



This project has received funding from  
The European Union's Horizon 2020 research and innovation programme under  
Grant agreement No 101017562

**Finding Endometriosis using Machine Learning**  
**FEMaLe**

Call/Topic: Digital transformation in Health and Care  
Type of action: RIA

**Date: 07.02.2024**

|                        |  |
|------------------------|--|
| DELIVERABLE NUMBER     | D5.5   |
| DELIVERABLE TITLE      | Patient global profile                                   |
| RESPONSIBLE AUTHOR     | RTU  |
| GRANT AGREEMENT No.    | 101017562  |
| DOCUMENT TYPE          | Report   |
| WORKPACKAGE N.   TITLE | 5   BIG DATA: Digital health monitoring and the Lucy App |
| LEAD CONTRACTOR        | RTU  |
| AUTHOR(S)              | RTU, SEMM, YC LAB KFT, and AU                            |
| PLANNED DELIVERY DATE  | December 31 <sup>st</sup> , 2023                         |
| ACTUAL DELIVERY DATE   | February 7 <sup>th</sup> , 2024                          |
| DISSEMINATION LEVEL    | Public   |
| STATUS                 | Reviewed and quality checked                             |
| VERSION                | Final version (1.5)                                      |
| REVIEWED BY            | FEMaLE PMO   |

Document history

| Version | Date <sup>1</sup> | Comment   | Author | Status <sup>2</sup> |
|---------|-------------------|---|--------|---------------------|
| 1.1     | 27-11-2023        | First draft created   | RTU    | Drafted             |
| 1.2     | 13-12-2023        | Second draft created, including recommendations from WP participants. | RTU    | Drafted             |
| 1.3     | 22-12-2023        | Third draft prepared for FEMaLe Review Panel.                         | RTU    | Drafted             |
| 1.4     | 02-02-2024        | Final draft created, based on FEMaLe Review Panel feedback.           | RTU    | Completed           |
| 1.5     | 06-02-2024        | Final version ready for submission, quality checked by FEMaLe PMO.    | AU     | Validated           |

<sup>1</sup> As per the project’s cloud storage or per email date if applicable.  
<sup>2</sup> Drafted, completed or validated as per the project's cloud storage or per email date if applicable.

## TABLE OF CONTENTS

|  |    |
|--|----|
| 1. INTRODUCTION .....                            | 3  |
| 2. DATASETS DESCRIPTION .....                    | 3  |
| 2.1 DAILY LOG .....                              | 4  |
| 2.2 QUESTIONARY .....                            | 7  |
| 3. DATA FILTERING .....                          | 8  |
| 4. DATA ANALYSIS .....                           | 9  |
| 5. DIETARY HABITS .....                          | 10 |
| 6. COHORTS ANALYSIS .....                        | 11 |
| 7. PATIENT CLASSIFICATION AND RISK SCORING ..... | 14 |
| 7.2 PATIENT CLASSIFICATION .....                 | 15 |
| 7.3 SURVIVAL ANALYSIS .....                      | 16 |
| 8. MULTIMODAL APPROACH .....                     | 19 |

## Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

All FEMaLe Consortium members are also committed to publish accurate and up to date information and take the greatest care to do so. However, the FEMaLe Consortium members cannot accept liability for any inaccuracies or omissions, nor do they accept liability for any direct, indirect, special, consequential, or other losses or damages of any kind arising out of the use of this information.

## Copyright notice



This work by parties of the FEMaLe Consortium will be licensed under a Creative Commons Attribution 4.0 International License; <http://creativecommons.org/licenses/by/4.0/>

## Acknowledgement



The project 'Finding Endometriosis using Machine Learning' (FEMaLe) has received funding from the European Union Horizon 2020 programme under grant number 101000640.

## Citation

Be so kind as to cite this work as: Finding Endometriosis using Machine Learning, 2023: Patient global profile – under the supervision of the Project's Coordinator.

## Legislation

Legislation H2020 Framework Programme – Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ 347, 20.12.2013, p. 104).

Euratom Research and Training Programme (2014-2018) – Council Regulation (Euratom) No 1314/2013 of 16 December 2013 on the Research and Training Programme of the European Atomic Energy Community (2014-2018) complementing the Horizon 2020 – The Framework Programme for Research and Innovation (OJ L 347, 20.12.2013, p. 948).

H2020 Specific Programme – Council Decision 2013/743/EU of 3 December 2013 establishing the Specific Programme Implementing Horizon 2020 - The Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p. 965).

Rules for Participation (RfP) – Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 of December 2013 laying down the rules for the participation and dissemination in Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) (OJ L 347, 20.12.2013, p.81).

Financial Regulation (FR) – Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council of 25 October 2012 on the financial rules applicable to the general budget of the European Union (OJ L 298, 26.10.2012, p.1).

Rules of Application (RAP) – Commission Regulation (EC, Euratom) No 1268/2012 of 29 October 2012 on the rules of application of I Regulation (EC, Euratom) No 966/2012 of the European Parliament and of the Council on the financial rules applicable to the general budget of the Union (OJ L 298, 26.10.2012, p.1).

## 1. Introduction

The information for this work package (WP) comes from the *Lucy App*, which is a medical smartphone application. *Lucy* lets people enter details about their menstrual cycle, health, medical history, what they eat, how they live, and any pain they feel (Fig. 1). *Lucy* gathers information that users share about their symptoms of endometriosis, their background, mental and physical health, diet, and way of life.

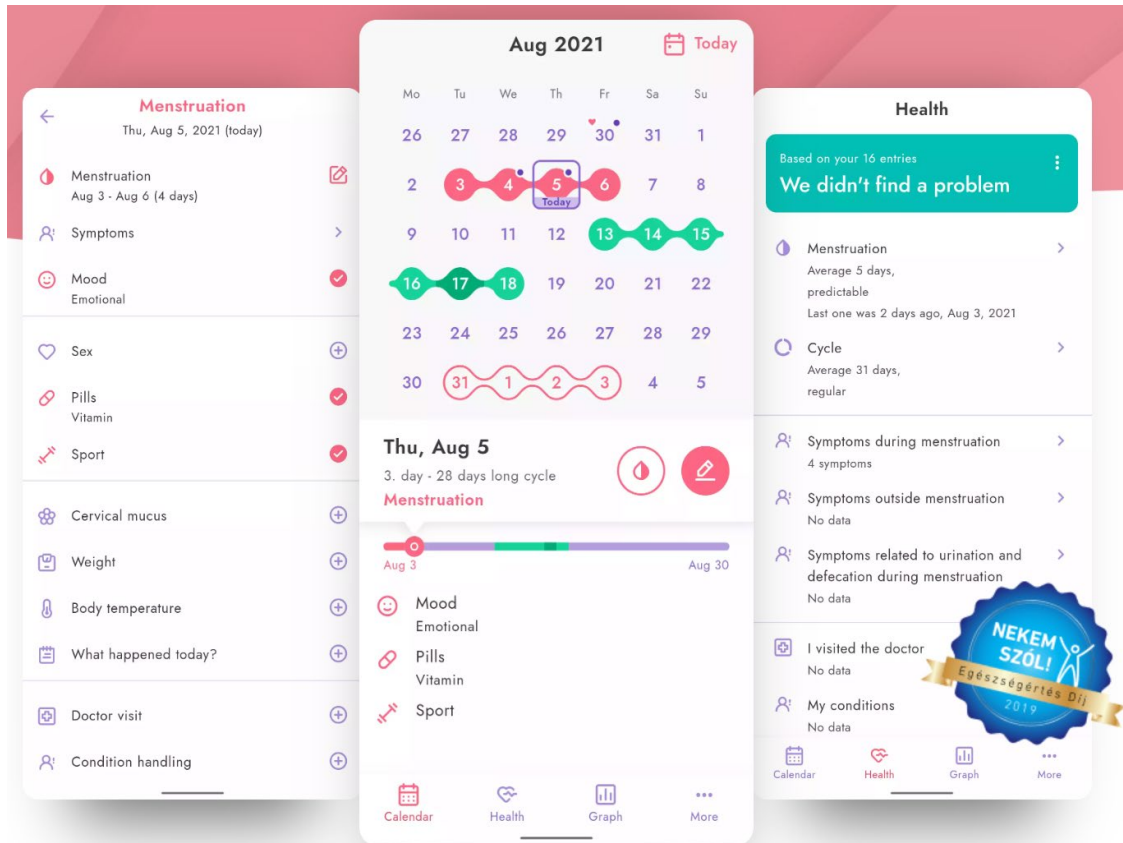


Figure 1: *Lucy App* main screen and daily log module.

By connecting this information with patients' records defining diagnoses, healthcare usage, and using a machine learning methods, an *Endometriosis Risk Score* was created for each patient. *Lucy* allows capturing a variety of additional data, such as dietary, mood, sports activities, and others, which will allow to develop more sophisticated diagnostic algorithms in the future. Current research in WP5 will focus on symptoms and their seriousness level along with specific dietary habits.

## 2. Datasets description

User data from *Lucy* is stored on a secure server in a PostgreSQL database. The relative nature of the database allows dynamically changing diseases, symptoms, adding questions and list of answers, as well as building a custom dietary to ease of entering recurring data. The following section will focus on two specific data sources: 1) daily log with symptoms, and 2) monthly questionnaire. The specific types were selected by medical scientists, using their knowledge and assumptions of most valuable data.

## 2.1 Daily log

Data is gathered from two tables, including 'daily registered symptoms' and 'diagnosed disease':

- The first table, 'symptoms', has the following fields: date of logged event, unique user id, symptom, and subjective evaluation of seriousness interpreted in a ten-point scale.
- The second table, 'diagnosed disease', holds: date when disease condition was changed, user id, disease name, and disease condition.

By using user id, it is possible to establish a relation between 'diseases' and 'symptoms' for each patient.

Total unique symptoms count is 42, having fewer common symptoms combined in 'other' type. Since total amount of unique users is close to 5.000, some of the symptoms are too rare (*Fig. 2*) for building reliable symptom to disease relation:

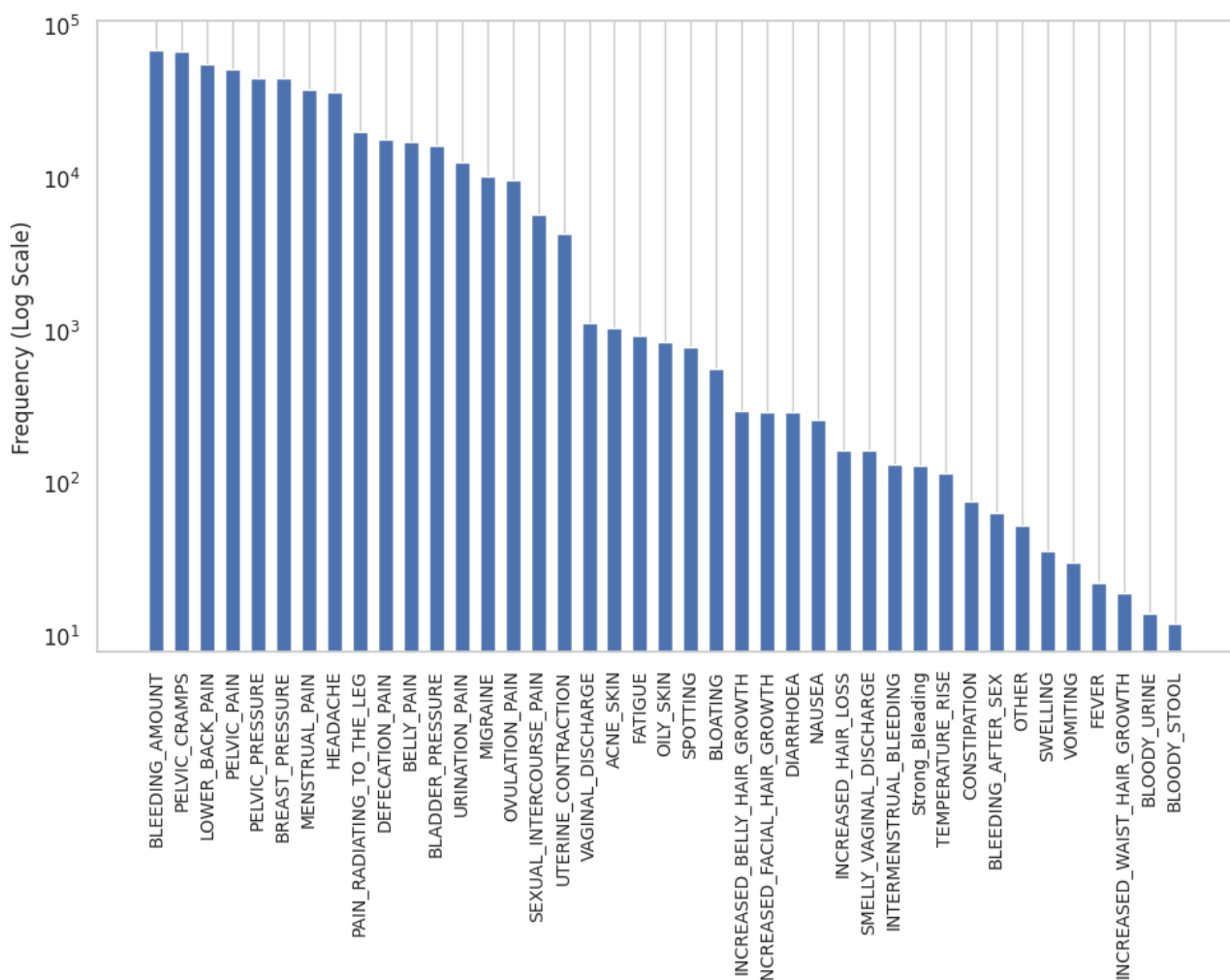
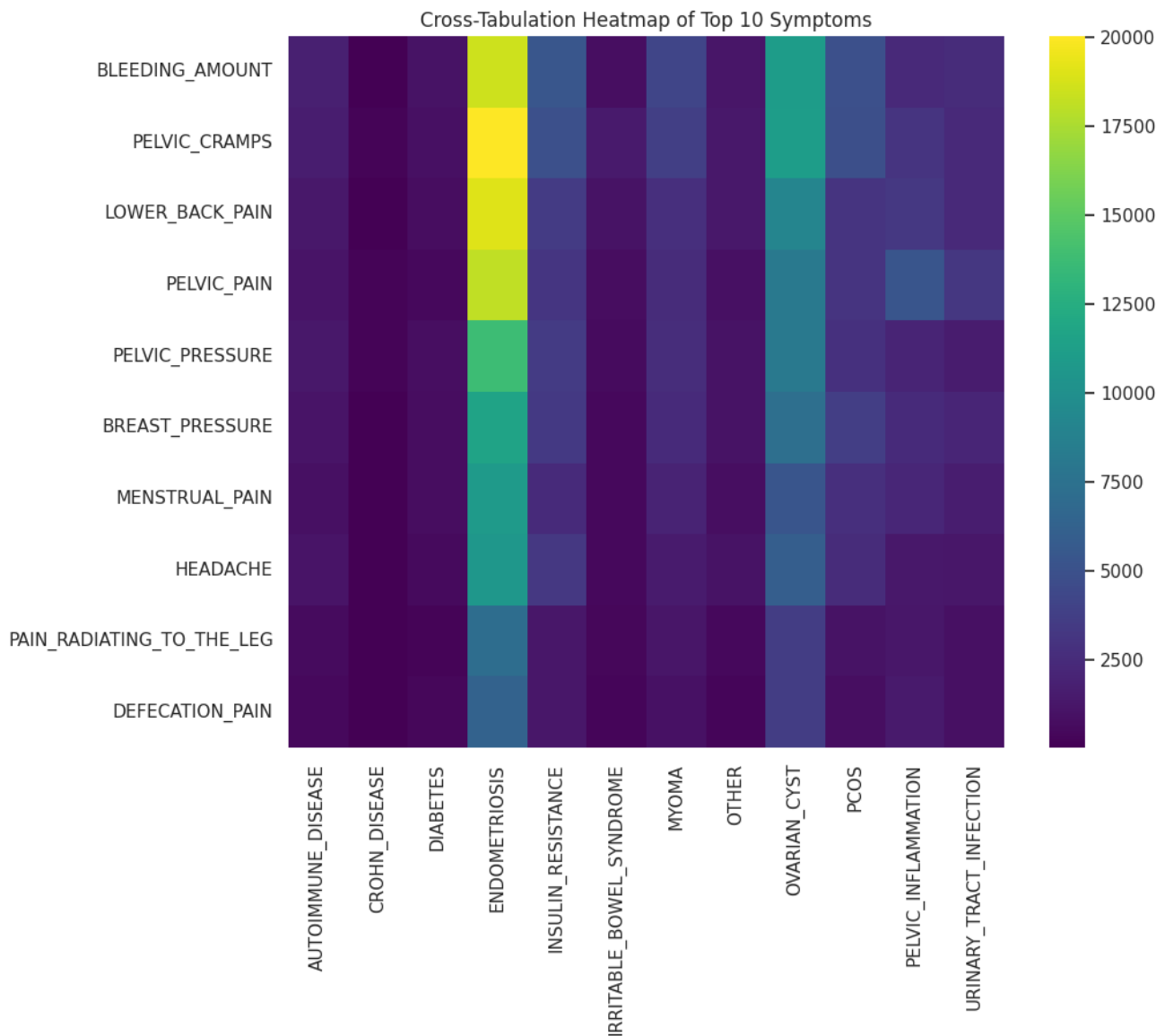


Figure 2: Frequencies of top ten symptoms.

To evaluate symptoms and disease correlation, a heatmap was constructed (*Fig. 3*). Plot shows medically plausible symptoms distribution. Ovarian cyst usually appears along endometriosis in practice; on *Fig. 3*, the distribution of the frequency of symptoms among these two diseases supports this medical correlation:

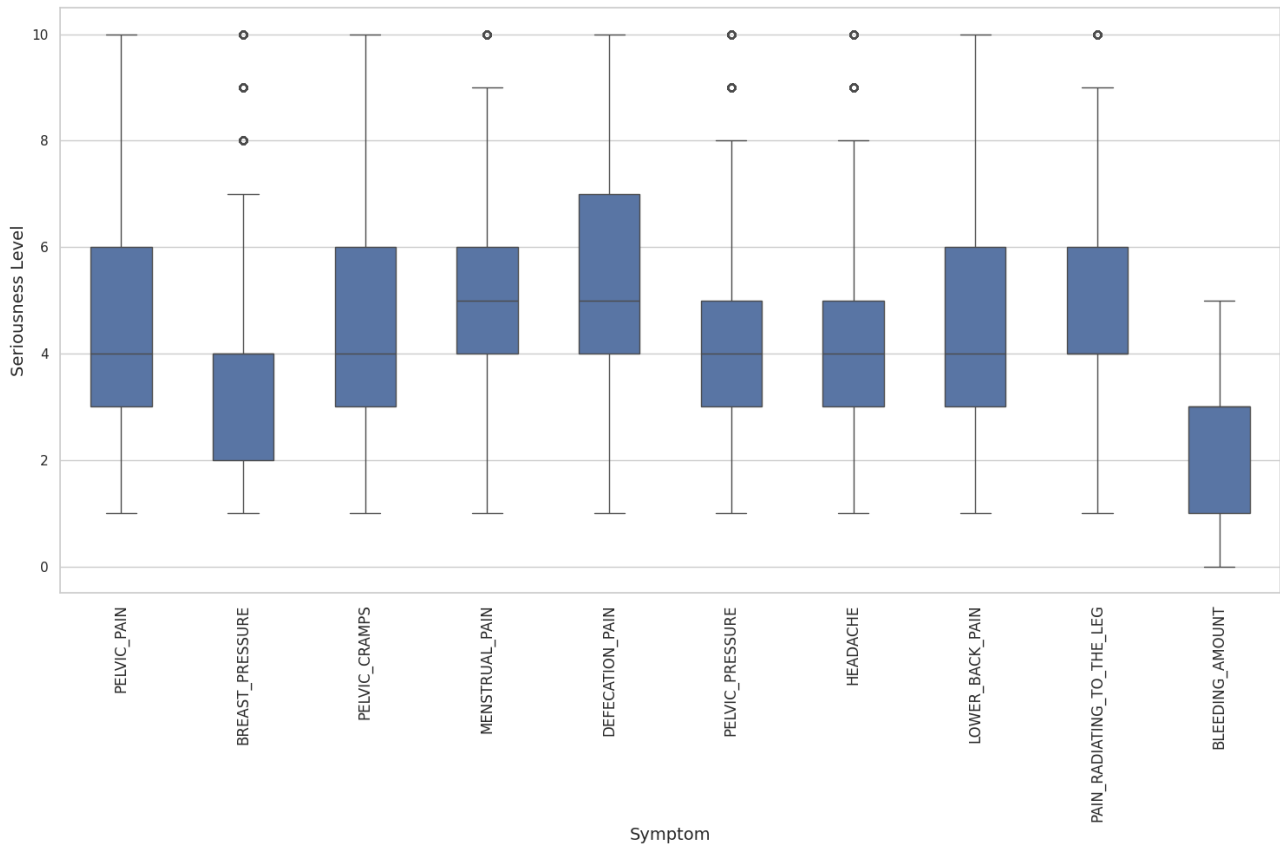


*Figure 3: Top ten symptoms in relation to diagnosed diseases. Colour defines frequency of each symptom.*

An interesting fact was observed during mapping between different languages. Since *Lucy* was developed in Hungary and has achieved strong support by local patients, a significant number of users come from Hungary. Early versions of *Lucy* had no automatic mapping of symptoms between languages, and data was exported in the original user language. Initially, this resulted in two groups of records: Hungarian and English.

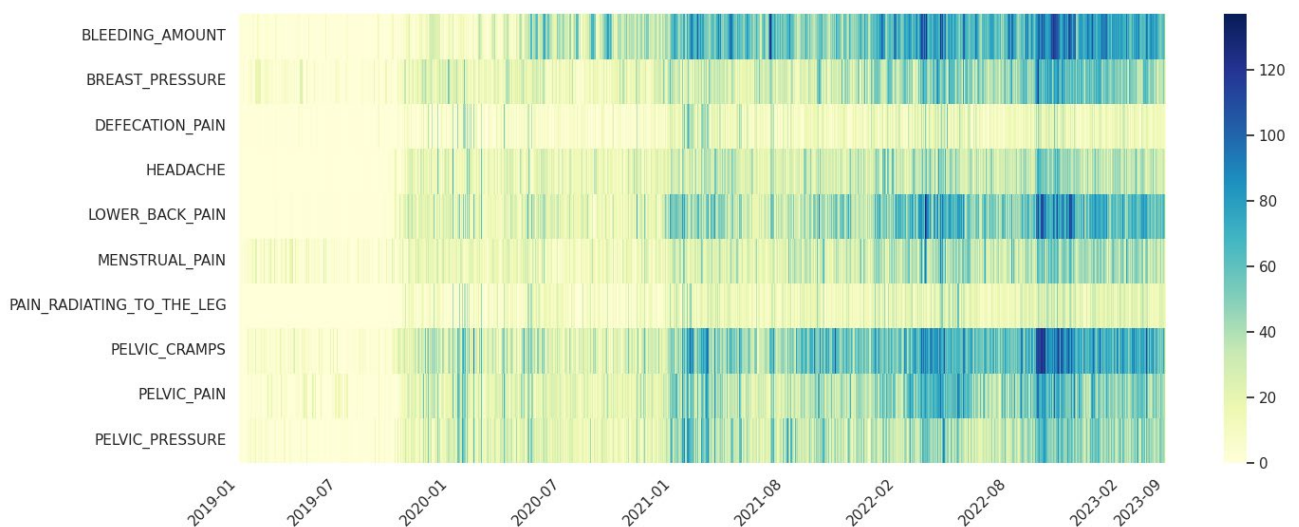
By observing symptoms statistics (*Fig. 3*), both groups (Hungarian and English) had almost identical frequency of symptoms correlated to disease. This supports that *Lucy* collected data is reliable and diseases share the same symptoms among diverse group of patients.

On *Fig. 4*, the distribution of the top ten symptoms and their level of seriousness can be seen. Most of them show realistic value distribution with a tendency to average value, due to the subjective nature of evaluation:



*Figure 4: Distribution of the top ten symptoms and their seriousness.*

On *Fig. 5*, the daily log records timeline shows plausible and realistic distribution. Several peaks (darker colour) are explainable by advertising campaign:



*Figure 5: Distribution of the top ten symptoms over time.*



## 2.2 Questionary

The main aim of the *Lucy* questionnaire module was to detect dietary habits and their correlation to endometriosis. Self-reported data of the patients were measured as follows:

- Assessing **pain scores** during the current or recent cycle, using the Visual Analogue Scale (VAS) in the questionnaire.
- Evaluating **quality of life** and **sexual function** in the four weeks prior to completing the questionnaire, using the Endometriosis Health Profile 5 (EHP-5).
- Examining **somatic** and **emotional symptoms** associated with **central sensitization**, using the Central Sensitization Inventory (CSI).

The structure of questionnaire tables and data is more complex than in the daily log (Section 3). To allow editing questions and answers, all answers and questions were encoded in single string variable, which allowed us to keep static SQL table structure, however, it also required additional decoding steps, as values cannot be filtered using SQL queries. Since questionnaire is filled once per month, records amount is lower than in daily log table.

Fig. 6 displays both *unfiltered* data (left) and filtered patient *without* endometriosis diagnosed (right). Since not every machine learning method could achieve a reliable result with unbalanced data, it is crucial during data analysis to evaluate the number of records being used:

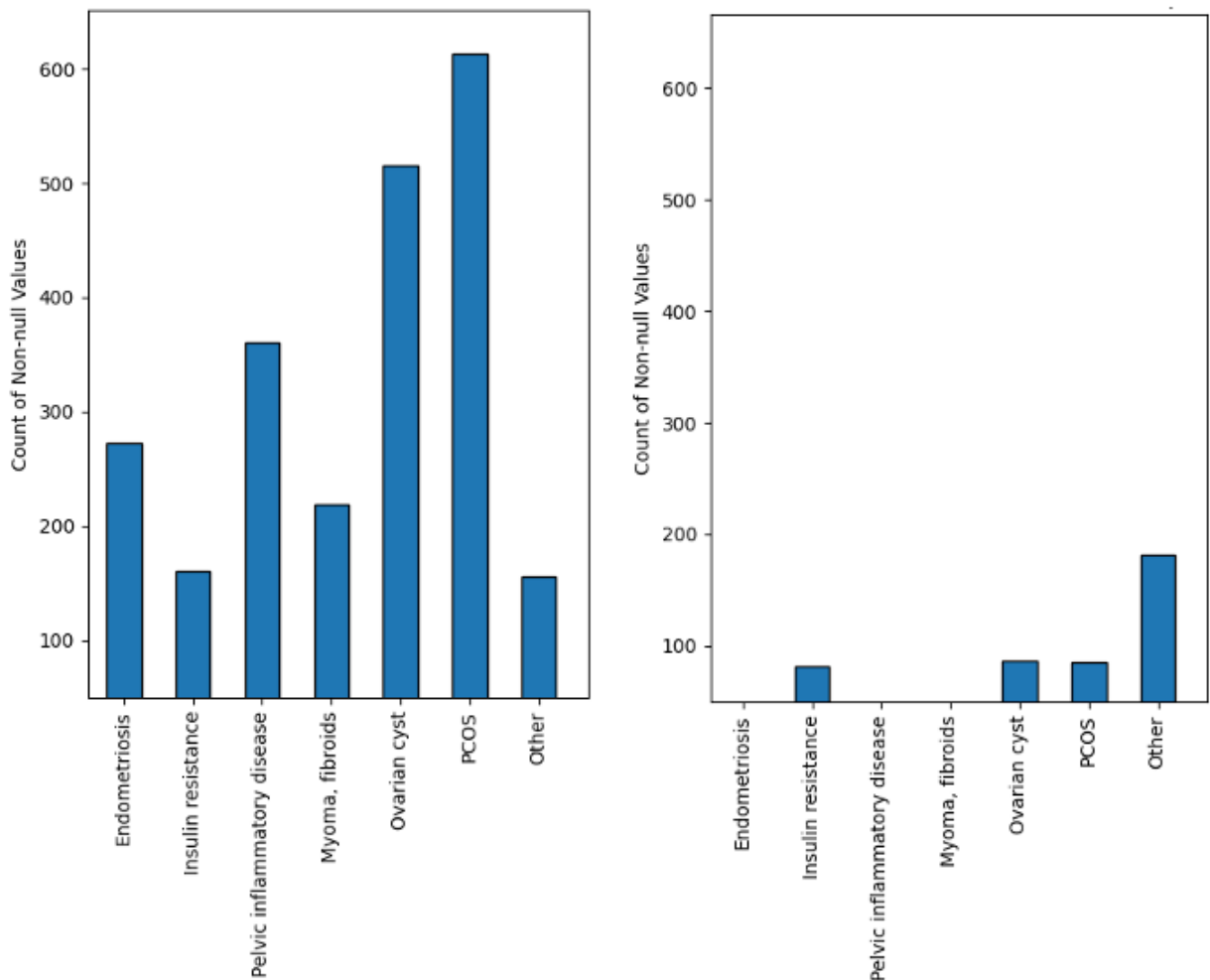


Figure 6: Records amount for each disease in questionnaire database. All records without filtering (left) and endometriosis non-diagnosed patient records (right), showing that some diseases are missing.

Some of the questions show high p-values (*Fig. 7*), meaning that some symptoms are very common and shared across different diseases. Nevertheless, this data is kept intact for further analysis, since machine learning algorithms can neglect insignificant data:

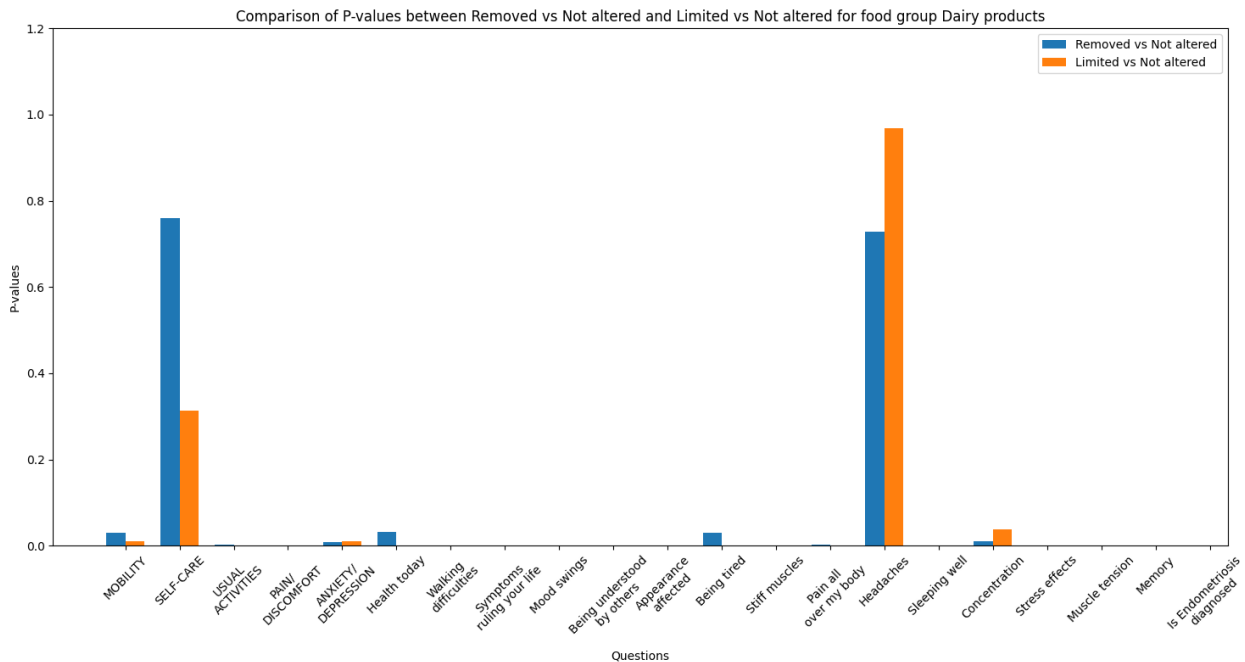


Figure 7: Self-assessed quality of life statistical analysis for three averaged patient groups with unaltered dietary habits to limited and restricted group.

### 3. Data filtering

Since some of the patients might enter erroneous data, in addition to the presence of unfiltered *Lucy* test data, records must be checked for logical compliance. For example, record date cannot be older than *Lucy* release for official use, and users with less than 30 records should also be removed, since there might be a bias during classification (*Fig. 8*):

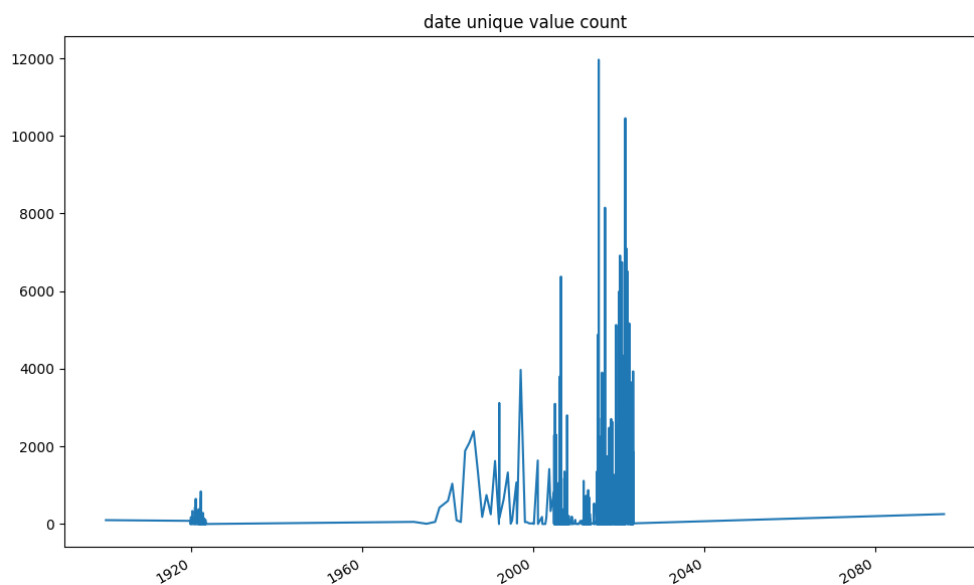


Figure 8: Records timeline, showing impossible log dates.

The initial number of *Lucy* records was over 700.000. After data cleaning, removing the abovementioned erroneous records and bias-risk users, the number of records decreased to around 500.000, which is still enough to study multi-symptom to multi-disease correlations and create a specific disease risk score.

The distribution of the number of records (*Fig. 9*) for each disease, 'chronic diseases' omitted, shows that we have access to an adequate amount of data for studying correlations and building cohorts. It is interesting that the studied disease – **endometriosis** – has the most records, which requires attention to dataset splitting during machine learning:

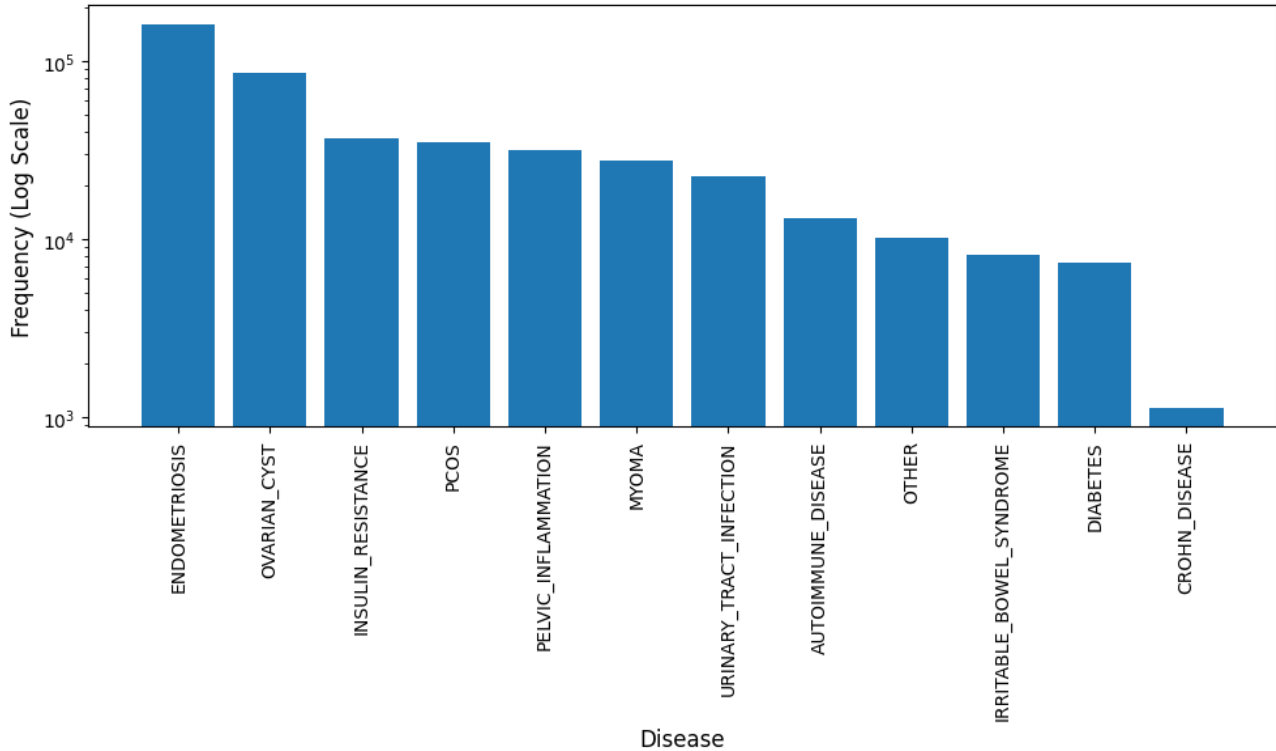


Figure 9: Number of records for each disease.

## 4. Data analysis

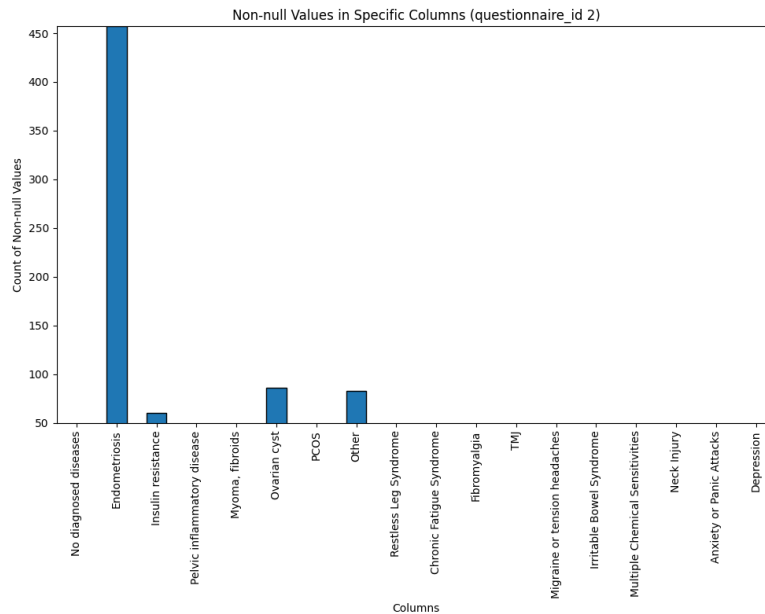
In addition to defining an *Endometriosis Risk Score*, we aim to establish *Endometriosis eCohorts* with the intention of analysing correlation between dietary habits and endometriosis. Both the *eCohorts* and *Risk Score* calculations were created, based on the daily records database (Section 2.1), since it holds enough records and is statistically significant. The dietary analyses were based on the questionnaires data (Section 2.2), as it holds exact and specific products (e.g., red meat, sugar) usage habits records.

To realise these aims, datasets should be split by disease/state; as a result, the following datasets were made:

- **Endometriosis group:** patients *with* diagnosed endometriosis, however, not cured.
- **Potential group:** patients *not* diagnosed with endometriosis, suffering from one of following diseases: 'Ovarian Cyst', 'Autoimmune Disease' or 'Irritable Bowel Syndrome'.
- **Negative group:** patients *not* having any of the mentioned diagnoses, suffering from one of following diseases: 'Myoma', 'PCOS', 'Insulin Resistance' or 'Diabetes'.
- **Others:** patients having *other* diagnoses.

## 5. Dietary habits

The effect on quality of life of dietary habits should be compared in groups. After filtering each dataset, the number of records for each disease (*Fig. 10*) was analysed and displayed:



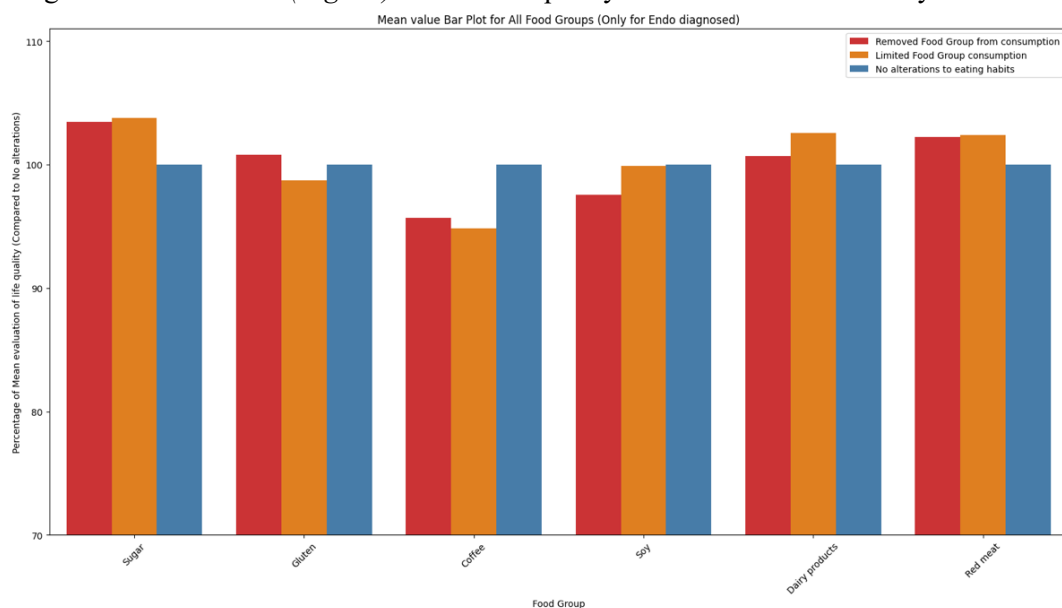
*Figure 10: Number of records for each disease in questionnaire dataset, filtered by patients with diagnosed endometriosis.*

To analyse dietary habits, questionnaire has three levels of certain food intake habits:

- No alteration made in specific food type for last period.
- Intake was limited, but not excluded.
- Food was completely removed.

Questionnaire holds 42 questions, 20 specifically related to quality of life, for six food groups.

First rough estimation reveal (*Fig. 11*) overall life quality effects of different dietary habits:



*Figure 11: Average quality of life effect of different dietary habits for endometriosis diagnosed group.*

As seen on *Figure 11*, there are **no significant difference** in the quality of life. Given the evaluation is done subjectively and score is an average of 20 specific effects, no major difference can be found; the biggest one is around 5% related to a coffee intake.

To analyse specific dietary effects on quality of life, the same calculation as in *Fig. 11* was performed for each product and on each answer, averaging by users. *Fig. 12* shows the potentially significant correlation of reducing red meat intake for endometriosis patients:

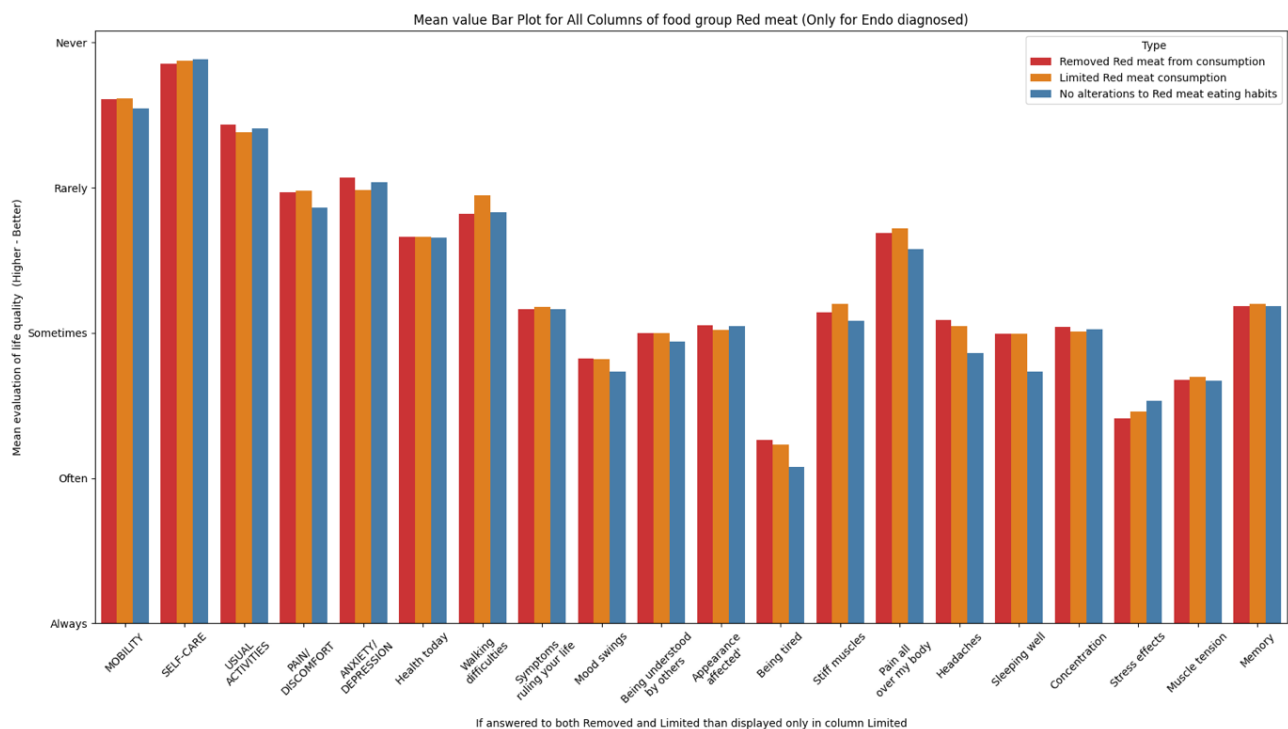


Figure 12: Average quality of life effect of specific food for endometriosis diagnosed group.

As seen on *Fig. 12* above, there are **no major difference** in quality of life; most prominent effect of reducing red meat is on *walking difficulties*. The effects will be studied further and might show stable difference in some specific dietary habit, when the number of users will be higher.

## 6. Cohorts analysis

There are known difficulties with strict patient groups separation, due to the lack of reliable information on undiagnosed patients. To deal with this, patient cohort analysis was performed. There are different clustering methods that allow visualising patient groups, when no strict classes can be defined and their amount unknown. UMAP (Uniform Manifold Approximation and Projection) is especially good in situations, where it is hard to clearly define group classes, because some patients might have a disease that has not yet been diagnosed. UMAP is a technique that helps to see patterns in large sets of data by simplifying and reducing the dimensions of the data, making it easier to visualise and analyse. Traditional methods might struggle to organise this data, because they rely on clear labels or categories, like 'diagnosed' and 'not diagnosed'. However, diseases often do not follow such neat lines, especially in early stages or when symptoms overlap with other conditions. One of the big benefits of using UMAP is its ability to bring out clusters or groups of patients with similar characteristics, even if we do not have clear labels for those groups. For example, it might reveal a cluster of patients whose symptoms and test results suggest they could have the examined disease, even though they have not been diagnosed.

On Fig. 13, a combination of different UMAP clustering results is shown. Data is coloured using three colours to help identify the predefined groups that are clustered closely. **Red** colour shows endometriosis diagnosed patients, **blue** is patients without any disease known to be related to endometriosis, and **green** is the potentially patients with endometriosis:

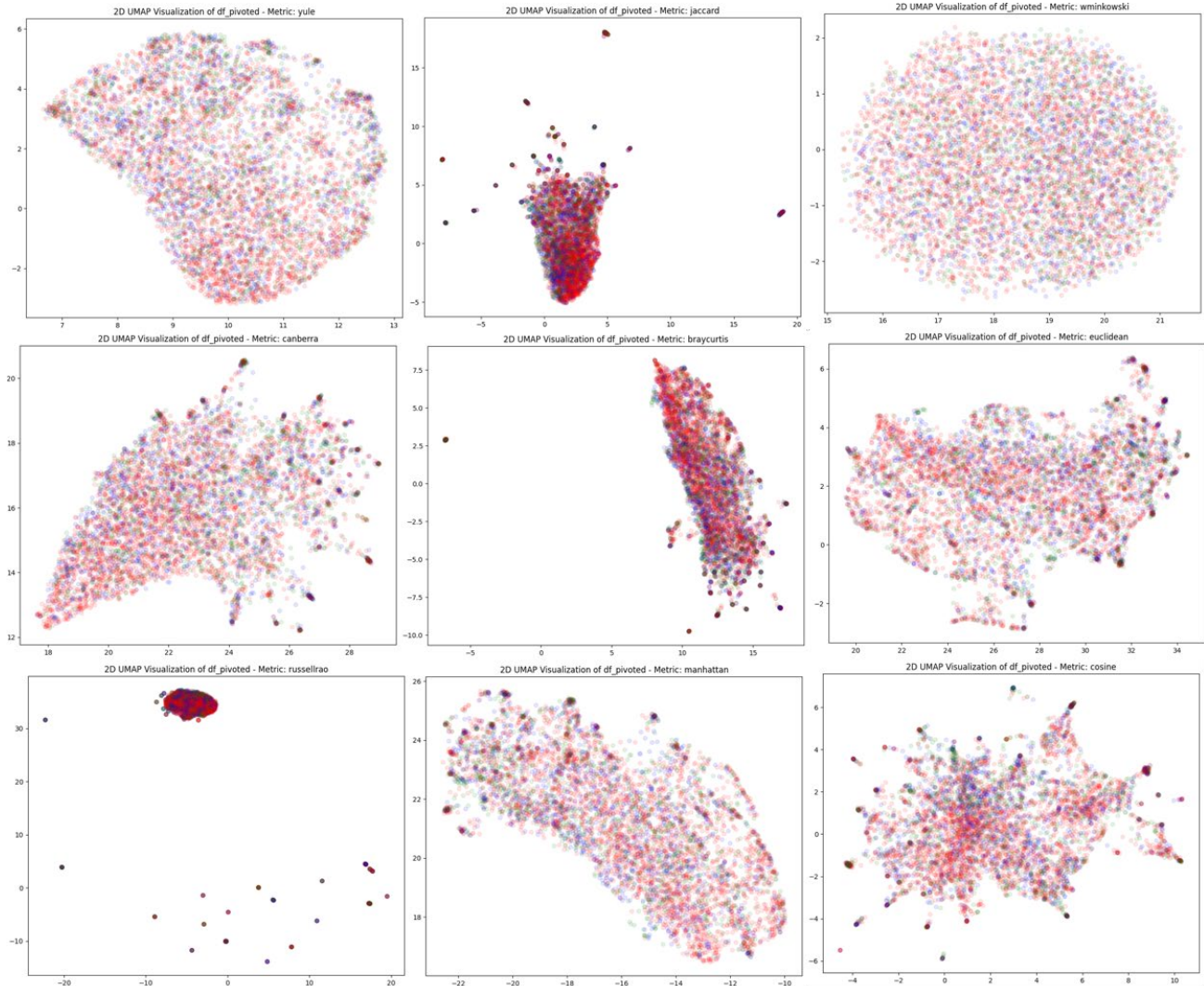


Figure 13. Patient clustering result, using UMAP method.

The effectiveness of UMAP in uncovering hidden patterns and structures within such datasets is largely influenced by its key parameters: 1) the metric used for calculating distance, 2) the number of neighbours, and 3) the minimum distance between points in the reduced space.

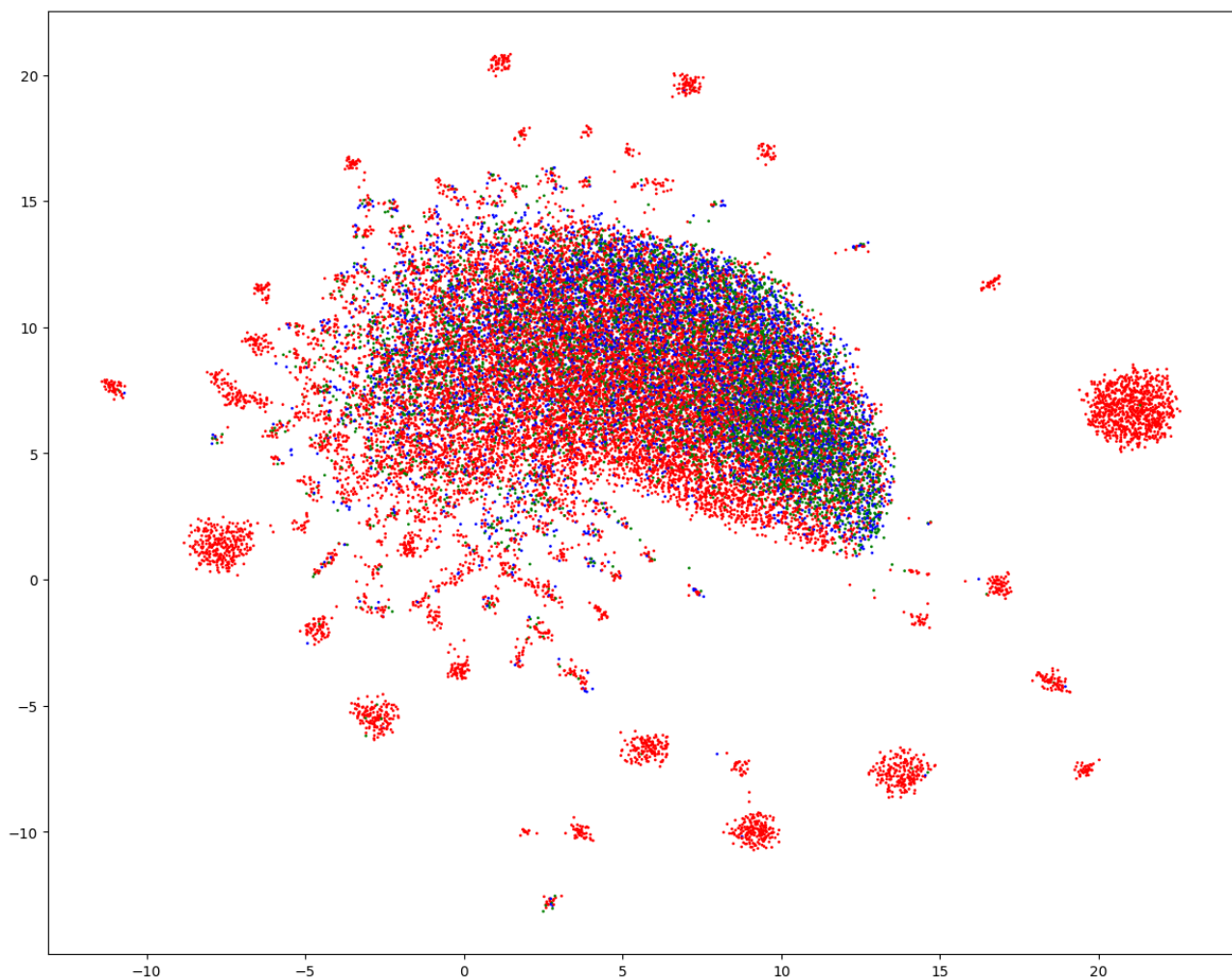
The *metric parameter* plays a crucial role in determining how distances between points are calculated in the high-dimensional space. Different metrics, like Euclidean, Manhattan, or cosine similarity, can dramatically change the resulting structure in the reduced dimensionality space. For example, in a dataset containing patient symptoms, the Euclidean distance might group patients, based on the overall similarity of symptom intensity, while cosine similarity could focus on the pattern of symptoms, such as which symptoms occur together, irrespective of their intensity.



The *number of neighbours parameter* affects how UMAP balances the dataset's local and global structure. A higher value considers a larger local neighbourhood for each point, emphasizing the dataset's global structure, whereas a lower value digs in on more detailed local structures. In the context of health data, a smaller value might reveal specific groups of patients with similar symptom patterns, possibly indicating different variants of an undiagnosed disease. Conversely, a larger value might identify broader distinctions, like the separation between patients with clear diagnostic signs and those potentially undiagnosed.

Finally, the *minimum distance parameter* influences how closely points can cluster together in the reduced space. A smaller value allows for tighter clustering, making it easier to distinguish between closely related groups, such as patients with different stages of an undiagnosed disease or similar but distinct conditions. A larger value, on the other hand, encourages a more spread-out representation, which might help in visualizing the overall structure of the data but could blur finer distinctions between patient groups.

As seen on *Fig. 13*, there are **no obvious clusters** in most of the trials. Nevertheless, by selecting visually the best metric and carefully adjusting parameters, some of the clusters are revealed. *Figure 14* shows distinct clusters of diagnosed patients as well as partial separation of red and blue dots, proving that there are certain patient groups that could be separated, based on their symptoms:



*Figure 14: Patient clustering result, using UMAP method, with dice metric.*

## 7. Patient Classification and Risk Scoring

Early attempts included basic statistical analyses and straightforward machine learning techniques to understand the relationships between individual symptoms and diagnoses. However, the weighted accuracy was around 60% and could *not* be used for reliable *Risk Score* calculation. Thus, a more sophisticated approach was needed to accurately evaluate patient risk. On *Figures 15* and *16*, Cramer's V correlation score is plotted:

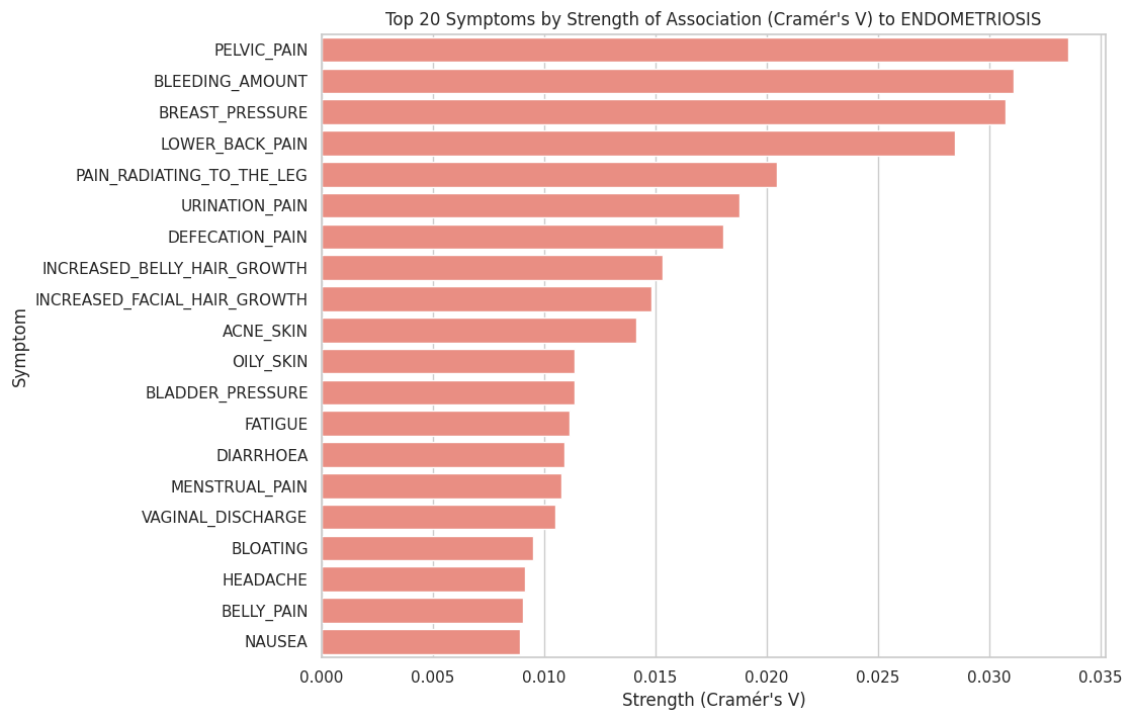


Figure 15: Cramer's V correlation of endometriosis group symptoms.

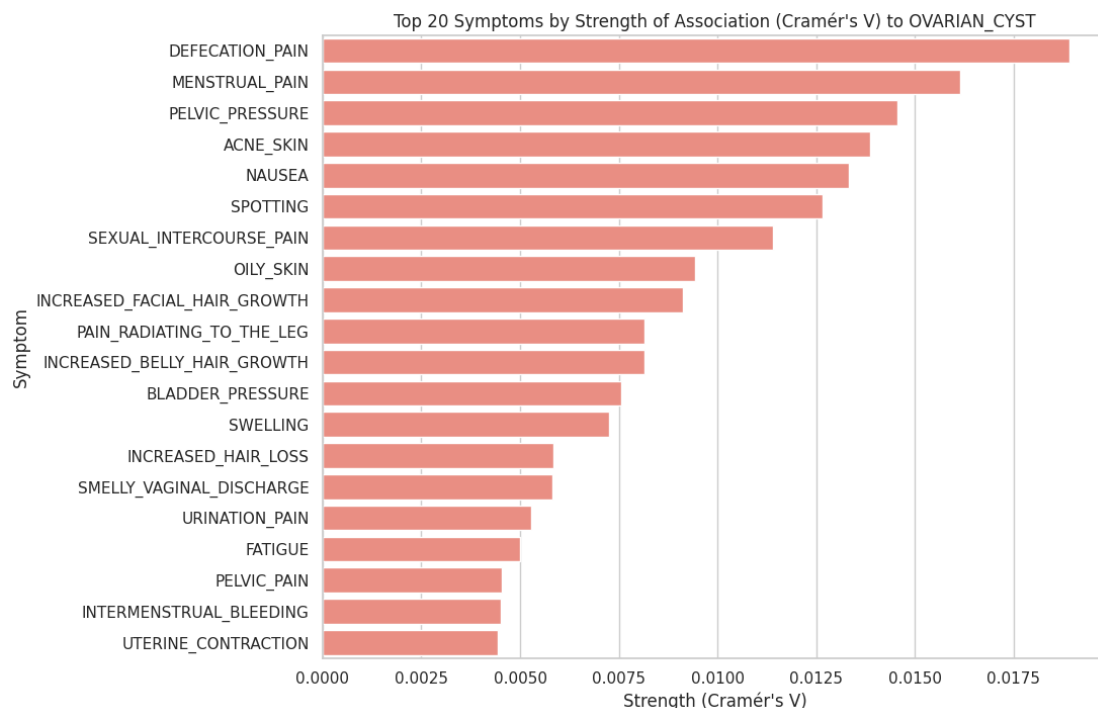


Figure 16: Cramer's V correlation of endometriosis group symptoms.



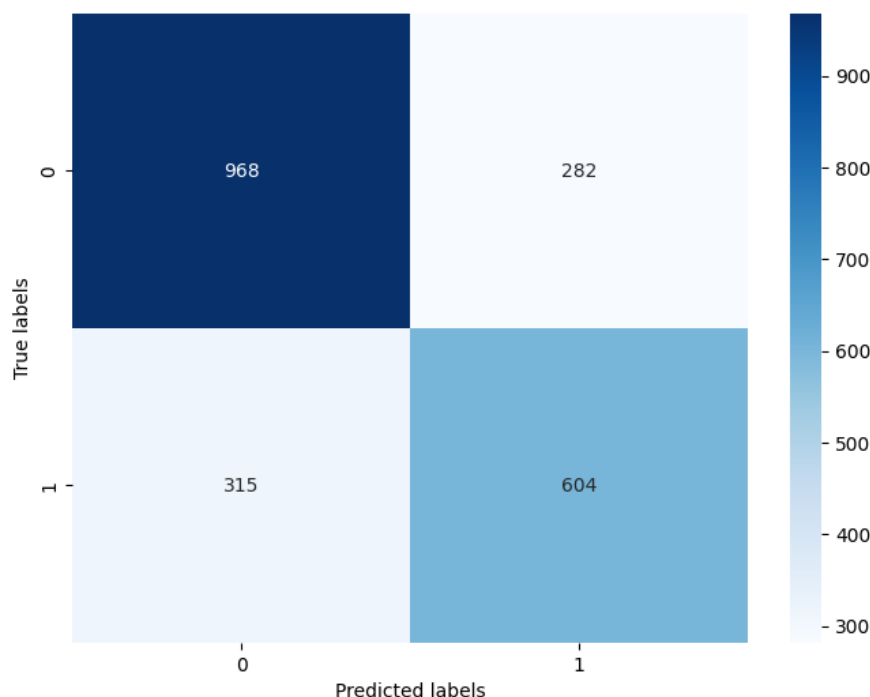
## 7.2 Patient classification

Overall correlation looks realistic to endometriosis and ovarian cyst diagnoses. Nevertheless, the correlation value seems too low to be used in a straightforward manner for further score calculation. This led to the adoption of a methodology that assessed symptoms not just by their occurrence but also by their severity, with this data being compiled monthly. The significance of each symptom in predicting a diagnosis was recognised, leading to the utilisation of a 'seriousness score' to each symptom. By multiplying the frequency of each reported symptom by its 'seriousness score', a weighted score for each symptom was generated. This score provided a more accurate representation of the symptom's significance in the context of the patient's overall health. These weighted scores were then aggregated on a monthly basis. The daily weighted scores for each symptom were summed up over a month to obtain a composite monthly score.

This approach allowed for the identification of broader trends in symptom presentation and their effects, while minimising the distraction of daily symptom variability. Monthly aggregation was found to be advantageous, offering a clear and manageable framework for analysis without overwhelming detail. In the table below, the accuracy of four machine learning model performances is shown. Despite having relatively high F1 score, by class performance is still low (~66%), making methods useless in a straightforward manner:

| Classifier   | ROC AUC | F1   |
|--------------|---------|------|
| XGB          | 0.66    | 0.83 |
| RandomForest | 0.62    | 0.81 |
| KNeighbors   | 0.60    | 0.79 |
| Perceptron   | 0.50    | 0.68 |

By applying hyper-parameter search over all patient dataset (*Fig. 17*), RandomForest classifier show higher results than XGBoost, reaching weighted score around 72%:



*Figure 17: RandomForest classificatory confusion matrix.*

Nevertheless, according to cluster data shown, there are many overlapping patients and it is not possible to accurately separate healthy patients. To help in solving this problem, datasets were filtered, to keep patients with high number of records, assuming that such patients have deep motivation, data is accurate, and symptoms are stable. In the table below, results of RandomForest classification are shown for users with at least 300 daily records. Final score increased to 83% which is adequate accuracy to build a risk scoring for a specific user:

|              | PRECISION | RECALL   | F1-SCORE |
|--------------|-----------|----------|----------|
| 0            | 0.905405  | 0.842767 | 0.872964 |
| 1            | 0.731183  | 0.829268 | 0.777143 |
| ACCURACY     | 0.838174  | 0.838174 | 0.838174 |
| MACRO AVG    | 0.818294  | 0.836018 | 0.825054 |
| WEIGHTED AVG | 0.846126  | 0.838174 | 0.840361 |

### 7.3 Survival analysis

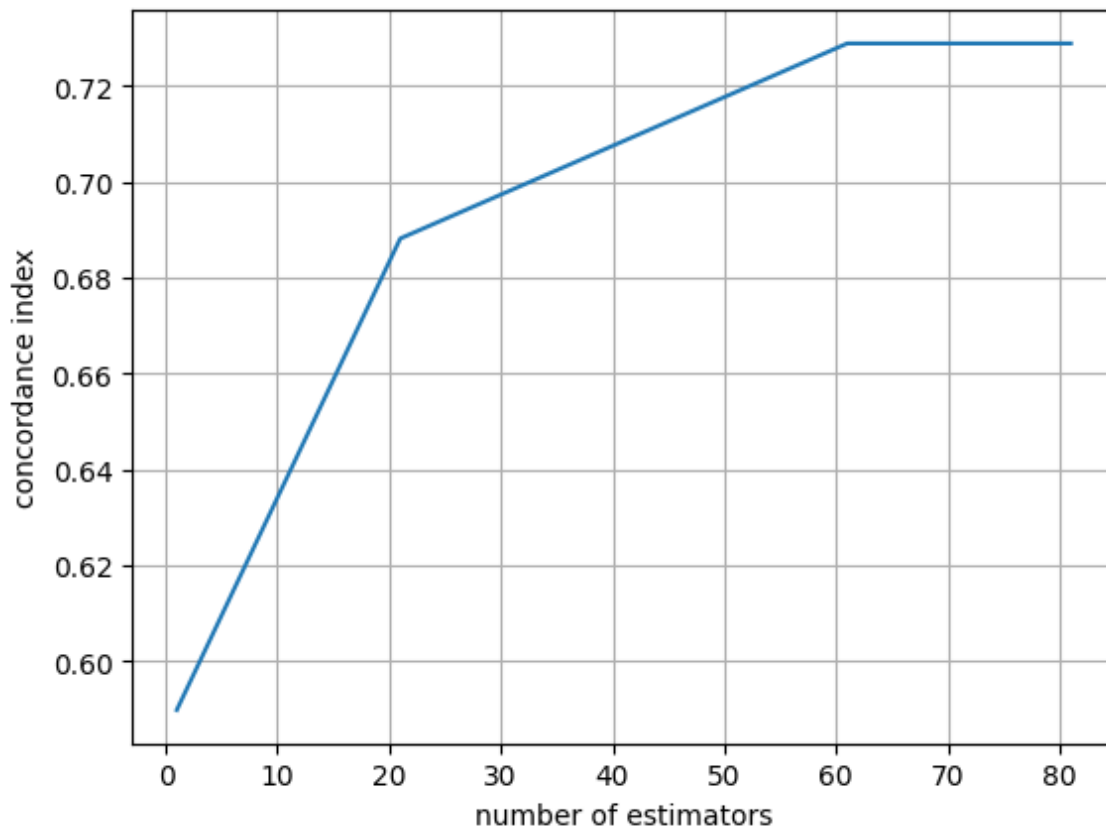
As a final method to evaluate *Risk Score*, survival analysis methods were used. The principle behind using survival analysis to evaluate an *Endometriosis Risk Score* is to consider symptoms, their seriousness, and the date of symptom logging.

Survival analysis is a statistical method used to predict the time until an event of interest (in this case, the diagnosis of endometriosis) occurs, particularly useful in medical research for analysing disease outcomes over time. In this context, each patient's symptoms and their severity are tracked from the day they are enrolled in the study. The 'day of enrolment' parameter serves as a timeline, marking the start point from which each patient's experience with potential symptoms leading up to a diagnosis is observed. This approach allows for a detailed examination of how symptoms evolve and impact the risk of developing endometriosis over time.

The core idea is to model the relationship between the symptoms (along with their assessed seriousness) and the time it takes for the disease to manifest. By analysing these relationships, it becomes possible to identify which symptoms are more strongly associated with a higher risk of endometriosis and how quickly those risks escalate following symptom onset.

Survival analysis techniques, such as the Cox Proportional Hazards model and Random Survival Forests, were used to accommodate the varying times at which patients might be diagnosed or remain diagnosis-free, accounting for both the occurrence and timing of the event. They allow to estimate the probability of diagnosis over time, given the presence of specific symptoms and their severity.

On *Fig. 18*, the concordance index relation to number of the survival model parameter amount ('estimators') is shown:



*Figure 18: Survival analysis concordance index on the full dataset.*

The concordance (C-index) is a measure used to evaluate the predictive accuracy of survival models. It quantifies the model's ability to correctly predict the order of event times. Specifically, for a pair of subjects, the model's predictions are considered concordant if the subject predicted to have an event (e.g., diagnosis of a disease) sooner than it actually does.

The C-index ranges from 0.5 (no better than random chance) to 1.0 (perfect prediction), with values closer to 1 indicating a model with strong predictive accuracy. Obtained result of ~68% shows average accuracy of the *Risk Score* prediction. To increase the C-index, several statistical values were added as features. Final result of a survival prediction model reached ~73% (see *Fig. 18* above).

The standard deviation of seriousness levels highlights the variability in symptom severity, indicating patients with fluctuating disease activity. Kurtosis offers insight into the distribution shape of symptom severity, identifying heavy tails that suggest the occurrence of rare but extremely severe symptoms. The maximum value of seriousness levels points to the peak severity a patient has experienced, crucial for understanding the potential impact on prognosis. Quantiles, specifically the 35<sup>th</sup> and 75<sup>th</sup>, help delineate the commonality and rarity of certain severity levels, respectively, offering a nuanced perspective on how often patients experience mild versus severe symptoms.

By integrating these features, survival analysis can more accurately predict disease outcomes, identify patients at higher risk, and facilitate targeted clinical interventions, based on detailed severity profiles. This comprehensive approach enhances the understanding of disease dynamics, improving patient management and care strategies. To analyse the effect of accuracy depending on enrolment day amount of user, three survival method results (Fig. 19) were plotted:

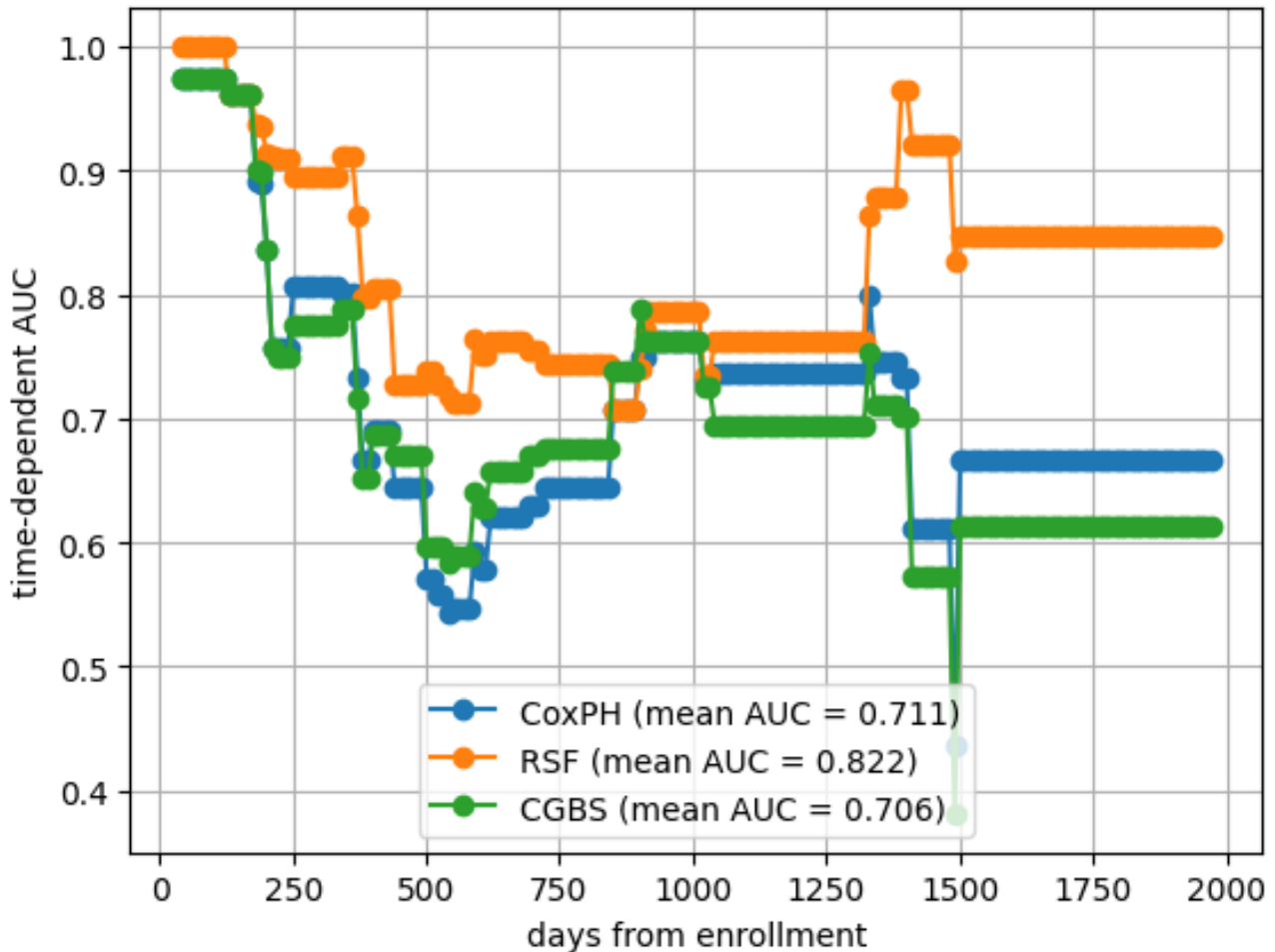


Figure 19: AUC curves for three survival rate evaluation methods.

As seen on Figure 19, accuracy is highly dependent on the specific user. There are several notable zones that should be carefully treated while building the final system:

- Zone >1,500 days of enrolment includes few patients and that result could not be treated as accurate.
- Zone <250 show inadequate high accuracy.
- Zone in the middle shows realistic accuracy, since after a year of logging, there are enough symptoms data to make robust prediction, neglecting the effects of random symptoms appearance not related to the *Endometriosis Risk Score*.

The overall survival rate accuracy and robustness are **not adequate** to be used in a real-world system, but according to RandomForest classification result, accuracy could be improved by selecting patients with known diseases (their absence) and qualitative data. Survival models require high amount of time-related data, as compared to RandomForest classificatory. To use survival model results for the *Risk Score* calculation, the database should be expanded, while keeping the patient diagnoses and their absence well known.

## 8. Multimodal approach

Combining the described methods, including Random Forest classifiers and survival analysis, to a multi-modal neural network offers a smart way to look at complex health problems, like how a disease gets worse over time. This setup is especially useful in healthcare because it can utilise many different sorts of data, such as notes from doctors, results from lab tests, and even MRI images. Each sort of data, like the text from doctor's notes or images from MRIs, is prepared in a very distinct way to make sure it is ready for the model to analyse. After preparing the data, the system can then use specific tools to pick out important information from each sort of data. For MRI scans, it could be possible to extract important patterns that could show signs of disease.

The long-term goal is to teach this system by showing it examples all at once, so it gets better at picking out and learning from the most helpful details. This could help the system to become really good at both figuring out health problems and predicting health outcomes. However, making a system that uses many different sorts of data is a complex and costly task. It is extremely hard to make sure the system understands and uses all the data correctly. The most difficult part in the FEMaLe project's WP5 would be combining MRI with *Lucy* data, as it requires personal involvement of each patient that has MRI data, and since *Lucy* is designed to keep patient data fully anonymized, it has not been possible for us to do so.