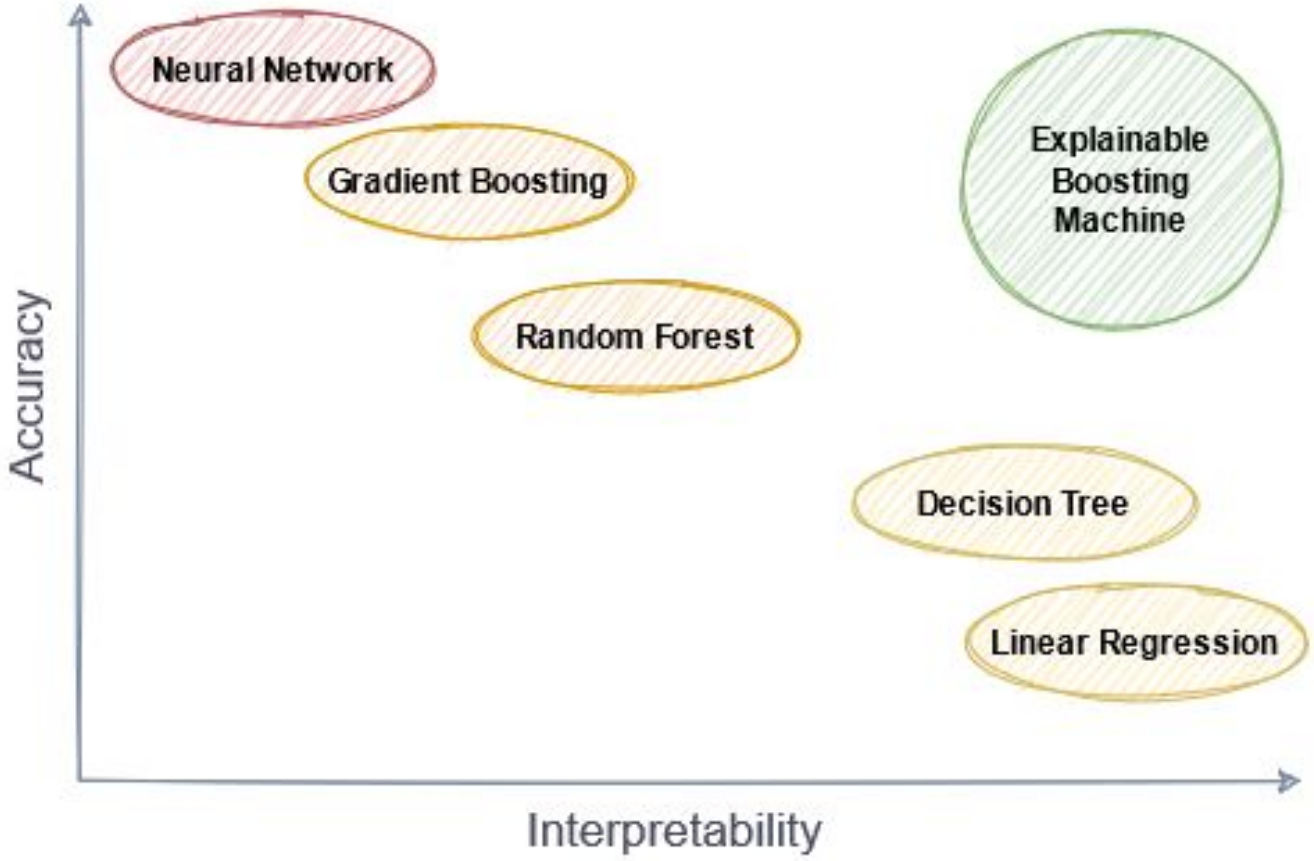
Uses gradient boosting and combines a number of decision trees (similar to XGBoost)

Generates “human-readable” results + provides ‘explanations’ for the predictions

Tests one feature at a time

Limitation(s): only explains the outer surface, not clear

Explainable Boosting Machine. interpret.ml. https://interpret.ml/docs/ebm.html
Kübler DR. The Explainable Boosting Machine. Medium. Published April 13, 2021. https://towardsdatascience.com/the-explainable-boosting-machine-f24152509ebb


Neural Network
Gradient Boosting
Random Forest
Explainable Boosting Machine
Decision Tree
Linear Regression
Accuracy
Interpretability

How the features are being interpreted

Phenotype groups:

1: 3 yrs - 2.5 yrs before diagnosis

2: 3 yrs - 2 yrs before diagnosis

3: 3 yrs - 1.5 yrs before diagnosis

4: 3 yrs - 1 yr before diagnosis

5: 3 yrs - 6 months before diagnosis

6: 3 yrs - Date of Diagnosis

Types of Features:

- Evident Features
- Prominent Features
- Progressive Features
- Eccentric Features
- Outliers

Prominent Features

Features that go directly with known symptoms/ideologies of endometriosis or point strongly to a diagnosis of endometriosis

****all found in the later groups; closer to the date of diagnosis****

- ★ Female Infertility
- ★ Single Live Birth
- ★ Pelvic Pain
- ★ Heart Rate/Diastolic Blood Pressure under anesthesia
- ★ Ultrasonography for fetal biophysical profile
- ★ Ultrasound
- ★ Cervical cytology
- ★

Evident Features

Features that have a direct impact on endometriosis diagnosis

****found in the vast majority of the groups****

- ★ Mean Platelet Volume
- ★ Carbon/Oxygen in Blood
- ★ Glomerular Filtration Rate
- ★ Glucose in Blood/Serum/Urine
- ★ Hemoglobin
- ★ Globulin
- ★ DNA Bacteria found by probe detection

Mean Platelet Volume

Leading feature - found in groups L1 - L5, X1, X3, X4

Measurement of the average size of blood platelets

Higher MPV = bigger platelets = faster platelet circulation = inflammation

Can be used as an ‘inexpensive’ biomarker

[22] Lale Bakir V, Dundar O, Bodur S, et al. Mean platelet volume as an inexpensive bio-marker of endometriosis. *Int J Clin Exp Med*. 2016;9(6):9431-9436. Accessed July 10, 2024. https://e-century.us/files/ijcem/9/6/ijcem0011915.pdf

Glucose in Serum/Urine/Blood

Serum - L1, X1
Urine - L3, X3
Blood - L5

- Endo patients often have lower glucose levels and higher insulin levels
  - Higher or lower than recommended range = risk of inflammation
  - High insulin (due to low glucose) = risk for type 1 diabetes (autoimmune)

Chen JP, Zhang YY, Jin JN, et al. Effects of dysregulated glucose metabolism on the occurrence and ART outcome of endometriosis. *European Journal of Medical Research*. 2023;28(1):305. doi:https://doi.org/10.1186/s40001-023-01280-7

Duffin J, read W min. How Dysregulated Blood Sugar Can Worsen Endo Symptoms. Endometriosis.net. Published March 9AD. Accessed July 10, 2024. https://endometriosis.net/living/diet-sugar

Carbon Dioxide/Oxygen in Blood

Carbon Dioxide - L1, L3, X1

- higher level indicates a higher hydrogen concentration (L1, x1)
  - Symptom: pH of arterial cord blood (L1, X1)

Oxygen -L2, L4, X2, X4

- insufficient amount = hypoxia
  - Direct correlation with endometriosis
- Venous cord: insufficient amounts can lead to abnormal SvO2 (L2, x2)
  - Too much oxygen can lead to increased risk of oxidative stress (indicator of endometriosis)

Donnez J, Mercedes M, Donnez O, Dolmans MM. Oxidative stress in the pelvic cavity and its role in the pathogenesis of endometriosis. ScienceDirect. Published October 2016. https://www.sciencedirect.com/science/article/pii/S0015028216625050#:~:text=Endometriosis%20is%20a%20disorder%20associated,known%20to%20have%20deleterious%20effec
Wu M, Hsiao K, Tsai S. Hypoxia: The force of endometriosis. *Journal of obstetrics and gynaecology research*. 2019;45(3):532-541. doi:https://doi.org/10.1111/jog.13900
Garcia AJ, Ramirez JM. Keeping carbon dioxide in check. *eLife*. 2017;6(e27563). doi:https://doi.org/10.7554/eLife.27563
Hu L, Li L, Li Y. Preliminary study on the effects of carbon dioxide and nitrogen pneumoperitoneums on endometriotic lesions. *Archives of Gynecology and Obstetrics*. 2012;286(2)
doi:https://doi.org/10.1007/s00404-011-2206-1

Glomerular Filtration Rate

Was at first positive in the first groups but then switched to negative as more time progressed (L1, L2, X1,)

Blood test that estimates how well the kidneys are filtering (per minute)

Glomerular Filtration Rate < 60 mL/min per 1.73 m^2 = chronic kidney disease

- Lower risk for women with endometriosis
  - Can be used to determine if the kidneys can be preserved during surgery for ureteral endometriosis

Ureteral endometriosis can result in loss of renal function = obstructed GFR

- Can result in high levels of urea nitrogen in serum/blood - L1, X1

Martínez-Zamora MA, Mensión E, Martínez-Egea J, et al. Risk factors for irreversible unilateral loss of renal function in patients with deep endometriosis. *Scientific Reports*. 2023;13(1):11940. doi:https://doi.org/10.1038/s41598-023-38728-z
Huang BS, Chang WH, Wang KC, et al. Endometriosis Might Be Inversely Associated with Developing Chronic Kidney Disease: A Population-Based Cohort Study in Taiwan. *International Journal of Molecular Sciences*. 2016;17(7):1079. doi:https://doi.org/10.3390/ijms17071079

Hemoglobin/Hematocrit

Hemoglobin - protein in blood that carries oxygen to cells - L4, X1

Hematocrit - percentage of red blood cells to total volume of blood - L4, L6

Blood markers are being used to help track the progression of endometriosis

- Lower levels of RBC indices = more severe endometriosis (after laparoscopy)

Excessive hemoglobin = promotes oxidative stress and inflammation

Cho HY, Park ST, Park SH. Red blood cell indices as an effective marker for the existence and severity of endometriosis (STROBE). *Medicine*. 2022;101(42):e31157. doi:https://doi.org/10.1097/MD.0000000000031157
Van Langendonckt A, Casanas-Roux F, Dolmans MM, Donnez J. Potential involvement of hemoglobin and heme in the pathogenesis of peritoneal endometriosis. ScienceDirect. Published September 14, 2001. https://www.sciencedirect.com/science/article/pii/S0015028201032113#:~:text=Result(s):%20Higher%20levels,hemoglobin
Nisenblat V, Bossuyt PM, Shaikh R, et al. Blood biomarkers for the non-invasive diagnosis of endometriosis. *Cochrane Database of Systematic Reviews*. 2016;(5). doi:https://doi.org/10.1002/14651858.cd012179

Bacterial pathogens found by Probe Detection

can be linked to inflammation

- ★ Streptococcus agalactiae - L2, X3, X4
  - Found more in women with endometriosis
- ★ Neisseria gonorrhoeae - L4
  - Interact with endometrial cells
  - Can cause pelvic inflammatory disease (linked to endometriosis)

Moreno I, Cicinelli E, Garcia-Grau I, et al. The diagnosis of chronic endometritis in infertile asymptomatic women: a comparative study of histology, microbial cultures, hysteroscopy, and molecular microbiology. *American Journal of Obstetrics and Gynecology*. 2018;218(6):602.e1-602.e16. doi:https://doi.org/10.1016/j.ajog.2018.02.012

Christodoulides M, Everson JS, Liu BL, et al. Interaction of primary human endometrial cells with Neisseria gonorrhoeae expressing green fluorescent protein. *Molecular Microbiology*. 2000;35(1):32-43. doi:https://doi.org/10.1046/j.1365-2958.2000.01694.x

Progressive Features

****Features that show an evident change as the date of diagnosis got closer****

- ★ features regarding medications and pain remedies grow in strength
  - Group 1 + 2: multivitamins, witch hazel pads vs. Group 5 + 6: benzocaine, lidocaine hydrochloride injection
- ★ more features regarding emotion/mental besides physical are added
  - Group 1: General examination vs. Group 5: Little interest or pleasure in doing things in the past 2 weeks
  - More severe endometriosis = worse quality of life
- ★ Features that lead to a certain condition with time
  - Presence of lymphocytes turning into lymphopenia

25.Moradi M, Parker M, Sneddon A, Lopez V, Ellwood D. Impact of endometriosis on women's lives: a qualitative study. *BMC Women's Health*. 2014;14(1). doi:https://doi.org/10.1186/1472-6874-14-123

Eccentric Features

****Features that seem irrelevant but still might have an effect on endometriosis diagnosis****

- ★ Body height - L1, L6, X1
  - taller women are at greater risk for endometriosis
- ★ Gestational age - L5, X2, X4, X5
  - Women with endometriosis are more likely to have premature births
- ★ Viruses
  - Influenza - endometriosis showing flu- like symptoms (L1, L2, X1)
  - Human papillomavirus - higher risk within women with endometriosis (L6, X5)
  - Cytomegalovirus - can infect endometrial cells
  - Hepatitis B - can lead to risk of endometrial carcinoma (
- ★ Women with endometriosis are more likely to be Rh-negative - L6, X3, X6
  - can lead to problems during pregnancy

Farland LV, Missmer SA, Bijon A, et al. Associations among body size across the life course, adult height and endometriosis. *Human Reproduction*. 2017;32(8):1732-1742. doi:https://doi.org/10.1093/humrep/dex207

Breintoft K, Arendt LH, Uldbjerg N, Glavind MT, Forman A, Henriksen TB. Endometriosis and preterm birth: A Danish cohort study. *Acta Obstetricia et Gynecologica Scandinavica*. 2022;101(4):417-423. doi:https://doi.org/10.1111/aogs.14336

Rees A. Endometriosis: how the condition may be linked to the immune system. *The Conversation*. https://theconversation.com/endometriosis-how-the-condition-may-be-linked-to-the-immune-system-203305. Published April 15, 2023.

Moslehi Z, Derakhshan R, Chaichian S, Mehdizadeh Kashi A, Sabet B, Rokhgireh S. Correlation of High-Risk Human Papilloma Virus with Deep Endometriosis: A Cross-Sectional Study. Silvestre S, ed. *BioMed Research International*. 2023;2023:1-6. doi:https://doi.org/10.1155/2023/6793898

Jiang XF, Tang QL, Zou Y, et al. Does HBV Infection Increase Risk of Endometrial Carcinoma? *Asian Pacific Journal of Cancer Prevention*. 2014;15(2):713-716. doi:https://doi.org/10.7314/apjcp.2014.15.2.713

Outliers

Measles

- Higher risk for women with endometrial cancer

Bordetella parapertussis

- No correlation with any gynecological issues?

Benharroch D, Kalinkovichi, Piura B, Shaco-Levy R, Gopas J. Evidence of measles virus antigens and RNA in endometrial cancer. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*. 2009;147(2):206-209. doi:https://doi.org/10.1016/j.ejogrb.2009.08.008

Thinking Bigger Picture...

Main categories: blood, kidney function, hormone production

- ★ Looking more into features related to blood (as an unofficial biomarker)
  - Endometriosis can cause a low red blood cell count - Indirect Bilirubin - L1, L2
- ★ Looking more into features that are hormone-related
  - Low levels of globulin resulting in higher levels of estrogen
  - Lower levels of mullerian inhibiting substance

Questions to explore further

- Why are features more prominent in one model than the other?
- If one model deems a factor as positive and the other deems it as negative, which one do we go with?
- If a feature goes from positive to negative, how do we attest it?
- Are there more prominent features that the models are not picking up on?

Conclusion

- ★ Endometriosis is an enigmatic, gynecological, condition.
  - Machine learning is paving the way for inexpensive and noninvasive ways of diagnosis
- ★ ML models, such as Logistic Regression and XGBoost, are being used to extract phenotypic features to better the understanding of endometriosis
- ★ Majority of the features can be organized into these categories: Prominent, Evident, Progressive, Eccentric, and Outliers