Endometriosis and Its Correlation with Lifestyle Factors and Health Indicators

A Data Mining Approach Using R and Python

JONAS STYLBÄCK

ELLA VILLFÖR
Endometriosis and Its Correlation with Lifestyle Factors and Health Indicators
A Data Mining Approach Using R and Python

Endometrios och dess samband med livsstilsfaktorer och hälsoindikatorer
En datautvinningsmetod med R och Python

JONAS STYLBÄCK
ELLA VILLFÖR

Degree project in medical engineering
First level, 15 credits
Supervisor at KTH: Tobias Nyberg, Mattias Mårtensson
Examiner: Mats Nilsson

KTH Royal Institute of Technology
School of Engineering Sciences in Chemistry, Biotechnology and Health
Hälsovägen 11 C
SE-141 57 Huddinge, Sweden
http://www.kth.se/cbh

2023
Abstract

Around 10% of women in fertile age have endometriosis, despite this there is little known about its origin. It can take years from the first experienced symptoms to an established diagnosis, which is done using invasive methods. A database from the Lucy application containing 11720 questionnaires with up to 41 questions each was examined for signs of associations between lifestyle factors and health indicators with endometriosis.

The database was restructured and unusable questionnaires were pruned, leaving 5719 questionnaires for analysis. Questionnaires were assigned a label depending on whether the respondent had an endometriosis diagnosis, some other diagnosis or no diagnosis. Histograms were created for 36 questions to visualize answer distributions among the labels while word clouds were created for 3 questions to determine commonly used words. Pearson’s chi-square test was performed on 24 questions to determine whether there were statistically significant differences in how respondents answered with regards to their diagnosis. Then, 26 questions were divided between seven question groups based on their similarities. The relative prevalence of endometriosis for every answer alternative was determined and compared with the average for that group. This was then followed up by a multi-correspondence analysis for every question group.

It was found that there are associations between endometriosis and 12 health indicators. These being: affected by infertility, always being tired, regarding own health as bad, having more than 100 birthmarks, severe chronic pelvic pain, dysmenorrhea, regularly whole-body pain, muscle stiffness, neck and shoulder tension, forgetfulness, concentration difficulties, regularly experiencing that stress worsens physical symptoms. No statistically significant association between lifestyle factors and endometriosis could be found in the survey database. There were findings of socioeconomic differences between respondents with endometriosis and those without a diagnosis which may have affected the result.

Keywords:

Endometriosis, lifestyle factors, correspondence analysis, data mining
Sammanfattning

Runt 10% av kvinnor i fertilt ålder har endometrios, trots detta är lite känt om dess ursprung. Det kan gå flera år från första upplevda symptom till en etablerad diagnos, som kräver invasiva metoder. En databas från Lucy applikationen med 11 720 enkäter upp till 41 frågor vardera undersöktes efter tecken på kopplingar mellan livsstilsfaktorer och hälsoindikatorer med endometrios.


Det hittades associationer mellan endometrios och 12 hälsofaktorer. Dessa var: infertilitet, alltid vara trött, betrakta sin egna hälsa som dålig, ha fler än 100 födelsemärken, uppleva allvarlig kronisk bäckensmärta, dysmenorré, regelbunden helkroppssmärta, muskelstelhet, spändhet i nacke och axlar, glömskhet, koncentrationssvårigheter, regelbundet uppleva att stress förvärnar fysiska symptom. Ingen statistiskt signifikant association kunde hittas mellan livsstilsfaktorer och endometrios från undersökningsdatabasen. Det hittades socioekonomiska skillnader mellan respondenter med endometrios och de utan diagnos vilket kan ha påverkat resultatet.

Nyckelord:

Endometrios, livsstil, korrespondensanalys, datautvinning
# Table of Contents

1 Introduction 1
   1.1 Aims 1

2 Background 3
   2.1 Endometriosis 3
   2.2 Data Structures 3
   2.3 Word Cloud 4
   2.4 Python Programming Language 4
   2.5 R Programing Language and Modules 5
   2.6 Jupyer Notebook 5
   2.7 Kotlin 5
   2.8 FEMaLe 5
   2.9 Lucy 5
   2.10 Pearson’s Chi-square Test of Homogeneity 8
   2.11 Corresponding Analysis 8
   2.12 Multiple Corresponding Analysis 11

3 Method 13
   3.1. Development Environment and Data 13
   3.2. Pipeline 1 - Python 13
   3.3. Pipeline 2 - R 15

4 Results 19
   4.1 Response Frequency 19
   4.2 Chi-square Test of Homogeneity 19
   4.3 Histograms of Questions 1-31, 55, 56, 101, 102 and 180 19
   4.4 Word Clouds 20
   4.5 Prevalence Diagrams 21
   4.6 Factor Maps and Scree Plots 23

5 Discussion 25
   5.1 There are Differences to Find 25
   5.2 Infertility a Common Affliction 25
   5.3 Stress Makes Physical Symptoms Worse 25
   5.4 Severe Pain in the Pelvic Region 25
   5.5 Health and Pain in General 26
   5.6 Mental Fitness and Health Suffers 26

vii
5.7 High Number of Birthmarks, a Surprise 27
5.8 Dietary Restrictions and Endometriosis 27
5.9 Activity and Selfcare 27
5.10 The Impact of Socioeconomic Factors 28
5.11 Data Shortcomings 28
5.12 Reliability of Factor Maps 29
5.13 Summery 29
6 Conclusion 31
7 References 33
Appendix 1: Introduction text to the Lucy Study
Appendix 2: Questionnaire in the Lucy Application
Appendix 3: Histograms of Relevant Questions
Appendix 4: Prevalence Diagrams
Appendix 5: Factor Maps and Scree Plots
Definitions

CA Correspondence analysis

Cos2 value Squared cosinus value

CSV Comma-separated values, a common way to structure data.

dataframe-image A Python module for Jupyter Notebooks which converts a DataFrame to image files.

deep-translator A Python module for translating text from one language to another using services such as Google Translate.

dplyr A R package for working with DataFrame-like objects.

Dysmenorrhea Medical term for painful menstruation periods

Dyspareunia Medical term consistent pain in genitalia during or after intercourse

Escape characters Characters that can lead to an alternative interpretation of succeeding characters.

factoextra A R package that provides functions to extract and visualize the output of multivariate data analyses such as factor maps.

FactoMineR A R package containing data analysis methods to summarize, visualize and describe sets of data.

ggplot2 A R package containing a system for creating.

json A Python module for encoding and decoding data in JSON format. Its main feature is the ability to convert JSON data to a Python Dictionary object.

Matplotlib A Python module for visualising data.

MCA Multi correspondent analysis

NumPy A scientific computing module for Python. It enables multiple forms of numerical computing such as shape manipulation, sorting, random simulation, Fourier transforms and more.

os A Python module that provides operating system dependent functionality such as getting and setting directory paths.

pandas A data analysis and manipulation module for Python. Its main feature is the "DataFrame" object, which is a two-dimensional table for data manipulation.
<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pickle</td>
<td>A Python module for serializing and de-serializing Python objects. Its main feature is being able to save objects such as lists or dictionaries to a file while preserving the structure.</td>
</tr>
<tr>
<td>plyr</td>
<td>A R package containing tools for splitting, applying and combining data.</td>
</tr>
<tr>
<td>re</td>
<td>A Python module for regular expression matching operations. A regular expression is a sequence of characters that specifies a pattern to match in text, for example it can be used to find a date on the YYYY-MM-DD format.</td>
</tr>
<tr>
<td>reshape2</td>
<td>A R package containing tools to restructure and aggregate.</td>
</tr>
<tr>
<td>SciPy</td>
<td>A Python module that provides algorithms for many types of problems such as optimization and statistics.</td>
</tr>
<tr>
<td>tidyr</td>
<td>A R package containing tools to help clean up in messy data and working with missing values.</td>
</tr>
</tbody>
</table>
1 Introduction

Endometriosis is a condition where the tissue lining of the uterus grows outside of the uterus. Approximately 10% of all women in reproductive age suffers from endometriosis [1]. The condition can become a source of pelvic pain during period, intercourse, urination and bowel movement, it can also lead to infertility and excessive bleeding [1]. The exact cause of endometriosis is unknown, but there are some studies suggesting that it is inherited [2].

Establishing diagnosis is challenging as endometriosis lack an accurate non-invasive biomarker [3]. To reliably establish diagnosis, a tissue sample must be extracted by an invasive procedure and examined [4]. In addition to being challenging to diagnose, women affected by the condition may have to wait more than six years before their diagnosis is surgically established [4]. This time gap can be explained by the lack of specialist referrals, with women having to visit their primary care giver seven times on average before receiving a diagnosis [4].

Finding Endometriosis using Machine Learning (FEMaLe), a project funded by the EU Horizon 2020 programme, aims to create tools to predict and prevent endometriosis [5]. Thus, increasing the quality of life for people living with the condition [5]. One of these tools, the Lucy application, was released in 2021 and is a menstruation tracking app, which amongst other things help its users to know when to expect their period or when they are fertile [6]. In addition to its normal functions, the application presents its users with the opportunity to opt-in for a research survey [7]. This survey contains 23-45 questions related to health indicators and were is filled out on a monthly basis over the course of a year.

As a member of the Horizon programme, KTH Royal Institute of Technology takes part in FEMaLe. The project at hand is to analyse survey answers provided by Lucy and look for statistically relevant connections between lifestyle factors and health indicators with endometriosis.

1.1 Aims

The aim of the project was to perform statistical analysis on a survey database provided by the Lucy application in order to decide whether there are statistically relevant connections between lifestyle factors and health indicators with endometriosis.
2 Background

In the following part background information is presented to give a thorough understanding of concepts that will be important in the report. The reader will be introduced to endometriosis, some data structures, word clouds, some programming languages, the FEMALe project, the Lucy application, Pearsons’ chi-square test, correspondence analysis and multi correspondence analysis.

2.1 Endometriosis

Endometriosis is a condition where the endometrium grows outside of the uterus and is often located on the outside of the pelvic organs, on the fallopian tubes, ovaries, and peritoneum [4]. However, it may also occur in other areas beyond the pelvic organs such as on the lungs and intestines [4]. The leading symptoms of endometriosis are chronic pelvic pain, fatigue, dysmenorrhea, infertility, and dyspareunia [1]. In a US study, patients affected by the condition also had a larger number of comorbidities including mental, endocrine-based, and autoimmune disorders compared to the general population [2]. Due to these symptoms women with endometriosis have reduced work efficiency, which can be translated to losing 10.8 hours of work weekly on average [4].

Diagnosing endometriosis is challenging due to the lack of an accurate, non-invasive, biomarker for the incidence and prevalence of superficial endometriosis [3]. This meaning that the endometrium, growing in the peritoneal cavity, have not left larger scars that can be felt or seen without laparoscopy [3]. In addition, a study by Kelechi E. Nnoaham et al. reveals an average time gap of 6.7 years from first experienced symptoms to established diagnosis [4]. According to the study, the main cause for the size of the time gap is the lack of specialist referrals [4]. Furthermore, the study points to an increase in this time gap in the presence of higher pelvic pain and increased body mass index (BMI) [4].

Treatment options for endometriosis are limited. Medical intervention and surgically removing tissue have shown to be effective in mitigating the symptoms, however, there is a 50% symptom recurrence post-treatment after a five-year period [1]. Previous research has shown that fish oil as well as antioxidant vitamins C, D and E may help in reducing pain [3]. An interview study by Jenny Vennberg Karlsson et al. [8] indicates that an individual-adapted dietary plan has the potential to reduce pain and fatigue for women affected by endometriosis. Specifically, decreasing the intake of gluten and dairy products while increasing the intake of fruits and vegetables may help to lessen some of the symptoms [8].

2.2 Data Structures

In the following section the reader will be introduced to two data structures that are used in this project, the dictionary, and JavaScript Object Notation (JSON). Both structures are members of the so called “name-value pairs” format common to programming.

Dictionary and Nested Dictionaries

A dictionary is a type of data structure with key-value pairs. For example, if you were to search for the definition of "apple" in the merriam-webster dictionary you would get "apple: the fleshy, usually rounded red, yellow, or green edible pome fruit of a usually cultivated tree (genus Malus) of the rose family" [9]. Here, apple is the key and the definition is the value associated with the key. In programming, if the value to a key is another dictionary the resulting data structure is called a nested dictionary.
JavaScript Object Notation

JavaScript Object Notation (JSON) is a data format that is designed to be readable by both humans and machines [10]. JSON is an interchangeable format, which means it is language independent [10]. JSON formatted data is built on one out of two universal data structures, a collection of name-value pairs often called dictionaries, or an ordered list of values often called arrays [10]. An example of the JSON format is shown in figure 1.

```json
{ "book": {
    "title": "Animal Farm",
    "author": "George Orwell",
    "category": "Political satire",
    "published": "1945-08-17"
}
}
```

Figure 1: Publisher details for the classical novel "Animal Farm" by George Orwell described using JSON name-value pairs format.

2.3 Word Cloud

A word cloud is a graphical representation of the occurrence of words in a text. For example, the tool can be applied to summarize text by visualizing the most occurring words in it [11]. Usually, to make frequency visual words that occur often will be presented with a larger text size than those occurring less often. The number of words is often limited to prevent the word cloud from becoming too muddled. Words that do not contribute to any context to the content of the text can be filtered out by being added to a list of stop words. Though word clouds do give a statistical summary it does not take linguistic knowledge into consideration and provide limited interaction capabilities [11]. An example of a word cloud can be seen in figure 2.

![Word Cloud Example](image.png)

Figure 2: A word cloud displaying names of colours. The bigger the word is the larger occurrence it has. The colour of the word does not have any implications.

2.4 Python Programming Language

Python is an open-source programming language developed by the Python Software Foundation and was first released in 1991 [11]. It is a multi-paradigm language,
which means that it supports different types of programming approaches. According to the TIOBE Index, Python is considered the most popular programming language as of May 2023 [12].

One of the main advantages of Python is that users can extend its functionality by installing code libraries called "modules". NumPy and SciPy are common modules for mathematical and scientific computing while pandas can be used to create a DataFrame object, a type of data structure. Word clouds can be generated with the help of the wordcloud module, that has a predefined list of stop words named STOPWORDS [12]. The full list of word that STOPWORDS contains of can be found at the GitHub repository for the word cloud python module [13]. The dataframe-image can be used to save tabulated data as images. There are also some useful modules that come packaged with a typical Python installation such as the os, re, json and pickle modules.

2.5 R Programming Language and Modules

R is a programming language developed by the R Core Team in 1993 and, as of May 2023, ranked the 16th most popular programming language by the TIOBE Index [14,15]. R provides a variety of statistical and graphical techniques, functionalities that can be extended with code libraries called "packages" [16]. Like Python modules, these packages can come bundled with a default R installation or can be imported from an external source. Theplyr, dplyr, tidyr, reshape2 and ggplot2 packages are common for data processing and analysis. The FactoMineR and factoextra packages are popular packages for correspondence analysis.

2.6 Jupyter Notebook

Jupyter Notebook is an interactive programming environment and was developed by Project Jupyter in 2014 [17]. In this programming environment, code is written in sections called "cells" which can be run independently from one another [18]. A Jupyter notebook environment includes features such as live code, plots, images, and videos. Jupyter Notebook supports multiple programming languages including Python and R [18].

2.7 Kotlin

Kotlin is a programming language developed by JetBrains in 2011 [19]. It is designed to interoperate with the Java programming languages and is mainly used for development of mobile applications [20].

2.8 FEMaLe

Finding Endometriosis using Machine Learning (FEMaLe) is a project funded by the European Union’s Horizon 2020 Research and Innovation Programme [21]. FEMaLe aims to improve intervention for individuals with endometriosis, combating the negative effects of the condition by utilizing data-driven tools [5].

The project has promised to deliver a mobile health application, three clinical-decision support tools and a computer vision-based software for real-time augmented reality guided surgery of endometriosis [5]. Ultimately, the project expects to reduce overall cost of endometriosis treatment by at least 20%, while improving patient outcomes [5].

2.9 Lucy

Lucy is a mobile women’s health application that was developed in collaboration with FEMaLe. The application act as a menstrual tracker and diary, letting users know when they
can expect menstruation or when they are fertile based on previous data, among other things [22].

A research survey has been developed in collaboration with FEMaLe that track women's health indicators and lifestyle on a monthly basis for twelve months and has been made available to users of the Lucy application [7]. The introductory text to the survey can be found in appendix 1. Participants that choose to opt-in for the survey will fill out 40-45 questions initially and 23-28 questions during following surveys [7]. Each question and its respective answer alternative can be found in appendix 2.

The questions are varied and range from age and education level to health indicators. Questions regarding the respondent’s health status are considered health indicator questions such as experienced pain during menstruation and intercourse or if the respondent is sun sensitive. The survey also included questions regarding the respondent’s lifestyle, which are factors the respondent can change or affect in their life, such as diet. Some questions are answered by choosing one or more pre-defined answers while some ask the participant to enter their own free-form answer. The surveys are anonymous and have no personal identifiable information, there is however a unique 36-character string to identify surveys from the same participant. Filled-out surveys are stored in a database using a combination of JSON and CSV structures as described in figure 3, which is accompanied by value details as described by table 1. A total of 11720 questionnaires has been filled out in the Lucy application to this date.

```
id,
{
    question_id:[question_type, answer, timestamp],
    question_id:[question_type, answer, timestamp],
    ....
    question_id:[question_type, answer, timestamp]
},
date,
language_code,
question_id_stack,
questionnaire_id,
connection_id,
completion_count
```

Figure 3: The data structure of a filled-out survey. Firstly, on row one is the survey id. Between the first braces, from row two until row seven, is the answers for the questionnaire following JSON format. For each question there is a question id, a question type, an answer and a timestamp. Accompanied information such as date and language code are stored in CSV format. Each field value is described in detail by table 1.
Table 1: Description of field values stored in a filled-out survey. The field names are listed in the left-most column while a short description can be found in the right-most column.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Survey specific identifier</td>
</tr>
<tr>
<td>question_id</td>
<td>Question specific identifier</td>
</tr>
<tr>
<td>question_type</td>
<td>One out of seven predefined question types</td>
</tr>
<tr>
<td>answer</td>
<td>One or more answers as a string or a list of strings</td>
</tr>
<tr>
<td>timestamp</td>
<td>Timestamp in Unix Epoch format</td>
</tr>
<tr>
<td>date</td>
<td>Date in YYYY-MM-DD format</td>
</tr>
<tr>
<td>language_code</td>
<td>Participant language code</td>
</tr>
<tr>
<td>question_id_stack</td>
<td>Groups of question identifiers</td>
</tr>
<tr>
<td>questionnaire_id</td>
<td>The survey number by this participant</td>
</tr>
<tr>
<td>connection_id</td>
<td>Participant-unique string for cross-survey connections</td>
</tr>
<tr>
<td>completion_count</td>
<td>Number of previously completed surveys by this participant</td>
</tr>
</tbody>
</table>
2.10 Pearson’s Chi-square Test of Homogeneity

To test if two or more groups is one homogenous population or not the chi-square test of homogeneity can be used [23]. If two or more distributions are statistically significantly different from each other or if they are a part of one homogenous distribution. The test is a non-parametric statistical method that compares the distribution of the responses from the groups to determine if they are dependent [24]. For statistical analysis a null hypothesis (H0), a specified idea on the appearance of the distribution, is created [23]. In the case of disproving the homogeneity of a group, H0 is set to be that the groups are a part of one homogenous population [23,24]. Therefore, if H0 can be disproven it can be said that all groups are not part of one homogenous population. If the groups are homogenous, the proportion of each set of responses should be the same in all groups. In the test a Q-value is calculated showing how much the recoded values differ from the expected values. Equation (1) describes how to calculate the Q-value [23]:

$$\displaystyle Q = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(x_{ij} - \frac{n_{ij}m_j}{N})^2}{\frac{n_{ij}m_j}{N}} \quad (1)$$

S is the number of columns, r is the number of rows, n_{ij} is the total number of respondents in a column, m_j is the total respondents to a row and N is the overall total number of the respondents. The term $\frac{n_{ij}m_j}{N}$ is therefore the expected value to an observation and $x_{ij}$ is the actual recorded value. The higher Q is the bigger difference there is between the distribution in the answers from the groups [23]. To reject H0 the p-value, the probability of H0, must be smaller than a set alpha value. The alpha value is the uncertainty of the test and is usually put to 0.05 [23]. This is equivalent to the $Q_{obs}$ value, the observed value of Q, being larger than the corresponding $x_\alpha^2$ distribution value, see equation (2) [23].

$$\alpha > p \iff Q_{obs} > x_\alpha^2((r - 1)(s - 1)) \quad (2)$$

2.11 Corresponding Analysis

Corresponding analysis (CA) is a technique used to try and detect underlying structures in tabulated categorical data, which is data that consists of discrete points rather than data that can take on any value within a range [25]. An example of categorical data is whether a person is male or female or whether they are right- or left-handed [25]. Imagine data as described by table 2, the first step to CA is to calculate the row and column averages as seen in the very right and bottom of the table followed by the overall average, as seen in the bottom right corner [25].

<table>
<thead>
<tr>
<th>Table 2: Original data and averages. Exemple data and their averages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Average</td>
</tr>
</tbody>
</table>

The next step is to calculate the expected value, which is the value if the data was homogenous [25]. For each row-column combination, take the average row value, multiply
by the column average, and divide the product with the overall average [25]. Applying these steps on table 2 results in table 3.

<table>
<thead>
<tr>
<th></th>
<th>Right</th>
<th>Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>26.47</td>
<td>18.53</td>
</tr>
<tr>
<td>Female</td>
<td>23.53</td>
<td>16.47</td>
</tr>
</tbody>
</table>

The third step is to calculate the residual values, which is the subtraction of the expected values from the original [25]. A large positive value corresponds to a strong positive association while a large negative value corresponds to a strong negative association [25]. Residuals represent the association between rows and columns in the data [25]. The residuals of table 2 can be seen in table 4.

<table>
<thead>
<tr>
<th></th>
<th>Right</th>
<th>Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3.53</td>
<td>-3.53</td>
</tr>
<tr>
<td>Female</td>
<td>-3.53</td>
<td>3.53</td>
</tr>
</tbody>
</table>

To represent these associations graphically the table of residuals must first be decomposed with a method called Singular value decomposition [26]. Singular value decomposition is beyond the scope of this report. It is sufficient to say that, as a result of the method, we are given a matrix of coordinates which represents the original values position in a multi-dimensional space [26]. The number of dimensions, also called principal axes, depends on the number of categories in the original data [26]. Coordinates on principal axes are called principal coordinates [26].

Categories can have a stronger or weaker presence on a principal axis. One category can have a strong presence on the first principal axis while having a weak presence on the remaining axes et cetera [26]. A measurement of how strongly a category is represented on an axis is determined by the cosine value of the category in relation to the axis, which is derived from its principal coordinates [26]. When considering a plane made up of two principal axes, one uses the squared cosine value from each axis, called squared cosine value (Cos2) to determine representation of the category on the plane [26]. Categories that have a high Cos2 value are well represented on the chosen axes [26]. As an example, imagine a category in 6 dimensions with the principal coordinates 6, -4, -7, 8, 11 and -4. To determine how well the category is represented in a principal plane made up of axis 3 and 6, plot the corresponding coordinates and draw right-angled triangle as described by figure 4. The cosine value is then calculated by taking the length of the adjacent side divided by the length of the hypotenuse. In our case, this results in the cosine value of 0.875. This is then squared to return the Cos2 value of 0.766, which would mean this category is well represented on the chosen plane.
How well a principal axis can represent the underlying structure of the data is measured in the variance covered by that axis [26]. Principal axes are ranked in such a way that the first axis covers the most variance, the second axis the second most variance and so on [26]. The variance covered rapidly plateaus beyond the first two or three axes, as such the first two axes are considered the most important in CA [26].

Not all variances might be necessary to describe the underlying structures in a dataset, as such one could consider removing axes that retains a low amount of variance [26]. This can be done by creating and inspecting a scree plot, which is a bar plot of eigenvalues calculated from the matrix of principal coordinates [26]. Variances have a linear relationship with eigenvalues, as such principal axes that have an eigenvalue below a chosen threshold can be removed from the data to reduce complexity in the analysis [26].

As a final step, categories are graphically illustrated on a two-dimensional scatter plot called a "factor map" or simply a "map". To interpret the results of CA on a map, consider the following [27]:

- Draw a line from the origin through a category, any other category along or close to the line can be interpreted to be associated with that category.
- Categories that are further from the origin are more differentiated than those that are closer to the origin. That is, their association or disassociation is stronger than if they were closer to the origin.
- Two categories that are negatively associated with one-another will be found on the opposite side of the origin.
- Two categories that are both close and well represented on the map can be considered both similar and associated.

See figure 5 for an example using movie genres.

Figure 5: An example in interpreting the results of a factor map in the first principal plane (dimensions 1 and 2), where the principal coordinates of 8 movie categories are plotted. In this example, multiple conclusions can be drawn. Action and Adventure are associated, at the same time they are negatively associated with Sci-Fi. Horror and Thriller are both similar and associated and negatively associated with Drama. Musical and Noir are more differentiated than Action and Drama, as can be seen by their distance from the origin.

### 2.12 Multiple Corresponding Analysis

Multiple correspondence analysis (MCA) is an extension of CA that is effective when analyzing data of more than two categorical variables [26]. The underlying mathematics of MCA is beyond the scope of this report and the reader is instead referred to the work of Roux et al. [26]. It is sufficient to say that MCA utilize the same principles as CA but is applicable to data with higher dimensionalities, such as a multiple-answer questionnaire described in previous sections.

The interpretation of the results of MCA is similar to the one done in CA. As a first step one consults the scree plot to choose which principal axes to retain or remove, relevant axes are then plotted using one or more factor maps [26]. Next step is to inspect the Cos2 value of
each category to see which are well represented on the chosen axes and which are not [26]. Associations can then be drawn on well represented categories with the same criteria as in CA [26].

If MCA is to be used to draw associations in relation to a category that is not itself used in the analysis, that category can instead be added as a supplement [26]. This might be the case when applying MCA on categories such as “taste in music” and trying to see if there are underlying associations with gender [26]. In this case gender would be added as a supplementary category [26].
3 Method

In the following part of the report is the steps to be able to perform our analysis. It is divided into three larger subsets, starting off with how to develop the environment that was used and then the two pipelines used to run the analysis. Pipeline 1 uses python to create histograms, wordclouds and performs the chi-square test. Pipeline 2 uses R to determine the relative prevalence among question groups and create factor maps. The code to run each analysis and preceding pre-processing of data can be found in a separate folder in the same location as this report. This folder also includes a text file outlining prerequisites and running instructions.

3.1. Development Environment and Data

The development environment was set up as a Jupyter Notebooks (Project Jupyter, https://jupyter.org) instance running code in both Python (Python Software Foundation, https://www.python.org/) and R (R Foundation, https://www.r-project.org). Additional modules and packages were installed to extend functionality of both languages. A list of software used in the development environment and their versions can be found in table 5.

<table>
<thead>
<tr>
<th>Python Software</th>
<th>Version</th>
<th>R Software</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>3.10.6</td>
<td>R</td>
<td>4.3.0</td>
</tr>
<tr>
<td>NumPy</td>
<td>1.24.2</td>
<td>plyr</td>
<td>1.8.8</td>
</tr>
<tr>
<td>pandas</td>
<td>2.0.0</td>
<td>dplyr</td>
<td>1.1.2</td>
</tr>
<tr>
<td>SciPy</td>
<td>1.10.1</td>
<td>tidyr</td>
<td>1.3.0</td>
</tr>
<tr>
<td>matplotlib</td>
<td>3.7.1</td>
<td>FactoMineR</td>
<td>2.8</td>
</tr>
<tr>
<td>dataframe-image</td>
<td>0.1.11</td>
<td>factoextra</td>
<td>1.0.7</td>
</tr>
<tr>
<td>wordcloud</td>
<td>1.8.2.2</td>
<td>reshape2</td>
<td>1.4.4</td>
</tr>
<tr>
<td>deep-translator</td>
<td>1.10.1</td>
<td>ggplot</td>
<td>3.4.2</td>
</tr>
</tbody>
</table>

Table 5: Software versions used. The os, re and pickle modules come bundled with Python and share its version number.

The Lucy data was provided by our supervisor in the form of two files. One was a Kotlin file outlining the structure of the questionnaire within the Lucy application. In this file, each question in the questionnaire had a unique identifier called question_id. If the question had pre-defined answers associated with it, the answers in turn also had a unique identifier called answer_id. The other file was a text file containing 11720 questionnaires in a CSV and JSON mixed format, see figure 3.

The code was split up into two pipelines, one using Python and another using R. Each pipeline was accompanied by a utility file containing custom functions for processing and analysing data.

3.2. Pipeline 1 - Python

As a first step, the Kotlin file was processed. Each question, answer and their unique identifiers were extracted and saved as a dictionary. This information was extended with the exact wording of the question and answer alternatives as it appeared to the respondent.

As a second step the text file containing questionnaires were processed. The data was restructured into pure JSON format and inspected for errors and duplicates. Out of the original 11720 questionnaires, 4554 were found to not follow the structure outlined by the Kotlin file and were discarded. Out of the remaining 7166 questionnaires, 1447 were found
to be duplicates of already existing questionnaires and were discarded. This left 5719 questionnaires to be used in the analysis.

As a third step, to see whether the data was rich enough to follow individuals over time, respondent-unique identifiers called connection_id:s were inspected. Out of 5719 questionnaires, 624 lacked this identifier. Out of the 5095 questionnaires with a connection_id, 3651 unique respondents were found.

As a fourth step it was necessary to know if a questionnaire came from a respondent with endometriosis, a respondent with some other diagnosis or from a respondent without a diagnosis. Questionnaires were first grouped by their connection_id to identify questionnaires from the same respondent and which questionnaires that lacked a connection_id. Questionnaires were then placed in one of three subsets, “no_diagnosis”, “endometriosis” or “other_diagnosis”. The criteria by which questionnaires were placed in subsets are described in figure 6. Out of 5719 questionnaires, 2066 questionnaires were placed in the "no_diagnosis" subset, 1241 in the "endometriosis" subset and 2412 in the "other_diagnosis" subset. Finally, all questionnaires were placed in a data frame.

Figure 6: Flowchart describing which criteria were used to place questionnaires in one of the three subsets, “no_diagnosis”, “endometriosis” or “other_diagnosis”. Green arrows show the flow for questionnaires from an identifiable respondent and red arrows shows the flow for questionnaires without an identifiable respondent. Numbers within parentheses represent the number of individual questionnaires at a point. Questionnaires were first grouped by their connection_id, a respondent-unique identifier. Questionnaires originating from the same respondent were all processed together while questionnaires that lacked this identifier were individually processed.

As a fifth step, both histograms and word clouds were generated. Histograms consisted of one staple per category and respondent type, resulting in three staples per category. The staple heights were scaled according to the answer distribution within each respondent type for that category. As such, a difference in staple heights between different respondent types reflects a difference in the answer distribution, not necessarily a difference in the number of respondents that chose that category. For the sake of transparency, and to aid interpretation
of the result, the number of respondents that chose a specific category was also added to the top of each staple as a label. Histograms were generated for questions 1-31, 55, 56, 101, 102, and 180 in appendix 2. Word clouds were generated using the wordcloud python module for questions 52-54. This was done in order to visualize the frequency of individual words and whether common words differed between subsets. The number of words to appear in each cloud were set to 25 and each word in the cloud was translated to English. The default list of stop-words was used, in addition the following words were used as stop words: lot, none, many, sometimes and little

As a sixth step, to know whether the data could be considered uniformly distributed or not, Pearson's chi-squared test of homogeneity was used on questions 2, 3, 6-16, 23-31, 101, and 102 in appendix 2. The significance level for the test was set to $\alpha = 0.05$.

As a final step in the Python pipeline, pre-processing was done to prepare the data for further analysis. Questions 18-22 were all endometriosis-specific questions and were subsequently removed. Question 0 and 120 were removed as they were not actual questions but rather contained information on the upcoming questions. Questions 5 and 180 were removed as they had already been used to determine subsets. Columns containing dates, timestamps, languages, and connection-IDs were also removed. Each remaining question in the data frame had their answer_id:s replaced by the text as the respondents see them, for example answer_id 60 was replaced by “often”.

The resulting data frame was split into two data frames, one containing questions 1-4, 6-17, 23-31, 101, 102 and another data frame containing 55 and 56. This split was made due to the nature of questions 55 and 56, which were multiple-answer questions and had to be expanded element-wise as described in figure 7.

![Figure 7: The expansion of question 55 and question 56. Cells containing more than one answer alternative were split row-wise to only leave a single answer per cell. Cells containing no answers were overwritten with cells containing the text "NAN".](image)

### 3.3. Pipeline 2 - R

As a first step, the “other diagnosis” subset was removed from the data frame. This was done to focus further analysis on the difference between the “no diagnosis” and “endometriosis”
subsets only. Next, to use factor maps, numerical values had to be re-coded as categories. The values that were re-coded and to what can be found in table 6.

Table 6: A list describing which questions contained numerical answer alternatives and how they were re-coded as categories.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
<th>Re-coded as</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10-19</td>
<td>Adolescence</td>
</tr>
<tr>
<td></td>
<td>20-29</td>
<td>Early adult</td>
</tr>
<tr>
<td></td>
<td>30-39</td>
<td>Adult</td>
</tr>
<tr>
<td></td>
<td>40-59</td>
<td>Middle aged</td>
</tr>
<tr>
<td></td>
<td>50-99</td>
<td>Elder</td>
</tr>
<tr>
<td>6-11</td>
<td>0-3</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>4-7</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>7-10</td>
<td>Severe</td>
</tr>
<tr>
<td>17</td>
<td>0-33</td>
<td>Bad</td>
</tr>
<tr>
<td></td>
<td>34-67</td>
<td>OK</td>
</tr>
<tr>
<td></td>
<td>68-100</td>
<td>Good</td>
</tr>
</tbody>
</table>

As a second step, each question was placed in one of seven different question groups. Questions that were considered similar to one-another were placed in the same group. Groups and included questions can be found in table 7.

Table 7: Details of which question belongs to which question group.

<table>
<thead>
<tr>
<th>Group</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>General health</td>
<td>15, 17, 25, 26, 29</td>
</tr>
<tr>
<td>Activity</td>
<td>12, 13, 14, 24, 30</td>
</tr>
<tr>
<td>Pelvic</td>
<td>6, 7, 8, 9, 10, 11</td>
</tr>
<tr>
<td>Socio-economic</td>
<td>1, 2, 3, 4</td>
</tr>
<tr>
<td>Mental</td>
<td>16, 23, 27, 28, 31</td>
</tr>
<tr>
<td>Food</td>
<td>55, 56</td>
</tr>
<tr>
<td>Other indicators</td>
<td>101, 102</td>
</tr>
</tbody>
</table>

As a third step, the prevalence of endometriosis was calculated for each question group. It was then also calculated for each combination of each question and answer alternative within the group. Question-answer combinations that were not present in both subsets were removed from this analysis. The relative increase or decrease in prevalence for each question-answer combination compared to the group was then visualized in a plot.
As a final step, factor maps and scree plots were generated using the R-packages FactoMineR and factoextra. For each question group, questions-answer combinations were chosen as active categories while “endometriosis” and “no_diagnosis” were chosen as supplementary categories. First a scree plot was generated for each question group to see the distribution of variance across principal axes, the maximum number of axes to visualize was set to 10 to reduce cluttering. Next, factor maps were generated in the first principal plane for each question group. To reduce cluttering, only 15 of the active categories with the highest Cos2 values were drawn on the factor map. Each remaining category was in turn coloured according to their Cos2 value while the supplementary categories were coloured purple to stand out from the active categories. A dotted line was drawn across the two supplementary categories through the origin to aid in identifying associations between active and supplementary categories.

The baseline for a “good enough” representation of an active category in the first principal plane was set to a Cos2 value of 0.25. An active category was considered to be associated with a supplementary category if it was well represented in the first principal plane and could be found within a 10-degree angle from the dotted line with regards to the origin. An active category was considered to be weakly associated with a supplementary category if it was well represented in the first principal plane and could be found within a 20-degree angle from the dotted line with regards to the origin. Categories that were either not well represented on the first principal plane or were found outside a 20-degree angle from the dotted line with regards to the origin were not considered. The criteria for evaluating associations between active and supplementary categories can be found in figure 8.
Figure 8: Evaluation of associations between active and supplementary categories on a factor map. Only active categories within the area of consideration are evaluated for possible association with a supplementary category. Categories that can be found between the supplementary categories and the origin are deemed “too close” and not considered. Thus, only the green active category in the figure can be associated with the rightmost supplementary category. The blue category has a Cos2 value below the threshold and is not considered. The yellow category is found too close to the origin to be considered. The red category is found outside the area of consideration and is not evaluated.
4 Results

In the following part relevant discoveries and results from the different tests are presented. These include response frequency, histograms, chi-square test of homogeneity, word clouds, prevalence diagrams and factor maps.

4.1 Response Frequency

The number of times each respondent had filled out a questionnaire can be found in table 8. Out of 3650 individual respondents, 74.8% only filled the questionnaire out once, 15.3% filled it out twice, 7% thrice and less than 3% filled it out four time or more. No respondent filled out the questionnaire more than nine times.

<table>
<thead>
<tr>
<th>Completed questionnaires</th>
<th>Number of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>258</td>
</tr>
<tr>
<td>2</td>
<td>560</td>
</tr>
<tr>
<td>1</td>
<td>2731</td>
</tr>
</tbody>
</table>

4.2 Chi-square Test of Homogeneity

The chi-square test of homogeneity showed that questions 2, 3, 6-16, 23-31, 101 and 102 did not follow a uniform distribution in the answers between the groups, with a significance of 0.05. Questions 26 (having headaches), 102 (number of birthmarks) and 7 (pain during bleeding days) had the closest distributions.

4.3 Histograms of Questions 1-31, 55, 56, 101, 102 and 180

All histograms can be found in appendix 3. For many of the questions, the answer distributions for the different groups followed similar shapes, but there were some categories that stood out. Histograms related to socioeconomic questions, such as age, where you live, education and relationship status, showed larger differences in distributions. Figure 9 shows the histogram for question 23 and is an example of this. It can be seen that “often” tired is the most common answer in all three groups. Looking at the answer “always” tired a distinguishably larger part of the endometriosis group answered compared to the other two groups and “sometimes” is likewise a more common answer in the no diagnosis group compared to the other two groups. The answer “never” tired had few numbers of answers.
Other categories that stood out to be more common with the endometriosis group, compared to the other two groups were, infertility, limiting or removing gluten, sugar and dairy, “often” having troubles concentrating, “always” having troubles sleeping, “rarely” having headaches, “moderate” pain and discomfort and answering 9 or 10 on pain during bleeding days.

The Categories that had a low number of answers include “extreme” pain and discomfort, “severe problems” and “unable” to do usual activities, “severe problems” and “unable” to walk about and “moderate problems”, “severe problems” and “unable” to perform selfcare. Also, answers higher than 7 on question 11 (dysuria), higher than 8 on question 10 (dyschezia) and answers higher than 8 on question 9 (dyspareunia). Lastly being older than 55 years, divorced, a widow or having a PhD had a low number of answers.

4.4 Word Clouds

For question 52, 53 and 54 the word clouds created can be seen in figure 10. For question 52, words related to stress occurred with a slightly higher frequency proportion in endometriosis respondents. Overall, in questions 52 and 54 no big differences were noted between the word clouds.

The word clouds for question 53 showed that “good” and “sick” were frequent answers in the no diagnosis group and other diagnosis group while less frequent in the endometriosis group. “Pain” is a uniquely common response by the endometriosis respondents for question 53.
4.5 Prevalence Diagrams

In appendix 4 diagrams from the prevalence analysis are shown. Regarding questions related to activity, there was a lower prevalence for endometriosis than the average and having “no”, “never” or “rarely” problems with stiffness or activities. People who described their different activity problems with the labels “always”, “often”, “moderate”, “slight” and “severe” had a higher prevalence of endometriosis with between 3.7 - 29.4% more than the average. The categories with low difference in prevalence of endometriosis are neck tension “sometimes”, stiffness “sometimes” and “no problems” with selfcare.

For food related questions the prevalence diagram reveals a tendency of limiting or removing gluten, dairy, and sugar in people with endometriosis. Labels with lower prevalence of endometriosis than the average were “limit” or “remove” coffee from their diet. There were ≤0.5% difference in the prevalence of endometriosis between the two groups among the “removed soy”, “limited soy” and “limited red meat” categories.

The general health related diagram showed higher prevalence of endometriosis than the average towards experiencing more severe pain. The exception for this is headaches where the tendency was shifted and prevalence of endometriosis for headaches “sometimes” and “always” were lower than the average. Having headaches “often”, “okey” health, “slight” pain and discomfort and “sometimes” having symptoms affected by stress shows little to no difference in prevalence of endometriosis.

The prevalence diagram for mentally related states can be seen in figure 11. Stats such as anxiety, forgetfulness, and concentration, showed that the group of people that perceived these more extremely had a higher prevalence of endometriosis. Categories with little tendencies at either way is “slight” anxiety and “often” tired. Both “always” and “never” unwell sleep has higher prevalence of endometriosis.
**Figure 11:** The prevalence diagrams for questions 17, 23, 27, 28 and 31 of mental health group. The blue coloured lines represent a relative decrease in prevalence of endometriosis compared to overall average. The red coloured lines represent an increase in prevalence of endometriosis compared to overall average.

The group perceiving pelvic related issues more severely have a higher prevalence of endometriosis, according to the prevalence diagram. There were small tendencies for low symptoms and not having a diagnosis. The prevalence of endometriosis was higher among those with infertility, meanwhile the prevalence of endometriosis was lower among those without infertility. The labels “medium” pelvic pain, “low” dysmenorrhea, “medium” dysmenorrhea, and “low” dysuria had small or no difference in prevalence of endometriosis than the average.

From the other group prevalence diagram, there was a higher prevalence of endometriosis than the average when having more than 100 birthmarks with 6.8%. All other birthmarks related labels had lower prevalence of endometriosis than the average. The other group prevalence diagram also showed that sun sensitivity had higher prevalence of endometriosis than the average and not being sun sensitive had lower prevalence.

The prevalence diagram for socioeconomic questions showed that labels representing middle age or adult age had both higher prevalence of endometriosis than average, 30% and 26.1% respectively. The labels married, college education and living in a capital had 14.4%, 18.1%
and 7.8% higher prevalence of endometriosis than the average respectively. The lower age categories, adolescence, and early adult, as well as lower educational status, elementary, vocational and high school graduation, have 11.3 - 35.5% lower prevalence of endometriosis than the average. Living in a city, town or village have 3.4 - 16% lower prevalence than the average.

4.6 Factor Maps and Scree Plots

All the factor maps and scree plots can be found in appendix 5. Endometriosis is found to be associated with being of adult age, having a college education, living in a capital, and being married. Furthermore, it was associated with experiencing slight problems with self-care, moderate pain and discomfort, severe pelvic pain, severe dysmenorrhea and often experiencing whole-body pain. The factor maps for the pelvic related categories can be seen in figure 12. In addition, there were signs of weak association between endometriosis and moderate activity problems, always experiencing headaches, always tired, always unwell sleep, often muscle stiffness, always stress symptoms and 100+ birthmarks.

In perspective, “no diagnosis” was found to be associated with no activity problems, no mobility problems and no selfcare problems. Furthermore, there were signs of weak associations of endometriosis with being of adolescence age, being single, highest education being an elementary degree, experiencing no pain or discomfort and experiencing general health as good. All other categories were either not well represented or not associated.
Figure 12: The factor maps with associated scree plot below, for pelvic related categories. Questions included were question 7, 8, 9, 10 and 11. The label of each category is shortened as described in method. The scale of the cos2 value can be found on the right in the factor maps. Categories that are close to the association line on the endometriosis side is “severe” dysmenorrhea, “severe” pelvic pain, “severe” dyspareunia and “severe” dyschezia. The scree plot shows that the two first dimensions hold 33.7% of the total variance.
5 Discussion

In this section we will discuss and evaluate the results from the previous section. The reader will be guided through each finding and the underlying reasoning for including or excluding a factor in the conclusion. The reader will also be introduced to some of the sources of error, the impact of socio-economic factors and what might be done in further studies.

5.1 There are Differences to Find

The results of the chi-square test found in 4.2 support the theory that the two groups, those with and those without endometriosis, do not belong to one homogenous group on the tested parameters. This indicates that there are significant statistical differences in how individuals with or without endometriosis answer the questionnaire. Furthermore, because the questions process how the respondents experience their health and lifestyle, there are differences between people with or without endometriosis in connection to lifestyle factors and health indicators.

5.2 Infertility a Common Affliction

As shown by the histogram for question 6 (Appendix 3), women diagnosed with endometriosis are affected by infertility to a larger degree than those without, 43.7% versus 11.8%. This is also supported by the prevalence diagram for the pelvic question group (Appendix 4), where the prevalence of endometriosis among individuals suffering from infertility is 33.7% higher than average. In contrast, prevalence of endometriosis is 24% lower than average for individuals not suffering from infertility. This association between infertility and endometriosis is not made apparent by the factor map for the pelvic question group (Appendix 5) as the category is discarded due to low Cos2 values, suggesting that it is not well represented in the first principal plane. Despite poor performance on the factor map, we consider the findings in the histogram and prevalence diagram are strong enough to support the association between infertility and endometriosis.

5.3 Stress Makes Physical Symptoms Worse

As observed in the prevalence diagram for the general health question group (Appendix 4), the prevalence of endometriosis is higher among individuals that have answered that stress makes their physical symptoms worse. Individuals answering “always” or “often” had 19.3% and 8.7% higher prevalence of endometriosis than average respectively. In contrast, individuals answering “never” or “rarely” had 20.1% and 10.2% lower prevalence of endometriosis than average respectively. This association between stress and worsening physical symptoms can also be observed in the histogram for question 29 (Appendix 3), where the answer distribution of the endometriosis group tends to have stress affecting physical symptoms more often, while the answer distribution of the no diagnosis group tend to not having stress affecting physical symptoms as often. Observing the factor map for the general health question group (Appendix 5), the label representing “always” is found in the cone of association and well represented, the same could not be said about the label representing “often”. These results are strong enough to support the association between endometriosis and regularly experiencing that stress makes physical symptoms worse.

5.4 Severe Pain in the Pelvic Region

As observed in the prevalence diagram for the pelvic question group (Appendix 4), the prevalence of endometriosis increases with severity of pain. The group of individuals that experience their dysuria, dyschezia or dyspareunia as “severe” had 23.4%, 17.2% or 15.6%
higher than average prevalence of endometriosis respectively. In addition, the group of individuals that experience their chronic pelvic pain as “severe” had 14.8% higher prevalence of endometriosis than average. It should be noted that this association between pain and endometriosis is not as differentiating within the “low” and “medium” labels, where the difference is within 1.1 – 7.7% higher or 1.4 – 5.0% lower than average.

The factor map and scree plot for the pelvic question group, shown in figure 12, supports the association between “severe” pain and endometriosis, as the “severe” label for dysmenorrhea, dyspareunia, dyschezia and chronic pelvic pain are all well represented and within the cone of association. However, inspecting the histograms (Appendix 4) of these categories give rise to caution as question 9, 10 and 11 have few respondents in the “severe” answer range. Due to this, the only association between endometriosis and pain in the pelvic area that can be reliably established is that of severe chronic pelvic pain and that of severe dysmenorrhea.

5.5 Health and Pain in General

Apart from headaches, the prevalence diagram for the general health question group (Appendix 4) indicates an association between endometriosis and negative health indicators. Labels representing “moderate”, “severe” and “extreme” discomfort and pain had 16.3%, 14.3% and 7.4% higher than average prevalence of endometriosis. This is also observed for whole-body pain as labels representing “sometimes”, “often” and “always” had 7.4%, 15.9% and 32.8% higher than average prevalence of endometriosis. In addition, the group of individuals describing their health as “bad” had a 10.7% higher than average prevalence of endometriosis. In contrast, labels representing neutral or positive health indicators had average or lower than average prevalence of endometriosis. From the word-cloud of question 53, figure 10, we can also see a tendency for those diagnosed with endometriosis to uniquely describe their everyday health with “pain”.

Some of these associations are supported by the factor map for the general health question group (Appendix 5), the labels representing “often” whole-body pain and bad health are both found well represented and associated with endometriosis while good health are found well represented and associated with no diagnosis. All other labels are discarded due to being outside the cone of association or not being well represented on the first principal plane. Once again, the histograms (Appendix 3) urge caution as question 15, relating to pain and discomfort, have relatively few answers in the “severe” and “extreme” labels. With these results in mind, it’s clear that there is an association between endometriosis and regularly experiencing whole-body pain as well as an association between endometriosis and describing one’s own health as bad.

5.6 Mental Fitness and Health Suffers

The prevalence diagram for the mental health question group figure 11 displays a few interesting points. The prevalence of endometriosis is higher than average for the labels representing severe and extreme anxiety or depression, 11.5% and 25% respectively. It is also higher for the labels representing forgetfulness and concentration difficulties, the labels “often” and “always” forgetful had an 10.9% and 10.1% increase while “often” and “always” having concentration difficulties had an 6.1% and 5.5% increase. In contrast, the prevalence of endometriosis is lower than average for the labels representing low amount or lack of anxiety or depression, forgetfulness, and concentration difficulties. Curiously, the label “always” tired had 19.6% higher than average prevalence of endometriosis while
“often”, “sometimes”, “rarely” and “never” all had lower than average prevalence of endometriosis.

The histograms for question 28 and 31 (Appendix 3) support the tendency for those with endometriosis to be more forgetful and have more concentration difficulties than their counterparts. This can be seen because a larger percentage of the endometriosis group tends more towards having more issues with forgetfulness and concentration, whereas a larger part of the no diagnosis group tends towards less problems with the two. Figure 9 also displays that people with endometriosis answer “always tired” in a higher degree than people without endometriosis.

Regarding the connection between endometriosis and anxiety or depression, care should be taken. This is because the histogram for question 16 (Appendix 3) shows that “severely” and “extremely” anxious and depressed have low number of answers which can skew the distribution. With regards to the factor map of the mental health question group (Appendix 5), the labels that are associated with endometriosis and well represented in the first principal plane are also the labels that represented a higher than average prevalence of endometriosis. The same can however not be said about labels representing a lower than average prevalence of endometriosis as they are not found inside the area of consideration with no diagnosis. An association can be drawn between endometriosis and regularly experience forgetfulness, regularly experience concentration difficulties as well as always tired.

5.7 High Number of Birthmarks, a Surprise

According to the results from the prevalence diagram for the “other” question group (Appendix 4), the prevalence of endometriosis is 6.8% higher than average for the label “100+” birthmarks. In contrast, labels representing 100 or less birthmarks had between 1.4 – 11.4% lower than average prevalence of endometriosis. The factor map for the “other” question group (Appendix 5) also shows an association between endometriosis and 100+ birthmarks, as the label is within the cone of association and well represented in the first principal plane. From the histogram for question 102 in appendix 3 we can observe the tendency for individuals within the endometriosis group to have more birthmarks than their counterparts. There is no evidence in these results of a linear relationship between endometriosis and the number of birthmarks, there appears however to be an association with endometriosis and having an excessive number of birthmarks.

5.8 Dietary Restrictions and Endometriosis

As can be seen in the prevalence diagram for the food question group (Appendix 4), labels representing limiting gluten, removing dairy products, and removing sugar had 8.0-11.9% higher than average prevalence of endometriosis. This relationship is supported by the histograms for question 55 and 56 in appendix 3. In contrast, the factor maps (Appendix 5) do not show any association between dietary limitations and endometriosis as none can be found within the cone of association. With this in mind, no reliable conclusion can be drawn with regards to an association between dietary lifestyle choices and endometriosis.

5.9 Activity and Selfcare

It can be observed in the prevalence diagram for activity related questions (Appendix 4) that people answering “often” and “always” on question 30, neck and shoulder tension, have 7.8% and 17.7% higher prevalence of endometriosis than the average. The histogram for question 30 also supports that people with endometriosis answer to having neck and shoulder
tension “often” and “always” to a higher extent. The factor map shows the association between endometriosis and “always” neck tension but with bad representation.

It can be seen in appendix 5 for the other activity related questions, 12, 13, 14 and 24, only “slight” selfcare problems and “often” muscle stiffness is inside of the association cone in the factor map. Both categories have fairly good representation according to the Cos2 value. These labels also have a high prevalence of endometriosis. When considering the histogram for question 13 (Appendix 3), regarding selfcare, the answer's response range is very narrow with 89%-96% answering “no problems”. Therefore, it cannot with certainty be concluded whether people with endometriosis have more selfcare problems than those without.

5.10 The Impact of Socioeconomic Factors

The prevalence diagram for socioeconomic questions (Appendix 4) revealed that labels representing “middle age” or “adult age” had both higher prevalence of endometriosis than average, 30% and 26.1% respectively. For socioeconomic related questions the results reveal that higher age has higher prevalence of endometriosis than the average. The histograms for question 1 (Appendix 3) and factor map for the socioeconomic question group (Appendix 5) support the association between endometriosis and being of “adult age” or “middle age”. This is not surprising as previously mentioned in the report, it may be several years between first experienced symptoms and establishing an endometriosis diagnosis [4].

Except age, other socioeconomic factors can be observed to have higher than average prevalence of endometriosis. These include the labels “married”, “college education” and “living in a capital”. Associations with endometriosis and these labels are also supported according to the histograms for question 2, 3 and 4 (appendix 3) and the factor map for socioeconomic related questions. Answering “divorced” and “doctoral education” also showed a high prevalence of endometriosis but the histograms for question 3 and 4 reveals that these have too few answers to support a conclusion.

Socioeconomic factors can play a role in previously mentioned health indicators. The tendency to be older could have an impact on questions regarding marital status, education level, infertility, activity and mental fitness. Living in a capital may increase healthcare access, especially for specialist referrals such as gynecologists, which would further impact health indicators.

5.11 Data Shortcomings

It is not feasible to follow how lifestyle factors and health indicators developed over time for a specific individual due to the low quality of the data. Ideally, each participant would have completed the entire survey of 12 questionnaires while in fact less than 10% of all respondents had completed 3 or more questionnaires with no-one completing more than 9, see table 8. It would have been interesting to see if changes in diet could have had an impact on perceived pain over time as suggested by Jenny Vennberg Karlsson et al [8].

With regards to the word-cloud in figure 10, little information could be extracted from the freeform answers using this method. This is mostly a limitation of the approach as phrases could not be preserved in the word-clouds. For example, "no pain" would be split up as "no" and "pain". As "no" is included in the list of stopwords, it would be removed and "no pain" would simply be represented as “pain” in the word-cloud. As generating word-clouds using this approach may not correctly represent the sentiment of the respondent, we believe it is
ineffective for further studies and instead suggest a separate qualitative study of the freeform answers.

5.12 Reliability of Factor Maps

There are areas of improvement to be done with regards to the factor maps. Scree plots, which can be found under each responding factor map in appendix 5, reveals that only between 19.6 - 35.2% of variance is retained in the first two principal axes. This means that 64.8 - 80.4% of the underlying features in the data remains unexplained in the first principal plane. As a majority of the features remain unaccounted for, there are almost certainly categories that are not well represented in the first principal plane that might have been better represented in another.

Future studies should consider extracting the Cos2 values of each category on each principal axis and determine the minimum number of planes to effectively describe a majority of variance, as outlined in La Roux et al [26]. This would reduce complexity and make it practically feasible to create factor maps only for those categories and axes that are of interest.

5.13 Summary

There are few surprises in the results. Infertility and pain in the pelvic area are already known ailments for women living with endometriosis. It is difficult to know if factors such as a reduction in activity and increase in whole-body pain is an indicator of the disease or of age. However, there are more underlying structures in the data to be uncovered and we consider the Lucy database to be a prime candidate for further studies.
6 Conclusion

12 health indicators in the survey database were found to have statistically significant associations with endometriosis. These are: being affected by infertility, always being tired, regarding own health as bad, having more than 100 birthmarks, experience severe chronic pelvic pain and dysmenorrhea, regularly experiencing whole-body pain, muscle stiffness, neck tension, forgetfulness and concentration difficulties and regularly experiencing that stress worsens physical symptoms. No lifestyle factors in the survey database were found to have statistically significant associations with endometriosis.
7 References


Appendix 1: Introduction Text to the Lucy Study

A copy of the text presented to users of the Lucy application, informing them of the study and research goals. Found amongst the data files given to us during the project. Another copy of this text can be found on the website for the Lucy application (https://hellolucy.app/en/female2020). Introduction text to the Lucy app:

We would like to introduce you our current study, which is part of the European collaboration of the FEMaLe Project (Finding Endometriosis using Machine Learning).

The aim of our study is to find out more about patients living with endometriosis. To achieve this, we also need to ask healthy women who do not have endometriosis about the characteristics of their menstrual cycle, quality of life and lifestyle.

In this survey, we would track your women's health indicators on a monthly basis for a year, by asking you to answer just 40-45 questions at the first time, and 23-28 questions later. The questions can be answered in a short time frame, there are no right or wrong answers, so we would encourage all participants to answer our questions honestly, according to their current menstruation cycle.

The survey uses data from 12 time points to track the health of participants, the stability, or changes in the health status. These values will allow us to identify different patterns and, based on these patterns, to further improve the quality of gynaecological care according to the health conditions and needs of women today.

Both the Lucy app and our research are anonymous, so neither the people running the study, nor our development engineers, nor even our staff analyzing the data will know which data comes from whom. There will be no final medical report or lab report on the results of the study. We are not able to provide feedback on individual results of participants, but we can provide feedback on aggregate results.

## Technical details of the study

Name of study not involving an intervention: FEMaLe – Finding Endometriosis using Machine Learning.

The investigator-in-charge: Dr. Borok Attila PhD, egyetemi docens (Semmelweis Egyetem Szülészeti és Nőgyógyászati Klinika Baross utcai részleg, 1082 Budapest, Baross utca 27.)

Questions about the app and how to use it can be asked here [lucyappinfo@gmail.com](mailto:lucyappinfo@gmail.com)


The investigator-in-charge thanks you in advance for your cooperation,
Dr. Attila Borok, MD, PhD.
Associate Professor, Semmelweis University
Faculty of Medicine, 1 st. Department of Obstetrics and Gynaecology
Appendix 2: Questionnaire in the Lucy application

A structured copy of the questionnaire presented to users of the Lucy application that opted in for participation in the study. Found amongst the data files given to us during the project. First column is the question identification number, next column is the question and its type. Lastly the third column is the answer alternatives.

<table>
<thead>
<tr>
<th>ID</th>
<th>Question [type of question]</th>
<th>Answer alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is your age? [Value]</td>
<td>Scale: 1-100 [years old]</td>
</tr>
<tr>
<td>2</td>
<td>Where do you live? [Single]</td>
<td>• Capital&lt;br&gt;• City&lt;br&gt;• Town&lt;br&gt;• Village</td>
</tr>
<tr>
<td>200</td>
<td>Please name the capital/city/town/village you live in. [Freeform]</td>
<td>Freeform answer (single)</td>
</tr>
<tr>
<td>3</td>
<td>What is the highest degree or level of education you have completed? [Single]</td>
<td>• Less than eight elementary school classes&lt;br&gt;• Eight elementary school classes,&lt;br&gt;• Vocational training,&lt;br&gt;• Graduation,&lt;br&gt;• College / university degree,&lt;br&gt;• Doctoral degree, PhD</td>
</tr>
<tr>
<td>4</td>
<td>What is your marital status? [Single]</td>
<td>• Single&lt;br&gt;• Living in a relationship&lt;br&gt;• Married, living in a cohabiting relationship&lt;br&gt;• Divorced&lt;br&gt;• Widowed</td>
</tr>
<tr>
<td>5</td>
<td>My diagnosed disease(s): [Multi]</td>
<td>• I have no diagnosed disease,&lt;br&gt;• Endometriosis,&lt;br&gt;• Insulin resistance (IR),&lt;br&gt;• Pelvic inflammatory disease (PID)&lt;br&gt;• Ovarian cyst&lt;br&gt;• Myoma, fibroids&lt;br&gt;• Polycystic ovary syndrome (PCOS)&lt;br&gt;• Other</td>
</tr>
<tr>
<td>No.</td>
<td>Question</td>
<td>Options</td>
</tr>
<tr>
<td>-----</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>500</td>
<td>If you had any disease marked in the previous question, please state the date of your diagnosis. [Freeform]</td>
<td>Freeform answer (single)</td>
</tr>
<tr>
<td>6</td>
<td>Are you currently affected by infertility, difficulty conceiving? [Single]</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Can not tell</td>
</tr>
<tr>
<td>7</td>
<td>Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle <strong>on the bleeding days (dysmenorrhea)</strong> [Value]</td>
<td>Scale: 0 [No pain] – 10 [Severe intolerable pain]</td>
</tr>
<tr>
<td>8</td>
<td>Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle <strong>during nonbleeding day (chronic pelvic pain)</strong> [Value]</td>
<td>Scale: 0 [No pain] – 10 [Severe intolerable pain]</td>
</tr>
<tr>
<td>9</td>
<td>Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle <strong>during or after sexual intercourses (dyspareunia)</strong> [Value]</td>
<td>Scale: 0 [No pain] – 10 [Severe intolerable pain]</td>
</tr>
<tr>
<td>10</td>
<td>Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle <strong>during defecation (dyschezia)</strong> [Value]</td>
<td>Scale: 0 [No pain] – 10 [Severe intolerable pain]</td>
</tr>
<tr>
<td>11</td>
<td>Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle <strong>during urination (dysuria)</strong> [Value]</td>
<td>Scale: 0 [No pain] – 10 [Severe intolerable pain]</td>
</tr>
<tr>
<td>12</td>
<td><strong>MOBILITY</strong> [Single]</td>
<td>• I have no problems in walking about</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• I have slight problems in walking about</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• I have moderate problems in walking about</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• I have severe problems in walking about</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• I am unable to walk about</td>
</tr>
<tr>
<td>13</td>
<td><strong>SELF-CARE</strong> [Single]</td>
<td>• I have no problems washing or dressing myself</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
| 14 | **USUAL ACTIVITIES** (e.g. work, study, housework, family or leisure activities) [Single] | - I have slight problems washing or dressing myself  
- I have moderate problems washing or dressing myself  
- I have severe problems washing or dressing myself  
- I am unable to wash or dress myself  
- I have no problems doing my usual activities  
- I have slight problems doing my usual activities  
- I have moderate problems doing my usual activities  
- I have severe problems doing my usual activities  
- I am unable to do my usual activities  |
| 15 | **PAIN/DISCOMFORT** [Single] | - I have no pain or discomfort  
- I have slight pain or discomfort  
- I have moderate pain or discomfort  
- I have severe pain or discomfort  
- I have extreme pain or discomfort  |
| 16 | **ANXIETY/DEPRESSION** [Single] | - I am not anxious or depressed  
- I am slightly anxious or depressed  
- I am moderately anxious or depressed  
- I am severely anxious or depressed  
- I am extremely anxious or depressed  |
| 17 | Please mark on the scale to indicate how your health is TODAY. [Value] | Scale 0 - 100 |
| 180 | Do you have a diagnosed case of **endometriosis***? [Single] | - Yes  
- No  |
| 18 | During the last 4 weeks, how often, because of your endometriosis **have you found it difficult to walk because of the pain**? [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always  |
<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Response Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>During the last 4 weeks, how often, because of your endometriosis <strong>have you felt as though your symptoms are ruling your life</strong>? [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>20</td>
<td>During the last 4 weeks, how often, because of your endometriosis <strong>have you had mood swings</strong>? [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>21</td>
<td>During the last 4 weeks, how often, because of your endometriosis <strong>have you felt others do not understand what you are going through</strong>? [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>22</td>
<td>During the last 4 weeks, how often, because of your endometriosis <strong>have you felt your appearance has been affected</strong>? [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>23</td>
<td>I feel tired and unrefreshed when I wake from sleeping. [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>24</td>
<td>My muscles feel stiff and achy. [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td>25</td>
<td>I feel pain all over my body. [Single]</td>
<td>Never, Rarely, Sometimes, Often, Always</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
| 26 | I have headaches. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
| 27 | I do not sleep well. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
| 28 | I have difficulty concentrating. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
| 29 | Stress makes my physical symptoms get worse. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
| 30 | I have muscle tension in my neck and shoulders. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
| 31 | I have difficulty remembering things. [Single] | - Never  
- Rarely  
- Sometimes  
- Often  
- Always |
<p>| 52 | Please describe your <strong>lifestyle (even your diet, exercise, sleep)</strong> with a few individual words. [Freeform] | Freeform answer (multi) |
| 53 | Please describe your <strong>health</strong> with a few individual words [Freeform] | Freeform answer (multi) |</p>
<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>54 Please describe <strong>yourself</strong> with a few individual words [Freeform]</td>
<td></td>
</tr>
<tr>
<td>55 Mark which foods and nutrients you have <strong>completely removed</strong> from your nutrition in the last month. [Multi]</td>
<td>• Sugar&lt;br&gt;• Gluten&lt;br&gt;• Coffee&lt;br&gt;• Soy&lt;br&gt;• Dairy&lt;br&gt;• Red meat</td>
</tr>
<tr>
<td>56 Mark which foods and nutrients you have <strong>limited</strong> in your diet in the past month. [Multi]</td>
<td>• Sugar&lt;br&gt;• Gluten&lt;br&gt;• Coffee&lt;br&gt;• Soy&lt;br&gt;• Dairy&lt;br&gt;• Red meat</td>
</tr>
<tr>
<td>101 Are you sensitive to the sun? [Single]</td>
<td>• Yes&lt;br&gt;• No&lt;br&gt;• Can not tell</td>
</tr>
<tr>
<td>102 Number of your birthmarks:</td>
<td>• ~10&lt;br&gt;• ~25&lt;br&gt;• ~50&lt;br&gt;• ~100&lt;br&gt;• 100+</td>
</tr>
</tbody>
</table>
Appendix 3: Histograms of Relevant Questions

The following appendix has the histograms of the answer distribution for questions 1-31, 55, 56, 101, 102, 180. The title for each answer histogram is the question, bars shows the proportional answers distribution within the group and the number on top of the bars is the amount of people who have answered that alternative from that group. The small box in each histogram presents the color division and the number of collected answers for each group for that question.
Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle during nonbleeding day (chronic pelvic pain) (Question 8)

Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle during or after sexual intercourse (dyspareunia) (Question 9)
Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle during defecation (dyschezia) (Question 10)

Please mark on the scale below how strong you have experienced YOUR PAIN in your current or recent cycle during urination (dysuria) (Question 11)
During the last 4 weeks, how often, because of your endometriosis have you found it difficult to walk because of the pain? (Question 18)

During the last 4 weeks, how often, because of your endometriosis have you felt as though your symptoms are ruling your life? (Question 19)
During the last 4 weeks, how often, because of your endometriosis have you had mood swings? (Question 20)

During the last 4 weeks, how often, because of your endometriosis have you felt others do not understand what you are going through? (Question 21)
During the last 4 weeks, how often, because of your endometriosis have you felt your appearance has been affected? (Question 22)

Being tired (Question 23)
I have headaches. (Question 26)

I do not sleep well. (Question 27)
I have muscle tension in my neck and shoulders. (Question 30)

I have difficulty remembering things. (Question 31)
Mark which foods and nutrients you have completely removed from your nutrition in the last month (Question 55)

Mark which foods and nutrients you have limited in your diet in the past month (Question 56)
Do you have a diagnosed case of endometriosis? (Question 180)

- No diagnosis: 2,066 answers
- Endometriosis: 1,241 answers
- Other diagnosis: 2,412 answers

Distribution:

- No, I don’t: 2,077
- Yes, I do: 1,106
Appendix 4: Prevalence Diagrams

The following appendix contains diagrams of the prevalence of categories between the endometriosis group and no diagnosis group. Positive relative prevalence describes increased prevalence from the endometriosis group and negative is higher prevalence from the no diagnosis group. Positive relative prevalence is represented with red and negative relative prevalence is represented with blue.

Socioeconomic group [questions: 1, 2, 3, 4]
Pelvic group [questions: 6, 7, 8, 9, 10, 11]
Activity group [questions: 12, 13, 14, 24, 30]
General health group [questions: 15, 17, 25, 26, 29]
Mental health group [questions: 16, 23, 27, 28, 31]
Food group [questions: 55, 56]
Other group [questions: 101, 102]
Appendix 5: Factor Maps and Scree Plots

This appendix contains the factor maps and scree plots for each group of questions. The categories in the groups are investigated to have associations with having endometriosis and no diagnosis. The scree plots shows the percentage of explained variance in the first 10 dimensions.

Socioeconomic group [questions: 1, 2, 3, 4]

![Factor Map and Scree Plot](image-url)
Pelvic group [questions: 6, 7, 8, 9, 10, 11]
Activity group [questions: 12, 13, 14, 24, 30]
Mental health group [questions: 16, 23, 27, 28, 31]

Variable categories - MCA

Dim2 (9%)

Dim1 (13.2%)

Scree plot

Percentage of explained variances

Dimensions
Other group [questions: 101, 102]