Finding patterns in irritable bowel syndrome patient data

Jakub Gdula Stephan Delhey Virginia Ardevol Martinez
Ariadna Saladrigas Pernias Julius Haring

September 2019

1 Introduction

Irritable Bowel Syndrom (IBS) is the most common gastrointestinal disorder in the population worldwide and it negatively impacts the quality of patients' lives. It is characterized by abdominal pain, diarrhea, constipation or a combination of both.[1] IBS diagnosis is symptom-based and there still exists a gap of reliable instruments which measure the effectiveness of the therapy. Currently, end of day questionnaires are the most common reporting method. In this method, patients answer a set of questions regarding the pain they encountered throughout the day and other disease-specific symptoms. It is argued, however, that their evaluations are affected by recall bias, i.e., the pain reported is overestimated. It has been proven that patients tend to report the peak rather than the average of pain encountered on that specific day. Taking into account those insights, a new method for recording therapy effectiveness, based on momentary reporting, has been proposed. [2] In this new method, called Electronic Experience Sampling Methodology (ESM), the patients are asked to answer a questionnaire which relates to their current pain perception and other disease specific factors in a momentary manner. The project, conducted with a close collaboration with the Department of Internal Medicine at the Maastricht University Medical Center, aims to investigate the presence of patterns in the data gathered from 73 subjects with the ESM method. The insights gained could improve IBS diagnosis, treatment and therefore patient's well-being.

1.1 Problem Statements

Irritable Bowel Syndrom (IBS) is a functional and benign disorder of the intestine. It is characterized by pain or abdominal discomfort associated with changes in the frequency and the consistency of depositions, either constipation, diarrhea or an alternation of both.

The symptoms of IBS are present in a 10 to 20% of the general population, with a greater incidence in (young) women. IBS seldomly appears after 45-50 years of lifetime. Irritable bowel syndrome accounts for 20 to 50% of Gastroenterology consultations and their cost to society is enormous. The main repercussions are greater school and work absenteeism, big consumption of health resources, limitations in

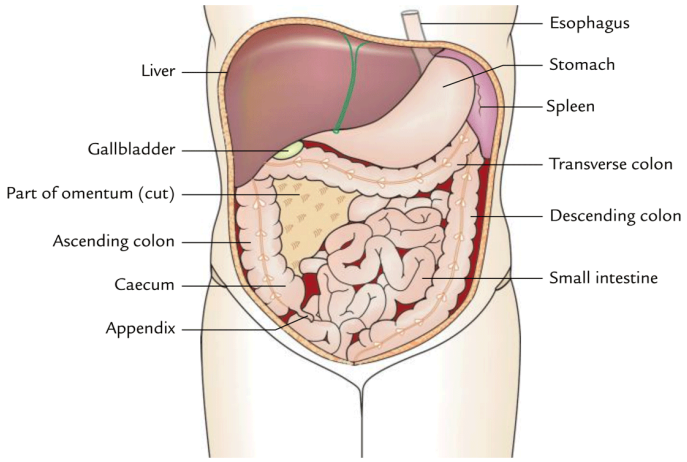

Figure 1: Abdominal intestines[1]

daily activities, inability and a decreased in the quality of patient's life.

The biggest problem of the IBS is that the cause or etiology of this syndrome is unknown. It is therefore the desired outcome, to explore and analyze data derived from a study in patients with irritable bowel syndrome using Experience Sampling Method (ESM). The study has been carried out in the Internal Medicine department in the Maastricht University Medical Center.

The main goal of this research project is to find out which factors are directly related to the symptoms and the pain to discover the principal causes and improve the well-being of the patients.

2 Research Goals

The main intention is to gain insight into correlations between markers provided by the ESM surveys, which are to be discussed in subsection 3.2, and the subjective abdominal pain rating. This is not limited to bivariate measurements, but can include interaction terms. We therefore strive to find a number of correlations under sufficient statistic evidence.

Additionally, we will try to show that these possible correlations are captured by ESM more efficiently than by daily diary questionnaires, which are also to be discussed in section 4.

The research goals can therefore be summarized as following:

- Provide statistical evidence of correlations between abdominal pain and ESM markers
- Provide statistical evidence of the efficiency of ESM in comparison with daily diaries

3 State of the Art

3.1 Irritable Bowel Syndrome

Irritable Bowel Syndrom (IBS) is the most common functional gastrointestinal disorder and it is characterized by recurrent abdominal pain and altered bowel habits, although other symptoms include bloating, cramping, diarrhea and constipation.[3] It is a chronic condition that affects up to 20 % of the population, being twice as frequent among women than men, and occurring typically before the age of 45. [4]

Due to lack of specific biological markers, IBS is a symptom-based diagnosis and reliable symptom assessment methods are essential to investigate natural disease course and evaluate response to therapy. The current standard for diagnosing IBS are the Rome IV criteria, according to which IBS is diagnosed on the basis of recurrent abdominal pain related to defecation or in association with a change in stool frequency or form. Additionally, symptoms must be chronic, occurring at least once per week, on average, in the previous 3 months, with a duration of at least 6 months.

On the basis of the Rome IV criteria, IBS is classified into four sub types, namely IBS with diarrhea (IBS-D), IBS with constipation (IBS-C), IBS with mixed symptoms of constipation and diarrhea (IBS-M), or unsubtyped IBS (IBS-U). [3]

Even though some factors appear to play a role in triggering the symptoms of IBS, such as diet or stress, the exact cause of this condition is still unknown. This is mostly due to the lack of reliable instruments to measure symptoms over time and quantify therapy response.

A multitude of different instruments is currently used to asses chronic abdominal pain in IBS patients, which makes it difficult to compare endpoints of published studies [5]. Furthermore, end-of-day questionnaires, which is the most used method for assessing IBS symptoms, are influenced by recall and ecological bias, i.e. systematic errors that occur when participants do not remember previous events or experiences accurately or omit details and when an inference is made about an individual based on aggregate data for a group [2]. However, the Experience Sampling Method (ESM) may overcome these limitations. [6]

3.2 Diagnosis Systems

At the moment, the standard method to asses IBS symptoms in IBS patients is retrospective self-reported paper questionnaires based on daily or weekly symptom monitoring, from here on referred to as daily/end-of-day diaries. This presents several problems, namely the influence on memory of pain by contextual or emotional and cognitive factors, distorted perception and memory of GI (gastrointernal) symptoms due to the presence of psychological comorbidity and lastly the inability to record temporal dynamics of symptoms. [2]

ESM is a new assessment method characterized by the following criteria: data are collected repeatedly and randomly during the day, for several days, using a digital

[2]https://catalogofbias.org/biases/recall-bias/

device, in the subject's natural environment, with a focus on the subject's momentary physical and mental state and behavior. [2]

According to a study where momentary symptom assessment by electronic ESM were compared with two retrospective paper questionnaires in 26 IBS patients [2], Irritable Bowel Syndrome patients reported higher abdominal pain and flatulence scores in end-of-day diaries, with a tendency of peak-pain reporting at the end of the day, and there were also smaller but significant differences for flatulence, nausea, and belching scores. Furthermore, the study also showed that ESM can provide more insight in the fluctuation of IBS symptoms during the day and may be useful for the identification of potential triggers of symptoms.

Another study [6] also proved that real-time ESM has the potential to capture treatment response more sensitively compared to a retrospective end-of-day GI symptom diary and the GSRS (Gastrointestinal Symptom Rating Scale, which is another disease-specific instrument based on questionnaires addressing 15 items combined into five symptom clusters: reflux, abdominal pain, indigestion, diarrhoea and constipation, that has a recall period of one week and where each item is rated on a 7-point Likert scale), by taking into account day-to-day symptom variability as well as momentary factors that might moderate treatment effect, such as anxiety.

4 Concepts

4.1 ESM

The Experience Sampling Method (ESM), also referred to as Ecological Momentary Assessment, is a new assessment technique, which tries to overcome the limitations of already well established methodologies like daily questionnaires. This assessment method can be characterized by the following criteria: data is collected repeatedly and randomly during the day, for several days, using a digital device, in the subject's natural environment, with a focus on the subject's in-the-moment (i.e. momentary) physical and mental state and behaviour [2].

ESM data for this research was collected 10 times a day, at completely random times with an additional constraint, that no two ESM questionnaires can be withing 15 minutes of each other. The time was constrained to be between 7.30 AM and 10.30 PM. Patients carried a digital device, the PsyMate (comparable to a mobile phone) which would give an auditory signal (a 'beep') that would inform the patient about the occurrence of a new questionnaire. Each questionnaire contains the same set of questions, takes 4-5 minutes on each occasion and patients were asked to rate different symptoms or pain intensity on a scale from 0 to 10. A questionnaire was considered as compliant, if at least 6 out of 10 daily questionnaires were answered, although incomplete questionnaires are still considered in the dataset this project uses. This process was repeated for exactly one week [2].

4.2 Aggregated data

The data consists of 73 patients, in which 37 suffer from IBS and the remaining 36 individuals are healthy and therefore act as a control group. (Figure 2 and 3).

	Population	Male	Female
Sample Age Mean	34.0	40.8	30.8
Netherlands Age Mean	42.6	41.5	43.6
Sample Age Deviation	14.3	12.2	14.3
Sample BMI Mean	24.5	25.7	23.9
Netherlands BMI Mean	25.4	25.9	25.0
Sample BMI Deviation	5.3	3.3	6.0

Figure 2: Patient Demographic Statistics; Source for data on Netherlands: WHO Global Status Report 2014 [7]

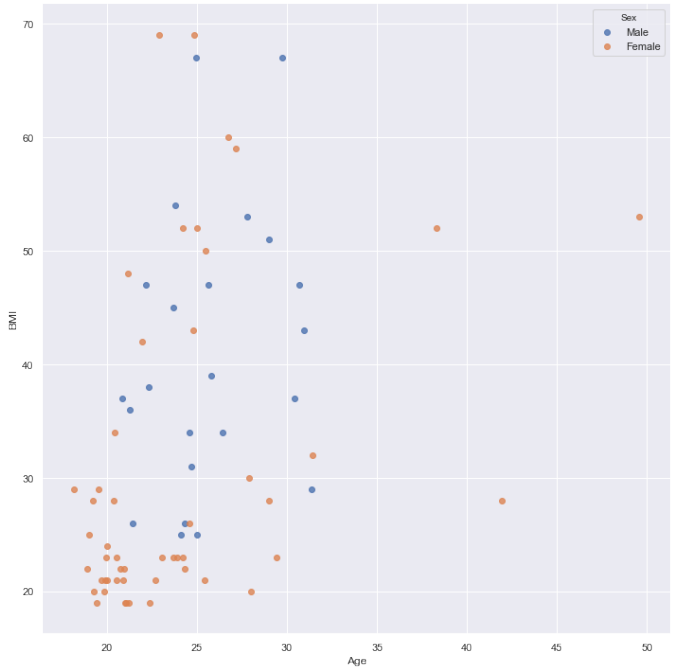

Figure 3: Patient Demographics

In total there are 3528 samples described by 142 features. These features can be separated into three main groups:

- Personal data - every sample consists of discrete data describing the persons

meta data (e.g. age and sex).

- Subjective data - discrete in nature, but in a range of 0 to 10 and describes the persons subjective feeling about specific symptoms (e.g. pain, diarrhea, etc.)
- Categorical data (e.g. specific drugs used) - some information (e.g. medication) can't be easily tracked in a discrete format and is therefore saved as plain text.

A big challenge is posed by the amounts of missing data. Roughly 44% of data entries are missing. Most of these are to be found in features that are not useful for this research, for example drugs for patients, that is not a focus of our project. Nevertheless, still a lot of data has to be imputed.

5 Approach

5.1 Soft- and Hardware

A variety of algorithms and packages are used in order to maximize the insight gained in this project. The Python programming environment provides a vast majority of algorithms needed and is a convenient tool for all sorts of analytic steps. Most of the machine learning algorithms employed in this project are implemented in scikit-learn (0.21.2) [8]. Additionally, R is used to compute statistical methods like explored in sub-subsection 5.5.1

The execution of the code is done on a MacBook Pro (2017), using 8GB of RAM (2133 Mhz LPDDR3), an Intel Core i5 Processor (2.3 Ghz) and the graphic chip Intel Iris Plus Graphics 640 (1536mb).

5.2 Domain Knowledge aggregation

The first step towards gaining insight, is gathering information about the domain of the data. Therefore, we consult the medical experts that supervise the ESM on a regular basis. This particularly includes meetings with Dr. Daniel Keszthelyi, as well as email correspondence to clarify any missing information, on the side of the authors of this project and on the side of Dr. Keszthelyi.

As the schedule of medical professionals is filled to extensive amounts and has to be matched with five authors, questions that arise are aggregated and taken care of when a meeting is possible.

5.3 Handling missing data

As explained in subsection 4.2, the data is of very sparse nature. Disregarding all instances with missing values would lead to almost empty dataset, so other methods of handling missing data are required. Patients who fill in the questionnaire might experience aches of different lengths. If one value is missing, but the surrounding values are indicating pain, it can be assumed that there had to be some kind of ache

during the time in between. Therefore, a moving average is employed to impute missing values. Another method explored is imputation using Principal Component Analysis. Additionally, the nature of the missing values will be researched regarding patterns that eventually emerge, showing a specific trend leading to missing data.

5.4 Exploratory Data Analysis

On the technical side, exploring the data set is mandatory in order to have a proper understanding of what information is available. This is a necessary first step to detect eventual outliers and prepare further usage of the data in an algorithmic sense. The exploration is done by employing multiple methods that are well established in the field of Data Science.

5.4.1 Data Visualization

As it is a key point to understand the sampling methodology employed by Keszthelyi et al., it is vital to understand the demographic data of the patients. Therefore, a visual exploration using histograms and kernel density estimations is employed so demographic data compared to the population of the Netherlands. Additionally, some features like the location of pain mapped to abdominal regions can conveniently be explored in a visual way.

5.4.2 Dimensionality Reduction

Principal Component Analysis (PCA), is applied with the goal of making the data visually more approachable. Therefore, the data analyses with respect to visualizing it in dimensions that can be plotted and visually analyzed. This way, outliers and clusters can be found before doing cluster analysis. This also is a technique used in Feature Selection, see section 5.4.4.

5.4.3 Cluster Analysis

As there could be hitherto undetected clusters in the data, eventually revealing insight into the symptomology of IBS, a cluster analysis is executed. As this is a preliminary exploration, the simple and fast KMeans algorithm is used. In order to find the right amount of clusters, the so called Elbow method [9] is employed.

5.4.4 Feature Selection

As stated in subsection 4.2, the number of features is rather high. It should therefore be a goal of this exploratory analysis to reveal some important features that are to be analyzed in later steps, keeping in mind their eventual correlation with the overall pain that patients experience. Therefore, different methods of feature selection are used. The Lasso Regression, implemented in scikit-learn, adds a L1 regularization term. As this can lead to dissapearing coefficients it can filter out some features that hence would have no correlation with the target variable. Furthermore, correlation matrices are calculated and visualized in the form of heat maps. This enables an intuitive visual way to detect correlations. Additional algorithms will be explored and compared with regards to their ability to describe the target variable, namely abdominal pain.

5.5 Statistical Research

5.5.1 Simplifying the problem

In order to get an overview of the feasability of the tasks proposed in this paper, the target variable will first be discretized into two bins. This will lead to a reduction in complexity, from a regression task to a binary classification task. If the outcome of our research on this data is successfull, we plan to move further to the original regression task of predicting the pain variable.

5.5.2 Multilevel models

As the nature of our dataset is hierarchical, with one patient filling out multiple questionnaires over multiple days, it is important to keep this structure in mind during statistical analysis. Therefore, Multilevel Models are employed, as an extension to General Linear Mixed Models, in order to exploit the nested structure of the data and gain further insight into possible correlations.[10]

5.5.3 Errors-in-variables models

The most important measurement taken during the ESM trial is the rating of abdominal pain, as this will be used as the target variable and give insight into how the patient feels at any given point. The problem, however, is that pain is a very subjective measurement. Hence, the measurement error of this variable, as well as other subjective ratings, must be taken into account. Errors-in-variables models are employed to account for this fact.[11]

6 Social Impact

There are two main areas in which the project could positively contribute to the society. Undoubtedly, our central intention is to find the dependencies between the pain suffered by the patients and the factors which affect it. In case of satisfactory results of the research, we hope it would contribute to the understanding of the relationship between pain factors which cause it. This in turn would help the doctors to propose more effective treatments of IBS. Better hospitalization could eliminate or minimize the ache felt by the patients.

The first societal impact described above could also lead to the second one, which is the evaluation of ESM's effectiveness. This novel method was already proven useful for patient's continuous monitoring and hope that our analysis could lead to a better comprehension of this approach for a statistical and data retrieval purposes. In case of positive outcome of the research, the ESM could gain even more credibility as a tool for treatment reporting. On the other hand, if our main research question remains unanswered, we could still propose the methods for ESM's improvement in terms of data mining and retrieval perspective. This could for example include recommendations how to handle unbalanced data, as the patient's responses are voluntary and their respective response rate varies.

Moreover, the ESM method of reporting could substantially reduce the costs of

healthcare systems in treating the patient's who are affected by IBS. The reporting system, which is fully automated with the software, eliminates the human operators in data collection procedure. As a result, healthcare workers could spend more time on different tasks, which are of more complex nature. Additionally, the amount of appointments a patient has could be reduced drastically, which would also reduce the overall cost of treating IBS in the population.

7 Risk Analysis

7.1 Statistical validity

As the nature of the research conducted in this project is statistical, there are some risks that might arise with regards to statistical validity. First of all, the sample could not reflect the population properly. This might be caused by insufficient amounts of data, some form of bias or missing data inside existing instances.

Insufficient amounts of data can only be overcome by addressing the issue with the domain experts and considering whether settling for a research goal that is achievable in an easier way, therefore less complicated, or getting more data.

If there is some form of bias in the sample, it is important to address this issue by taking sufficient statistical steps. If we encounter any of these, we plan to employ models that can account for such a bias, as explained for example in sub-section 5.5.3. If, however, a selection bias (or similar) fully permits a confident research result, it is important to communicate that to the domain experts and settle exploratory insight instead of exact statistical results.

In case of missing data, there are some strategies that are well established in the statistical and data science communities, such as explained in subsection 5.3. If these strategies can not deliver sufficient results for some variables, it could be possible to explore further algorithms for imputation. If this doesn't lead to acceptable results, the variables must be disregarded.

7.2 Lack of correlation

Another potential risk to the success of this project could be a lack of correlation between the predictors and the target variable. Since, in the context of this research, we only have access to a limited amount of data, finding patterns in this particular data might be a challenging task.

If it isn't possible to find any correlations at all, there should be two main reasons for it. Firstly, feature selection. The amount of data might be sufficient to analyze and find patterns in it, but the features selected during the ESM process might simply have no real correlation. In this case, further analysis will be nearly impossible and for better results the ESM process could be repeated in the future with additional or adjusted features to take a look at. Secondly, data sparsity. The selected features might contain valuable information, but due to the limited amount of accessible entries, finding these correlations might become really difficult. Implementing more advanced imputing techniques like grouping time slots (Moving

Average) or Principal Component Analysis (PCA) might be helpful in overcoming these difficulties.

8 Deliverables

8.1 Statistical report of correlations

The final product of the project would be a report. All the research goals will be explored with regards to their success, together with a detailed description of the methods used, results and discussion of those. Based on the data provided, we expect to find a correlation between the various features and the subjective abdominal pain ratings.

8.2 Comparison with daily reports

Another goal is to prove the usefulness of ESM over daily questionnaires. Therefore, any potential findings with regards to this will be taken into account in the report delivered.

9 Project Plan

As the research done in this project consists of multiple steps, it is important to plan ahead in order to avoid mismanagement.

The project can be divided into several main tasks, denoted by different colors:

- Exploratory tasks (blue)
- Statistical analysis (orange)
- General tasks (green)

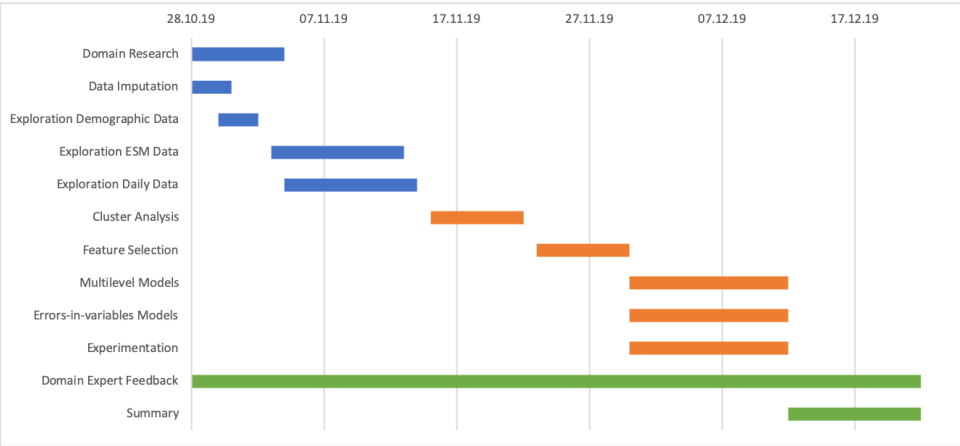

Figure 4: GANTT chart depicting project plan

References

[1] M. Mir-Nasseri H. Vahedi, R. Ansari and E. Jafari. Irritable bowel syndrome: A review article. *Middle East J Dig Dis.*, 2(2):66–77, 2010.

[2] L. Vork-R. Lousberg D. M. A. E. Jonkers D. Keszthelyi M. A. Hesselink T. J. C. Van Schagen J. Van Os A. A. M. Masclee J. W. Kruimel Z. Mujagic, C. Leue. The experience sampling method - a new digital tool for symptom assessment in ibs: an exploratory study. *Neurogastroenterol Motil*, 12624, 2015.

[3] Brian E. Lacy M.D. A. C. Ford, M.D. and Ph.D. N. J. Talley, M.D. Irritable bowel syndrome. *The New England journal of medicine*, 376(26):2566, 2017.

[4] Keszthelyi D. Vork L. and Mujagic Z. Development, content validity, and cross-cultural adaptation of a patient-reported outcome measure for real-time symptom assessment in irritable bowel syndrome. *Neurogastroenterol Motil*, 2017.

[5] Q. Aziz W. Reinisch E. G. Quetglas F. De Leonardis M. Segerdahl A. A. M. Masclee Z. Mujagic, D. Keszthelyi. Systematic review: instruments to assess abdominal pain in irritable bowel syndrome. *Aliment Pharmacol Ther*, 42, 2015.

[6] Drukker M. Vork L., Mujagic Z. The experience sampling method—evaluation of treatment effect of escitalopram in ibs with comorbid panic disorder. *Neurogastroenterol Motil*, 31:e13515, 2019.

[7] World Health Organization. WHO Global Status Report on noncommunicable diseases. 2014.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] R. L. Thorndike. Who belongs to the family? *Psychometrika*, 18(4):267–276, 1953.

[10] Anthony S. Bryk S., Raudenbush W. Hierarchical linear models : applications and data analysis methods. *Journal of Machine Learning Research*, 2002.

[11] Stefanski L. A. Carroll R. J., Ruppert D. and Crainiceanu C. Measurement Error in Nonlinear Models: A Modern Perspective. 2006.