,,

Universiteit Leiden
The Netherlands

# Opleiding Informatica

## Early Detection of Celiac Disease in Children

Tomke Meyer

Supervisors:
Dr. K.J. Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/10/2022

## Abstract

Celiac disease is an autoimmune disorder that affects more than 1% of the population all around the world. The only reliable diagnosis at the moment is an endoscopy or biopsy which is stressful and expensive. The patient has to keep on eating gluten leading up to the biopsy and has to undergo anesthesia for the procedure. To help diagnosing celiac disease without the inconvenience of a biopsy, a questionnaire has been developed for filtering for celiac disease. This research examined the correlation of multiple symptoms with the risk of having celiac disease, using data from 3014 patients. This had been done by analyzing data of a survey over 9 different symptoms. The results show that abdominal distension, diarrhea, constipation and fatigue are corresponding with a higher risk for having celiac disease, while symptoms like vomiting are not crucial for the diagnosis. The results have been visualized in a web application.

# Contents

# 1 Introduction

## 1.1 Medical background

### 1.1.1 Celiac disease

Celiac disease (CD) is an autoimmune disorder caused by the ingestion of gluten in genetically predisposed individuals and is one of the most common chronic diseases, affecting about 1% of the population. Gluten is a protein found in grains like wheat, rye or barley, and its ingestion causes inflammation of the small intestine. In recent years there have been many changes in the diagnosis, pathophysiology and natural history of this condition as well as a constant increase in the number of identified diagnoses. This has led to greater availability of sensitive and specific screening tests to identify risk groups for CD and also to earlier diagnoses and more awareness of the disease. But although there have been more diagnostic and medical developments in recent years, there is not yet a cure found. The following paragraphs explain the immune response to gluten, the genetic influence and environmental triggers.

#### 1.1.1.1 Pathophysiology

The gluten protein consists of different components including gliadin (in wheat), an alcohol-soluble part that contains most of the toxic components. Undigested molecules of gliadin can not be degraded by the gastric, pancreatic and intestinal brush-border membrane proteases. In individuals with celiac disease, gliadin interacts with the intestinal cells which activate the disassembling of the inter-enterocyte tight junction. This disablement of the tight junctions results in an up-regulation of zonulin, a peptide that is involved in the regulation of tight junctions and leads to increased intestinal permeablity. Because of that, partially digested gliadin fragments can pass through the epithelial barrier of the intestine and activate T-lymphocytes that are located in the lamina propria. These activated CD4+-T-lymphocytes produce very high levels of pro-inflammatory cytokines, inducing either a T-helper-1 pattern dominated by interferon-$\gamma$ or a T-helper-2 pattern, which causes a clonal expansion of B-lymphocytes that subsequently differentiate in plasma-cells secreting anti-gliadin and anti-tissue-transglutaminase antibodies. Some of the gliadin peptides, that are not recognized by the T-lymphocytes, activate antigen-presenting cells and intestinal epithelial cells. This leads to an increased expression of interleukin-15, resulting in the activation of intraepithelial lymphocytes that express the activating receptor NK-G2D, which is a natural-killer-cell marker. These activated cells become cytotoxic resulting in the killing of enterocytes.[GC07][CVS$^+$19]

#### 1.1.1.2 Genetic and environmental influence

There is a lot of genetic influence in the possibility of getting celiac disease as different factors like occurrence in the family increase the possibility of developing celiac disease. The prevalence for celiac disease with an affected first-degree relative is around 10% to 15%, with type-1 diabetes 3% to 16% and for women the prevalence is approximately two times higher. This has to do with the genetics behind the disease as it only develops in individuals with alleles that encode for the HLA-DQ2 or HLA-DQ8 proteins. More than 90% of the affected individuals carry the gene for HLA-DQ2 molecules and the rest for HLA-DQ8. Although around 30% of all people carry these alleles, only about 1%-3% develop celiac disease, so the presence is a necessary but not sufficient factor for the development of the disease.

Another possible influence in the development of celiac disease are the environmental factors, like the feeding pattern in the first years and viral infections. Different studies [C$^+$06] [PIO02] suggest that there is a protective effect of breastfeeding and also the introduction of gluten corresponding to weaning. The introduction of gluten before 4 months of age is associated with an increased risk, while the introduction of gluten after 7 months is associated with a marginal risk. Still, the overlap of breastfeeding with the introduction of gluten may be a bigger factor in minimizing the risk of celiac disease. The occurrence of rota-virus and other gastrointestinal infections can also increase the risk of children getting celiac disease [CVS$^+$19].

#### 1.1.1.3 Diagnosis

Diagnosis of celiac disease is often difficult and there is still a lack of awareness about the different presentations of celiac disease and its diagnostic criteria. The symptoms vary from patient to patient and the disease can occur at any age which makes it very difficult to diagnose celiac disease early on. There are however two peaks, one in the first 2 years of life after weaning, as then gluten is introduced to the diet, and in the second or third decades of life. This research was done only with children between 1 and 4, thus focusing on the first peak. The clinical presentation of celiac disease can best be classified into intestinal and extraintestinal. The intestinal form is mostly detected in children and is based on symptoms like diarrhea, loss of appetite or abdominal distention. Extraintestinal symptoms are for example vitamin, calcium or iron deficiencies causing microcytic anemia or osteoporosis. The standard celiac disease diagnosis is a combination of both a serological test, such as for anti-tTG antibodies, and an intestinal biopsy. But all antibody tests can not provide a specificity of 100%, which means that a biopsy is still necessary. Often a ”four out of five” rule is used, stating that four out of five criteria are sufficient to diagnose the patient with celiac disease. These five criteria are: 1. typical signs and symptoms, 2. antibody positivity, 3. HLA-DQ2 or HLA-DQ8 positivity, 4. intestinal damage, 5. clinical response to a gluten-free diet. [CVS+19]

So in most cases, celiac disease develops early in life but because of a great variety of symptoms, it is often unrecognized. These symptoms include the most common malabsorption, diarrhea, poor growth in children and weight loss but also other symptoms like chronic fatigue, irritability or vomiting. When untreated, celiac disease can lead to many long-term complications like malnutrition, osteoporosis, other autoimmune diseases or even cancer. Up to this day, there is no cure for the disease but people with celiac disease can be treated by following a strict gluten-free diet. This means that to prevent severe health problems, an early diagnosis and treatment of celiac disease is the only way. But early diagnosis can only be achieved by active case finding, meaning testing all patients with celiac disease-associated symptoms.

## 1.2 Research Background

### 1.2.1 Related Work

Celiac disease was severely underdiagnosed in children in the Netherlands according to the Leiden University Medical Center (LUMC) research group in 1999. For every diagnosed child there were seven children where the disease was unrecognized and therefore untreated [CMvB+99]. Back then, celiac disease was still a mostly unknown and uncommon disease and there were not many ways to rightfully recognize the disease. Because of CD being unrecognized that often, there was another study from the Dutch Paediatric Surveillance Unit (DPSU) from 2010 to 2013 to gain insight into the prevalence of diseases in youths [QPW+14]. During this period all Dutch pediatricians were asked to report new cases of selected conditions, in this case CD, on a monthly basis, as well as a questionnaire collecting patient information. To further research how to help with earlier detection of celiac disease, the LUMC started a new research project called GLUTENSCREEN[M+21b]. Similar research with children and adolescents was also done in five other European countries (Croatia, Germany, Hungary, Italy and Slovenia), using a questionnaire to analyze the clinical presentation of CD, resulting in abdominal pain as the most common symptom overall, but abdominal distensin and diarrhea in children younger than 3 [RDLD+21].

### 1.2.2 GLUTENSCREEN research

This research is based on the data collected as part of the GLUTENSCREEN study. This LUMC study aims to develop a new case-finding project to identify celiac disease as early as possible in 1 to 4-year-old children. This project started on 4 February 2019 and will end on 1 February 2023. The parents/legal guardians of children in the chosen age group, patients at the Youth Health Care Centres (YHCCs) Kennemerland in the Netherlands, filled in a standardised questionnaire with CD-related symptoms and marked whether the children had had these symptoms or not. The symptoms or signs are: (1) abdominal pain, (2) abdominal distension, (3) constipation, (4) diarrhea, (5) vomiting, (6) fatigue, (7) aphthous, (8) moodiness or irritation and (9) growth restriction, with weight and growth being controlled at the YHCCs. Additionally, the family case history was checked. The chosen CD-related symptoms are based on the recommendations of CD testing in symptomatic children and adolescents in the Guideline Coeliac Disease of the European Society for Pediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN). There is a national control group based on data reported by the National Dutch Paediatric Surveillance Unit (DPSU). If the child had one of these symptoms a point-of-care-test (POCT) was performed to detect CD-specific antibodies (tissue transglutaminase(TG2A)). If the POCT was positive the child was referred to the hospital for a definitive diagnosis by biopsy. [M+21a]

## 1.3 Research questions

For this thesis, the dataset from the GLUTENSCREEN project has been analyzed. The dataset contains all the information concerning the CD-related questionnaire and demographic details for around 15000 patients. In addition, the dataset contains the result of the POCT and whether the child has already been diagnosed with CD or not. The goal of this research is to find out whether the GLUTENSCREEN project is feasible and if the questionnaire is useful for an early detection of celiac disease. So to determine if the project should keep on running, it is important to analyse the data and use the data to answer questions about celiac disease in children. These are the two main research questions:

**RQ1:** Is there a correlation between specific symptoms and celiac disease?

Here it is important to compare the symptoms of non-celiac and celiac patients. So one subquestion to answer is, what is the probability of a child having celiac disease in regards to different symptoms. It is also important to have a closer look at the more common symptoms for celiac disease and if they are characteristic for the disease or if they are as common as in the non-celiac group. Additionally, many of the symptoms by themselves are non-specific for celiac disease and could be an indication of a number of different disorders, so it is important to look into those as well.

**RQ2:** Is there a significant difference in risk of CD for different age groups and different genders?

The age group for the GLUTENSCREEN project is 1 to 4 years. It would be interesting to look at the differences in symptoms and the relevance for the probability of CD and the age of the children. The children from the first age group (from 1-2 years) for example cannot report their symptoms and according to the literature, there is a lower percentage of affected children. So by grouping the data in different age groups and analyzing that data, the question if GLUTENSCREEN should keep on screening very young children or not, could be answered. Additionally there could be big differences between the symptoms boys and girls express, leading to the question if there should be different questionnaires according to the gender.

# 2 Methods

This section describes how the research was done. The processes of data pre-processing, data analysis, clustering and the prediction model will be fully explained in their subsections. For the data pre-processing and the data analysis Python (`https://www.python.org`, release 3.6.9)was used and the web application was written in Python as well, using Shiny for Python (`https://shiny.rstudio.com/py`).

## 2.1 Dataset

For the Glutenscreen-project the researchers received a dataset from the Youth Health Care Centers Kennemerland which has been analyzed for this thesis. This dataset contains all the information concerning the CD-related questionnaire, thus the answers from the parents of the participants visiting the YHCC, and also demographic details of the participants. The answer choices for the different symptoms were only yes or no, depicted in the dataset as "16.0" for yes and "17.0" for no, and the dataset contains the data of around 17000 participants. Some other information in the dataset includes the time of being in care at the YHCC, other related diseases, already existing CD diagnosis, the family case history and most importantly the result of the POCT. A biopsy was done on the children with a positive POCT, resulting in three cases that didn't have celiac disease. These were added to the symptomatic group and not the celiac group.

## 2.2 Data pre-processing

The dataset used in this research contained a lot of additional information that was not necessary for this research. So most columns were not used, the important columns were the different symptoms, birth date, the date on that the survey was done, gender and the testresults. Some parents did not want to take part in the survey but were added to the dataset. These were also not included in the research as there was no further usable information. To answer the second research question the age of the children was needed, so a new column with the age was added, calculated from the birth date and the day the survey was conducted. In most cases the results of the diagnosis were later added to the dataset, which resulted in many participants having two rows with data that had to be merged together. For easier use, some columns were also renamed. After removing the patients that didn't take part in the survey, a total number of 15358 participants remained. Table 1 below shows the number of participants, number of symptomatic participants and the number of patients with celiac disease also split by gender.

| | |
|---|---|
| Total number of participants | 15358 |
| Number of participants with 1 or more symptoms | 3014 |
| Number of patients with CD | 58 |
| Patients with CD (w) | 41 |
| Patients with CD (m) | 17 |

Table 1: Participants involved in the study

## 2.3 Data analysis

To analyze the data and answer the research questions, different methods have been used. To get an overview of the data, the instances of the symptoms in both the celiac group and the symptomatic group have been counted. To put the different symptoms as well as the two groups (celiac and symptomatic) in comparison, the occurrences in percent were calculated. The symptomatic group includes the celiac patients as well, as they are all symptomatic as well. For further research in the age groups 0-1, 1-2, 2-3 and 3-4, the occurrences in those ages were counted and then the relative frequency of these symptoms was calculated. This was done to better visualize the development, as well as split into girls and boys. Table 2 gives an outline of the numbers of participants per age group involved in the study and table 3 shows the participants per age group and per gender.

| Agegroups | Participants with CD | Symptomatic participants |
|---|---|---|
| 0-1 | 1 | 104 |
| 1-2 | 11 | 1041 |
| 2-3 | 21 | 518 |
| 3-4 | 25 | 1416 |

Table 2: Number of participants in the age groups

## 2.4 Relative frequency

To compare the different symptoms and their correlation to celiac disease, the relative frequency of the symptoms in different age groups was calculated. The relative frequency is the ratio of the frequency of the symptoms in the celiac group

| Agegroups | Girls with CD | Symptomatic Girls | Boys with CD | Symptomatic Boys |
|---|---|---|---|---|
| 0-1 | 1 | 41 | 0 | 63 |
| 1-2 | 9 | 505 | 2 | 536 |
| 2-3 | 13 | 235 | 8 | 283 |
| 3-4 | 18 | 690 | 7 | 726 |

Table 3: Number of participants in the age groups split by gender

to the entire symptomatic group. To calculate it, the following formula was used:

$$P(CD) = \frac{\text{Occurrence with CD}}{\text{Occurrence in symptomatic group}}.$$

This is the best way to estimate the specificity of a symptom or a combination of symptoms for celiac disease with the available data and methods. Although it is estimated with a small dataset the calculated ratios could be seen as a probability of having celiac disease when having specific symptoms.

## 2.5 $k$-means Clustering

To better understand which symptoms are more typical for celiac disease and also in which age groups they occur, $k$-means clustering was used on the data set. Clustering is used to classify data into a set of categories or clusters. This means that the data in each subset shares some common features. The $k$-means clustering method is an evolutionary algorithm and the most commonly used clustering method in medical fields [BG13]. It is used to find clusters that have not been explicitly labeled in the data. The algorithm works by having $N$ data points and a number of clusters $k$ as an input, and then giving an output of the data points classified into $k$ subclusters. These are the steps taken by the algorithm:
1: Initialize $k$ random means
2: Associate each data point in $N$ with the nearest mean, resulting in $k$ clusters
3: Recalculate the position of the means according to the found clusters
4: Repeat steps 2 and 3 until the clusters converge. [ZHAS$^+$21]

In this project the

`sklearn.cluster.KMeans`

was used [Lea]. It accepts a 2D-array of how often the symptom occurs per age group together with the relative frequency per age group and symptom. To determine the optimal $k$, the elbow method was used. It is a method that runs the k-means clustering method for a range of clusters $k$(e.g. from 1 to 5) and calculates the sum of squared distances from each point to the assigned centers for each value. These so-called distortions are then plotted and the inflection point or 'elbow' is the best value for $k$. The figure 1 below shows the result of the elbow method for the k-means clustering in this project.

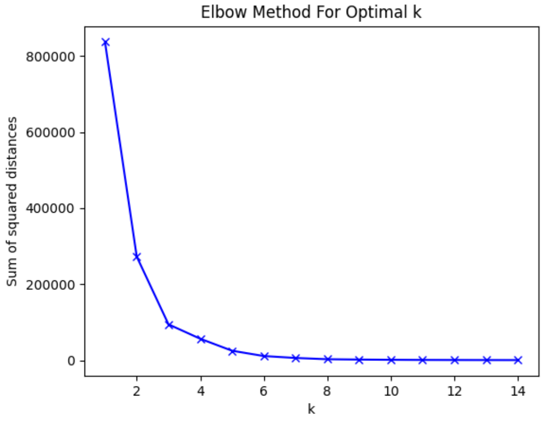


Figure 1: Elbow method to determine the best $k$

From the elbow method, $k = 3$ was chosen as that is the point where the line starts to become more linear. It was then used to compute the $k$-means clusters from the 2D-array with

```
KMeans(n_clusters=3, random_state=0).fit(coordinates)
```

and coordinates being the 2D-array.

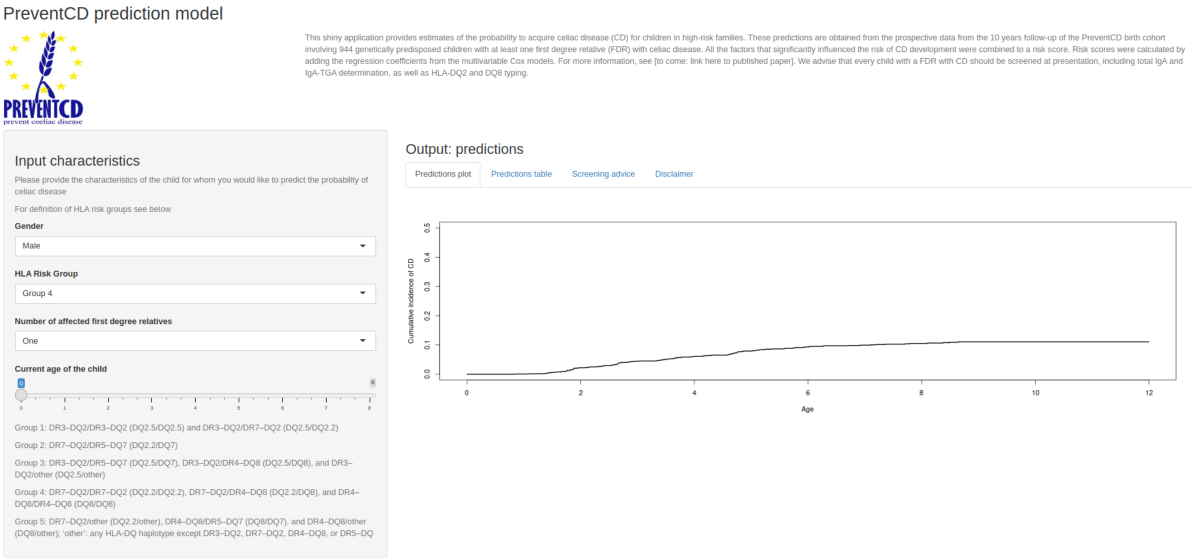
## 2.6 Heatmap of symptom combinations

One of the sub-research questions was, which combinations of symptoms are more typical for celiac disease, as most patients do not suffer from only one symptom but a multitude of symptoms. To display the frequency of these combinations a heatmap was used. A heatmap is a plot of data as a color-encoded matrix. In this project, a 2D-array was used as input for the heatmap-function from the Seaborn Data Visualization library. The 2D-array consists of all symptoms on the x-axis and all symptoms on the y-axis resulting in a heatmap that shows the number of occurring combinations of symptoms as further explained in the results section.

```
ax2 = sns.heatmap(dataheatmap, xticklabels=x_axis_labels,
yticklabels=y_axis_labels, annot=True)
```

## 2.7 Prediction Model

After analysing the data, the results have to be visualized in a prediction model that shows the relative frequency of celiac disease, given a symptom, and gives an estimate of the probability to express celiac disease. The requirement for the prediction model is an easy-to-use tool to compare different symptoms as well as the differences between the genders. The figure 2 below shows a prediction model built by Hein Putter from LUMC for the PreventCD study[HEAR+10], which has been used as an inspiration for the prediction model in this research.



PreventCD prediction model

This shiny application provides estimates of the probability to acquire celiac disease (CD) for children in high-risk families. These predictions are obtained from the prospective data from the 10 years follow-up of the PreventCD birth cohort involving 944 genetically predisposed children with at least one first degree relative (FDR) with celiac disease. All the factors that significantly influenced the risk of CD development were combined to a risk score. Risk scores were calculated by adding the regression coefficients from the multivariable Cox models. For more information, see [to come: link here to published paper]. We advise that every child with a FDR with CD should be screened at presentation, including total IgA and IgA-TGA determination, as well as HLA-DQ2 and DQ8 typing.

Input characteristics

Please provide the characteristics of the child for whom you would like to predict the probability of celiac disease

For definition of HLA risk groups see below

Gender: Male

HLA Risk Group: Group 4

Number of affected first degree relatives: One

Current age of the child: 0

Group 1: DR3–DQ2/DR3–DQ2 (DQ2.5/DQ2.5) and DR3–DQ2/DR7–DQ2 (DQ2.5/DQ2.2)

Group 2: DR7–DQ2/DR5–DQ7 (DQ2.2/DQ7)

Group 3: DR3–DQ2/DR5–DQ7 (DQ2.5/DQ7), DR3–DQ2/DR4–DQ8 (DQ2.5/DQ8), and DR3–DQ2/other (DQ2.5/other)

Group 4: DR7–DQ2/DR7–DQ2 (DQ2.2/DQ2.2), DR7–DQ2/DR4–DQ8 (DQ2.2/DQ8), and DR4–DQ8/DR4–DQ8 (DQ8/DQ8)

Group 5: DR7–DQ2/other (DQ2.2/other), DR4–DQ8/DR5–DQ7 (DQ8/DQ7), and DR4–DQ8/other (DQ8/other); ‘other’: any HLA-DQ haplotype except DR3–DQ2, DR7–DQ2, DR4–DQ8, or DR5–DQ

Output: predictions

Figure 2: Example prediction model, used as a foundation for the new prediction model

To build the web application for the prognosis of celiac disease, the Shiny application for Python from Rstudio has been used. Shiny applications consist of two different parts, the user interface and the server function.

The user interface has to be easy to use and should give different choices on what to display in the prediction model. The server part contains all data calculations for the model, divided into reading the input and assigning the corresponding variables, calculating the frequencies for all possible inputs, rendering a prediction plot and rendering a prediction table.

To calculate all necessary frequencies, for-loops are used to count the occurrences of the chosen symptom or of the chosen symptom combinations in the chosen group (celiac, boys and celiac, girls and celiac) and the corresponding symptomatic group. This procedure supports the choice of one to nine symptoms. The resulting occurrence lists were divided by each other to get the relative frequency used for a prediction plot and table.

# 3 Results

The following section presents the results of the data analysis where different aspects of the data were examined using the methods described above. For the examination different measures of correlation were used. The occurrence of a symptom reflects the number of patients having said symptom, used in both, the celiac and the symptomatic group. The percentage of a symptom says how often a symptom occurs in comparison to the group. For most comparisons between different symptoms or different genders the relative frequency was used. Only the heatmaps use the relative frequency as a percentage for better readability.
Table 2 shows the occurrence of specific symptoms in patients with celiac disease and the occurrence of specific symptoms in the entirety of the symptomatic patients in the study. A patient was counted as part of the celiac group if both the POCT and the biopsy were positive, while every patient with at least one symptom is part of the symptomatic group.

| | Symptom | Participants with CD | | Symptomatic participants | |
|---|---|---|---|---|---|
| | | Occurrence | Percentage | Occurrence | Percentage |
| 1 | abdominal pain | 13 | 22.41% | 484 | 16.06% |
| 2 | abdominal distension | 23 | 39.65% | 662 | 21.96% |
| 3 | constipation | 5 | 8.62% | 290 | 9.62% |
| 4 | diarrhea | 7 | 12.07% | 239 | 7.93% |
| 5 | vomiting | 0 | 0.0% | 56 | 1.86% |
| 6 | fatigue | 9 | 15.52% | 271 | 8.99% |
| 7 | aphthous | 2 | 3.45% | 92 | 3.05% |
| 8 | moodiness or irritation | 13 | 22.41% | 468 | 15.53% |
| 9 | growth restriction | 42 | 72.41% | 1632 | 54.15% |
| | Total | 58 | | 3014 | |

Table 4: Occurrence of symptoms and their percentage

The following figure 3 is a representation of the occurrence of the different symptoms in symptomatic patients with and without celiac disease, displayed as a percentage of occurrence in their respective groups. It clearly illustrates that in the celiac group a higher percentage suffers from most symptoms.

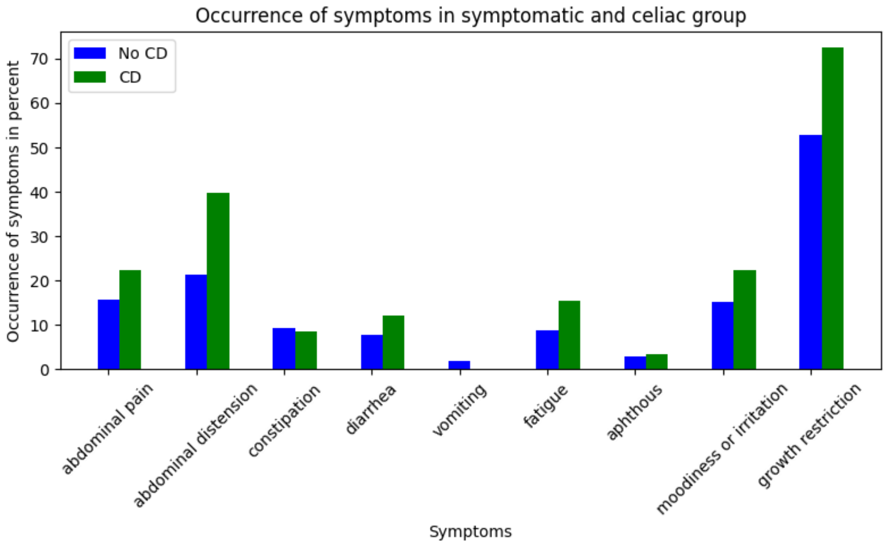


Figure 3: Occurrence of symptoms in the symptomatic group and in CD-patients

From the figure above it seems that the two most prominent symptoms in patients with celiac disease are growth restriction and abdominal distension. This could lead to the assumption that these two symptoms are the most significant for celiac disease, but after looking at the entire symptomatic group it is clear that they are also the most common there. To prevent misreading these statistics it is important to calculate the relative frequency of these symptoms and their correlation to celiac disease. This gives a different insight into which symptoms have a higher specificity for celiac disease. To determine the relative frequency $P(CD)$ for each symptom the following formula was used:

$$P(CD|A) = \frac{\text{Occurrence of symptom A in patients with CD}}{\text{Occurrence of symptom A in patients in symptomatic group}}.$$

Table 5 shows both the relative frequency of the symptoms in patients that have CD but also the percentage of the celiac group that has said symptom.

| | Symptom | Relative frequency | Percentage of occurrence |
|---|---|---|---|
| 1 | abdominal pain | 0.026 9 | 22.41% |
| 2 | abdominal distension | 0.034 7 | 39.65% |
| 3 | constipation | 0.017 2 | 8.62% |
| 4 | diarrhea | 0.029 3 | 12.07% |
| 5 | vomiting | 0.0 | 0.0% |
| 6 | fatigue | 0.033 2 | 15.52% |
| 7 | aphthous | 0.021 7 | 3.45% |
| 8 | moodiness or irritation | 0.027 8 | 22.41% |
| 9 | growth restriction | 0.025 7 | 72.41% |

Table 5: Possibility of having celiac disease when having one symptom and the percentage of the symptom occurring within the celiac group

From the table above it is obvious that most of the CD patients have growth restriction (72.41%) and abdominal distension (39.65%). But when comparing with the entire symptomatic group and calculating the relative frequency, it is clear that while growth restriction is very common among celiac disease patients it is not an identifying symptom as it is also a very common symptom in the entire symptomatic group. The symptoms with the highest relative frequency are abdominal distension (0.0347), fatigue (0.0332) and diarrhea (0.0293), although abdominal pain, moodiness and growth restriction are still relatively frequent. In comparison, abdominal pain and moodiness or irritation are occurring more in the celiac group than fatigue and diarrhea, but this means they also occur more in the entire symptomatic group.
Figure 4 shows this as well, now split into the different age groups 0-1, 1-2, 2-3 and 3-4 years.

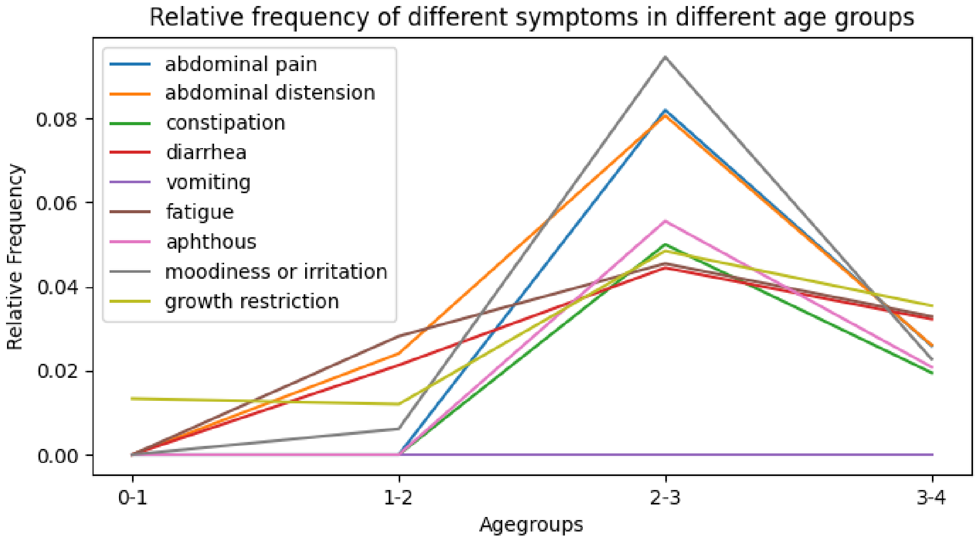


Figure 4: Relative frequency of symptoms per age group

There is a peak in symptoms at age 2-3 for all symptoms, with moodiness, abdominal distension and abdominal pain having the highest peak. Interestingly both fatigue and diarrhea don't show such a high peak and their occurrence is rather evenly distributed among the ages from 0-4.

Figures 5 and 6 show the relative frequencies of the different symptoms for patients having CD split on gender. The girls are represented in figure 5 and the boys in figure 6. There are many interesting differences, as the frequency of symptoms is much more distributed for the girls among the age groups 1-2 and 2-3. The most decisive factors are diarrhea, abdominal distension, fatigue and moodiness or irritation. This is very different from the boys as there are almost only cases in age group 2-3 and a few in age group 3-4. Here the symptoms with the highest relative frequency are aphthous, abdominal distension, moodiness or irritation and abdominal pain. Interesting are also differences in least occurring symptoms, as both do not suffer from vomiting, while girls also don't seem to have aphthous, which has a rather high frequency in boys. Additionally both constipation and diarrhea have a very low frequency in boys, with only a few occurrences in the last age group.

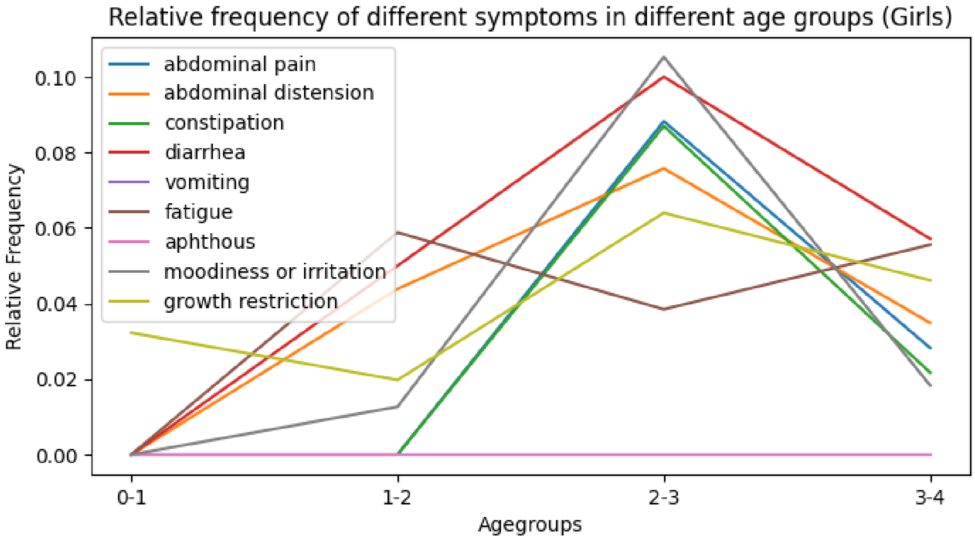


Figure 5: Relative frequency of symptoms per age group

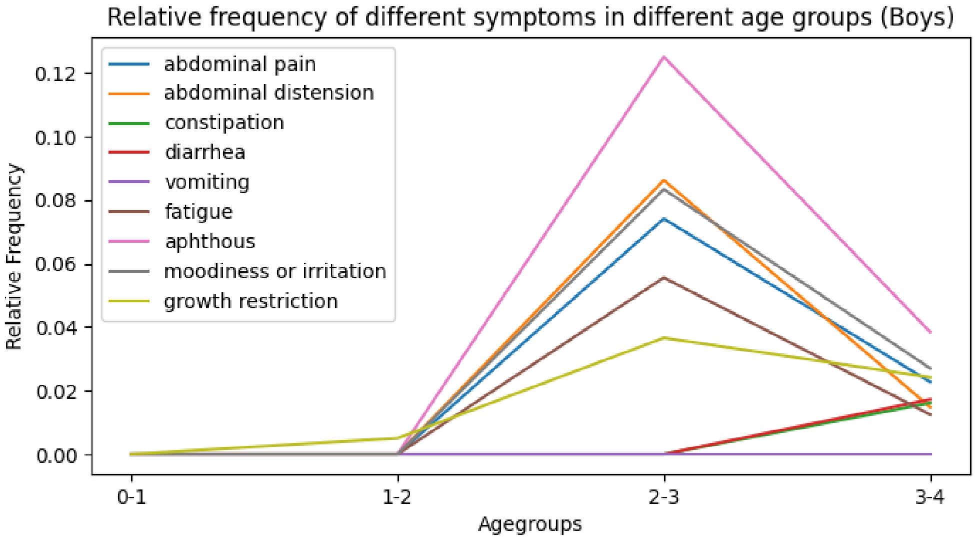


Figure 6: Relative frequency of symptoms per age group

Figures 7 and 8 are the results of the $k$-means clustering method and classify the symptoms that occur into different subclusters. For this depiction abbreviations were used to make the figures more readable. The assigned abbreviations can be seen in table 4 below. The $k$-means clustering assigned horizontal clusters, which resulted in these figures not giving much additional information.

| | Symptom | Abbreviation |
|---|---|---|
| 1 | abdominal pain | ap |
| 2 | abdominal distension | ad |
| 3 | constipation | c |
| 4 | diarrhea | d |
| 5 | vomiting | v |
| 6 | fatigue | f |
| 7 | aphthous | a |
| 8 | moodiness or irritation | mi |
| 9 | growth restriction | gr |

Table 6: Participants involved in the study

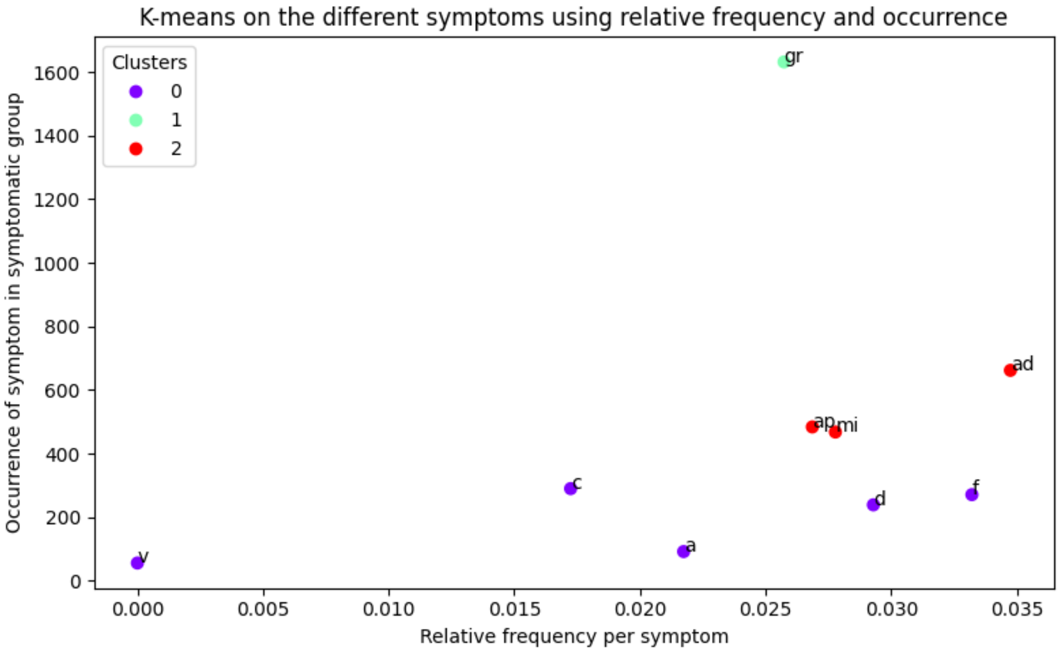


Figure 7: K-means clustering of symptoms

The $k$-means clustering algorithm was used on a dataset depicting the relative frequencies per symptom against the occurrence of the symptoms in the symptomatic group. The further a symptom is on the right side of the x-axis, the more specific it is for celiac disease while being higher on the y-axis means it is a more frequent symptom in the symptomatic group. So this means that for example abdominal distension and fatigue are very CD-specific, while growth restriction is rather common for both the CD group and the symptomatic group making it less decisive in a diagnosis. The $k$-means algorithm split the symptoms into three subclusters, 0: growth restriction; 1: abdominal pain, moodiness or irritation, abdominal distension; 2: constipation, diarrhea, fatigue, vomiting, aphthous. Subcluster 0 represent the outlier in symptoms as growth restriction occurs very often in the entire symptomatic group but is also not a CD-specific symptom. The symptoms in group 2 are occurring less in the symptomatic group but are also relatively uncommon and not specific for celiac diseases. Cluster 1 then depicts the common and specific symptoms of celiac disease with a certain variation between the clusters.

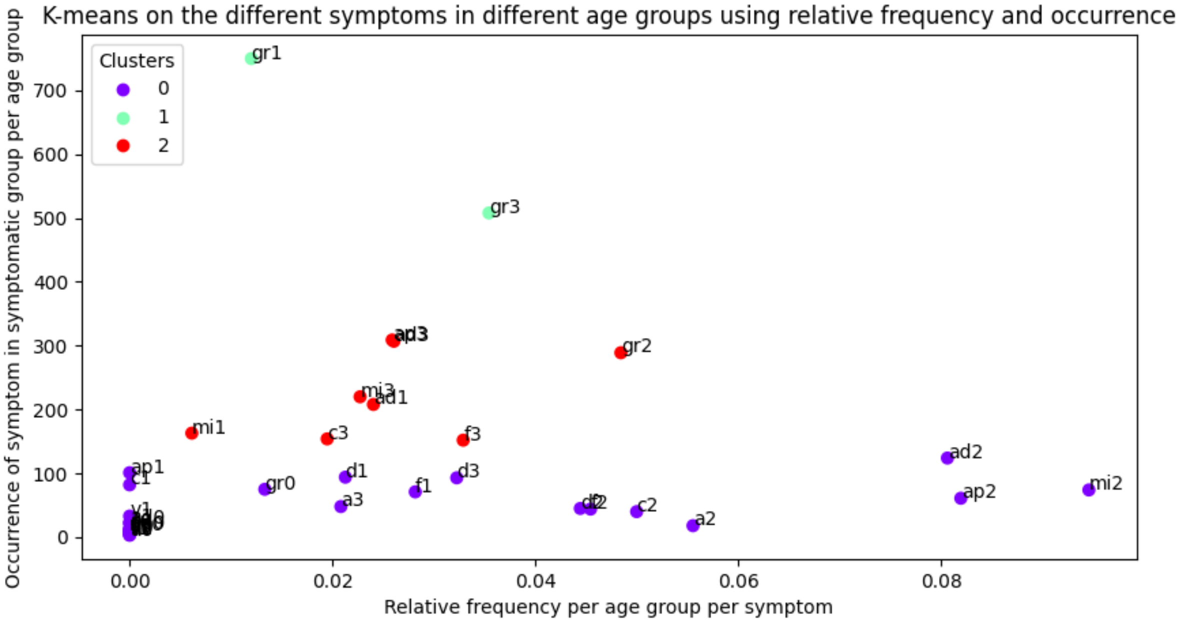


Figure 8: K-means clustering of symptoms in different age groups

To further look into the different symptoms and their specificity, the dataset was split into the four age groups. The result can be seen in figure 8. Here it is very obvious that in the age group 2 (2 to 3-year-olds) there is a much higher relative frequency for the occurrence of most symptoms while also not occurring that much in the entire symptomatic group. This leads to the assumption that in this group the symptoms are both occurring more and being better detectable.

The previous results all depicted the occurrence and relative frequency of all symptoms within the celiac group. To look further into which symptoms have a higher correlation with celiac disease, the symptoms were mapped against each other. This is shown in the figure 9 below, which shows how many celiac patients suffer from two different symptoms. From this heatmap it is clear that growth restriction is a very dominant feature and occurs more in frequent combinations, for example together with abdominal distension, abdominal pain and moodiness.

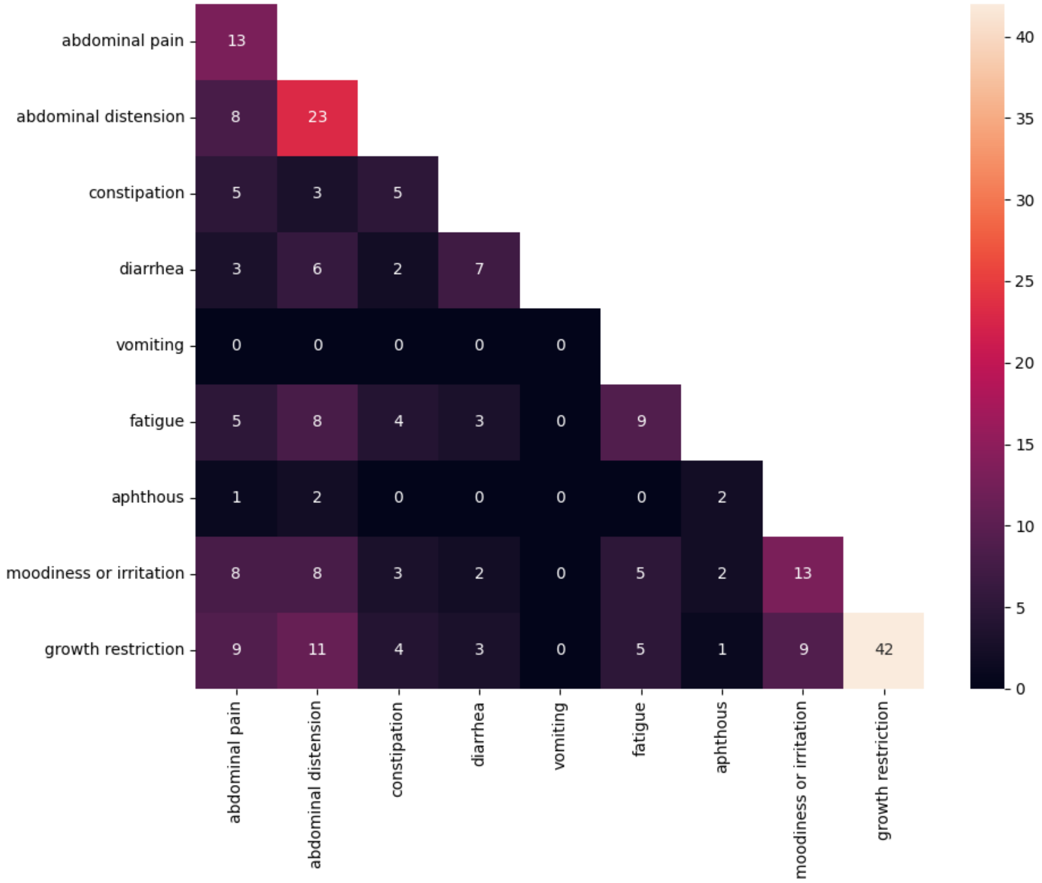


Figure 9: Heatmap of correlation of the different symptoms

To better compare the different symptoms, again the relative frequency (in percent) per combination of symptoms was calculated, resulting in figure 10. Interestingly the combination with the highest frequency is diarrhea and constipation (10.53%), which are two symptoms that medically rarely occur together but sometimes occur alternating as result of an irritated bowel [Cam21]. Other combinations with rather high frequencies are fatigue and constipation (7.02%), fatigue and abdominal distension (6.72%), diarrhea and abdominal distension (6.52%), and moodiness or irritation and aphthous (6.06%). These are the combinations that occur relatively frequent in the celiac group compared to the entire symptomatic group, making them more typical for celiac disease.

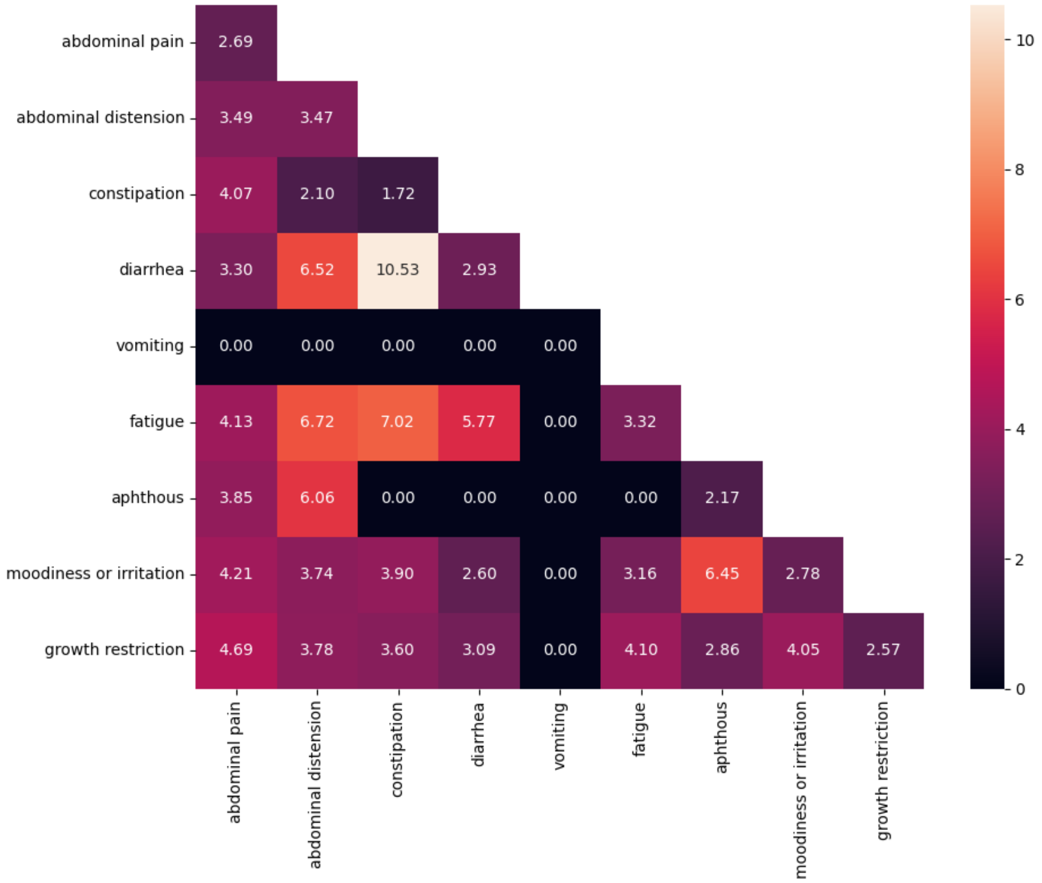


Figure 10: Heatmap of relative frequency in percent of correlation of the different symptoms

Figure 11 and 12 also show the relative frequencies in percent for the different symptom combinations, but now split by genders. This makes it possible to further examine celiac disease specific symptoms and the clear differences between boys and girls. Although figure 4 and 5 already showed those immense differences between boys and girls, these figures show which combinations of symptoms are more specific for which gender. For girls the combination with the highest frequency is again diarrhea and constipation (16.67%), followed by diarrhea and abdominal distension (13.89%), fatigue and diarrhea (9.52%), and fatigue and abdominal distension (9.23%). This differs a lot from the boys where the combination with the highest frequency is aphthous and abdominal distension (12.5%), followed by moodiness or irritation and aphthous (11.11%), aphthous and abdominal pain (8.33%), and constipation together with diarrhea or fatigue (7.69%).

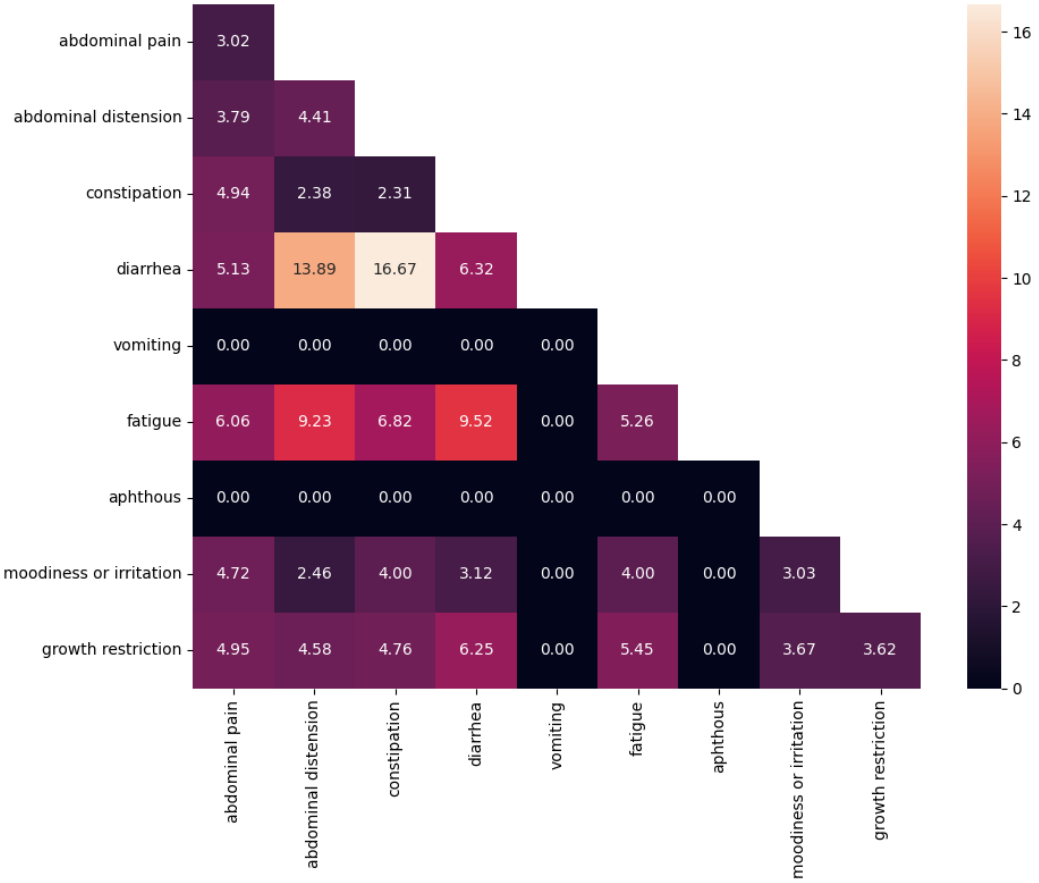


Figure 11: Heatmap of relative frequency in percent of correlation of the different symptoms (Girls)

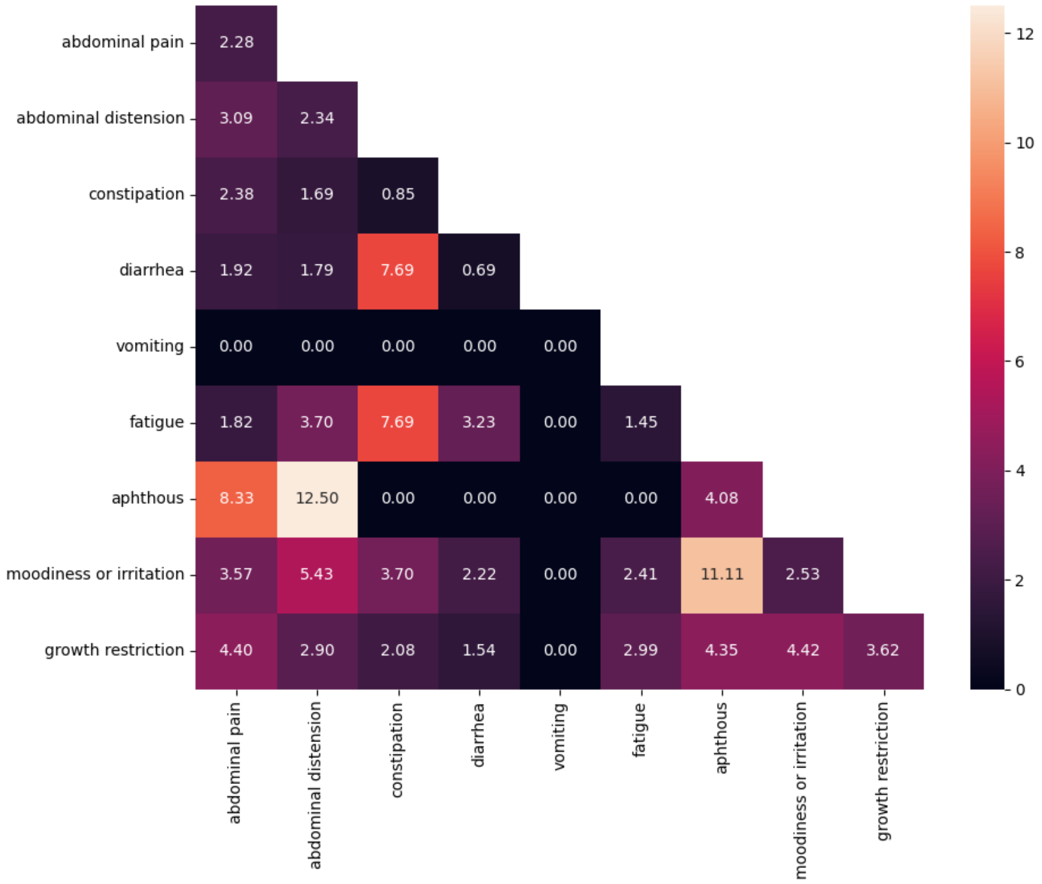
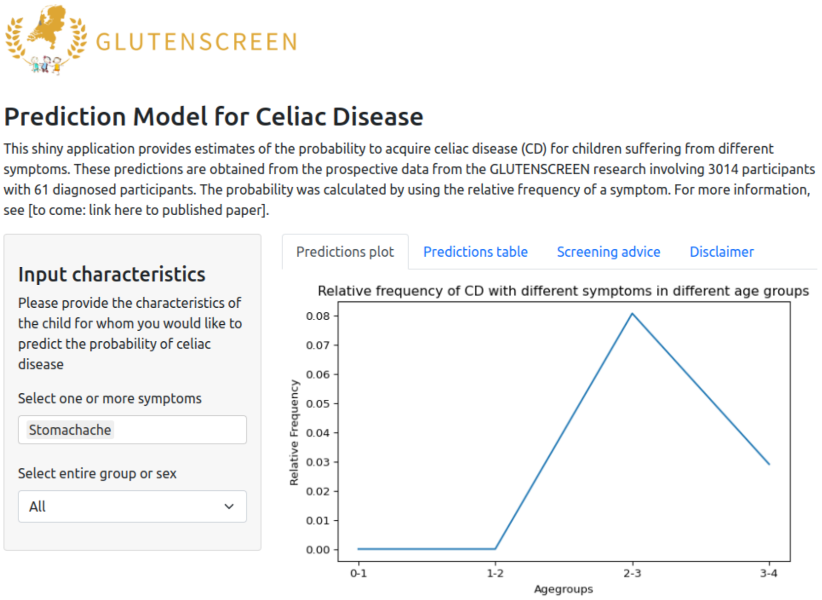


Figure 12: Heatmap of relative frequency in percent of correlation of the different symptoms (Boys)

## 3.1 Web-Application

In order to make the analyses more accessible to celiac disease researchers, a web application was created. This web-application displaying the prediction model consists of two columns, one being the input component where the user can choose one or more symptoms and choose between male, female or both together. The other is the output component that displays the generated plot of the relative frequency of the chosen symptoms in the celiac group over the four age groups. Focusing on either the different symptoms or combinations of them, or the differences between the entire celiac group, the girls and the boys, the plot gives an estimate for the risk of having celiac disease given a number of symptoms. Additionally, a prediction table showing the frequency and relative frequency of the symptoms is generated as well as a screening advice according to those outcomes. Additionally, some information about how to use this tool and what the outcomes mean was added. Figure 13 shows the default settings, when opening the webpage. In this case the prediction for having abdominal pain is shown, divided into the age groups.



Prediction Model for Celiac Disease

This shiny application provides estimates of the probability to acquire celiac disease (CD) for children suffering from different symptoms. These predictions are obtained from the prospective data from the GLUTENSCREEN research involving 3014 participants with 61 diagnosed participants. The probability was calculated by using the relative frequency of a symptom. For more information, see [to come: link here to published paper].

Input characteristics

Please provide the characteristics of the child for whom you would like to predict the probability of celiac disease

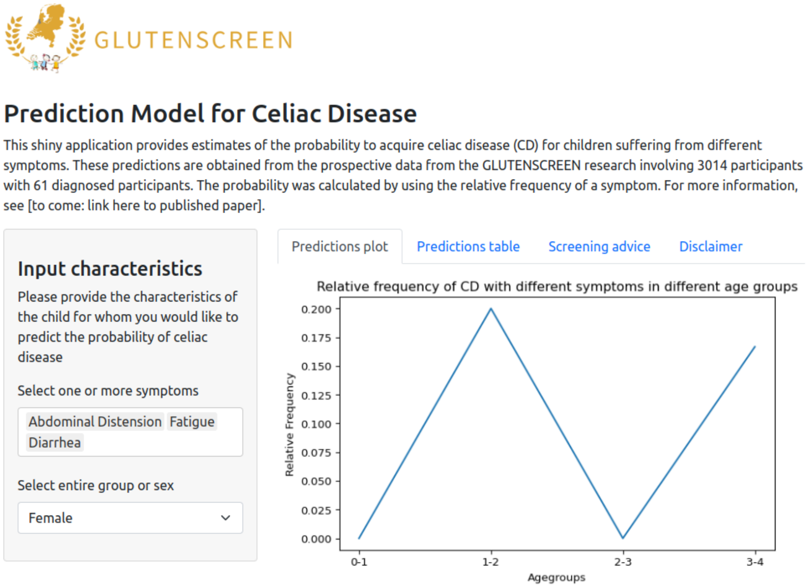
Select one or more symptoms: Stomachache

Select entire group or sex: All

Figure 13: Prediction plot for abdominal pain and all participants

By selecting different symptoms and a gender, as seen in figure 14, the predictions can be updated and an estimate of the risk of having celiac disease can be read off of the prediction plot. Here a screening advice was added to give an indication as to whether the child presumably has celiac disease or maybe some other disease.

For better readability of the relative frequency, there is a tab on the web app for the prediction table (figure 15). That prediction table also includes the occurrences in the celiac group and the symptomatic group, because the risk is only an estimate and not representative enough as there was not enough data.



GLUTENSCREEN

## Prediction Model for Celiac Disease

This shiny application provides estimates of the probability to acquire celiac disease (CD) for children suffering from different symptoms. These predictions are obtained from the prospective data from the GLUTENSCREEN research involving 3014 participants with 61 diagnosed participants. The probability was calculated by using the relative frequency of a symptom. For more information, see [to come: link here to published paper].

### Input characteristics

Please provide the characteristics of the child for whom you would like to predict the probability of celiac disease

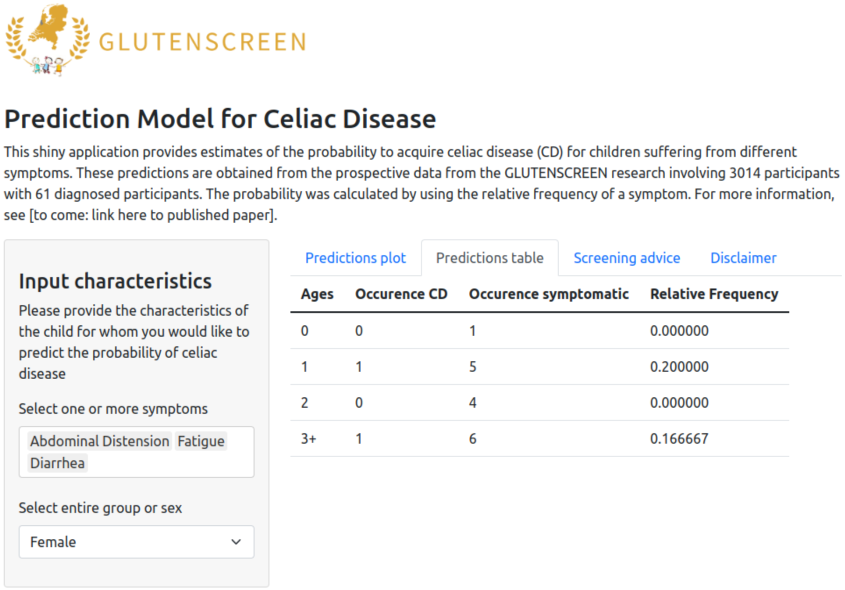
Select one or more symptoms

Abdominal Distension Fatigue Diarrhea

Select entire group or sex

Female

Figure 14: Prediction plot for girls with abdominal distension, fatigue, diarrhea



GLUTENSCREEN

## Prediction Model for Celiac Disease

This shiny application provides estimates of the probability to acquire celiac disease (CD) for children suffering from different symptoms. These predictions are obtained from the prospective data from the GLUTENSCREEN research involving 3014 participants with 61 diagnosed participants. The probability was calculated by using the relative frequency of a symptom. For more information, see [to come: link here to published paper].

### Input characteristics

Please provide the characteristics of the child for whom you would like to predict the probability of celiac disease

Select one or more symptoms

Abdominal Distension Fatigue Diarrhea

Select entire group or sex

Female

| Ages | Occurence CD | Occurence symptomatic | Relative Frequency |
|---|---|---|---|
| 0 | 0 | 1 | 0.000000 |
| 1 | 1 | 5 | 0.200000 |
| 2 | 0 | 4 | 0.000000 |
| 3+ | 1 | 6 | 0.166667 |

Figure 15: Prediction table for girls with abdominal distension, fatigue, diarrhea

# 4 Discussion and Conclusion

## 4.1 Medical

The data analysis has shown that there seems to be a clear correlation between some specific symptoms and celiac disease. The symptoms that appear very closely related to celiac disease are abdominal distension, fatigue, diarrhea and abdominal pain. Although growth restriction cannot be counted as a decisive factor in the detection of celiac disease, it is very common in patients with celiac disease and should always be regarded as an additional factor. Furthermore, the calculated relative frequencies, which give an estimate of how great the chances are of a patient having celiac disease when having a symptom, are giving new insights into which symptoms are more important to look at. Especially when looking at the differences in the age groups and the combinations of symptoms.

The analysis of the data split into age groups clearly shows that from the age of 2 up to 3 more symptoms are expressed and celiac disease is best detected. A possible explanation for this could be that at this age the participants are able to explain how they feel. While a few symptoms (fatigue, abdominal distension and diarrhea) are starting to show a bit earlier, a clear expression of all symptoms starts only from the age of 2. This differs a bit between boys and girls, which could be because girls already have a much higher chance of developing the disease, but there are also clear differences in which symptoms they have. Girls in general start expressing symptoms from the age of 1, although there is still a high peak in symptom expression at a later age, while there is close to no data of boys having symptoms before reaching 2 years. Aside from the higher prevalence for girls, girls also might develop faster than boys, meaning they can possibly communicate earlier which could result in the earlier expression of symptoms. Other important differences are the symptoms both groups express, with the most common symptom being diarrhea and moodiness or irritation for girls, and abdominal distension and aphthous for boys. These differences suggest that using a different questionnaire for boys and girls could be useful.

The combinations of symptoms as plotted in a heatmap show that abdominal distension and fatigue are very CD-specific symptoms as there are many combinations with these that have a relatively high frequency of 5% to 10% compared to the frequency of other combinations between 2% and 4%. Especially fatigue in combination with abdominal distension, constipation and diarrhea and abdominal distension in combination with diarrhea and aphthous are those combinations with a higher frequency. An interesting high-risk combination is diarrhea and constipation which usually does not occur together that often. This could either be very specific for celiac disease or be rooted in a misinterpretation of the parents

of the different symptoms. To prevent this confusion and because both symptoms occur fairly often with people with celiac disease, it could be a good idea to combine the two symptoms into into bowel irritation or abnormal feces.
After comparing the two symptom combination heatmaps for boys and girls there are a few more notable differences. As said above the more specific symptoms differ, but this applies also to the symptom combinations. For the girls bowel irritation has again the highest frequency, of around 16 %. The other important symptoms are diarrhea, fatigue and abdominal constipation where the combinations have a relative frequency of around 10%. The combination of these three symptoms even has a relative frequency of 36%. For boys the two symptoms with the highest frequency in combinations are aphthous and constipation. Although they do not occur together, combinations with those symptoms have a relatively high frequency of around 10%.
While analyzing the dataset a few problems were encountered. The biggest issue was that there just is not enough data yet, as there were only 58 patients diagnosed with celiac disease. This is not enough data to make very precise statements about the expressed symptoms and their correlation with celiac disease, although the results can be read as an estimate. Additionally, to there only being a restricted number of participants and symptomatic participants, there were also some parents of symptomatic patients that did not authorize a point-of-care-test, resulting in some symptomatic patients that could not be used for this research. From the symptoms that were used in the questionnaire from GLUTENSCREEN, some symptoms were probably too general, especially growth, as more than 50% of the symptomatic group had growth restriction. For future usage of this questionnaire, it would be better to expand the boundaries for regular growth so that small deviations from the normal growth curves won't be immediately counted as growth restriction. As for the rest of the symptoms, there are many other diseases with similar symptoms. These are mostly other autoimmune disorders like Irritable Bowel Syndrome, Crohn's disease, Type 1 Diabetes or Autoimmune Hepatitis[KO17][Fou], making it difficult to identify celiac disease specific symptoms from this limited number of symptoms.

The goal of this thesis was to investigate the correlation between celiac disease and different symptoms as well as to conclude whether there are significant differences in the development of celiac disease among the different age groups and the two genders. From the available data these questions, with reservations, have been answered. There is a correlation between specific symptoms and celiac disease and although there are other diseases with similar symptoms, as named above a few specific symptoms have been found. But there is also one symptom that did not occur at all with celiac patients, namely vomiting. This symptom could be either excluded from the questionnaire or be used as a way discarding the diagnosis for celiac disease. Another interesting result is that although there is some data of a few months old children, celiac disease can not be presumed yet, as the symptoms are either not yet distinct enough or gluten has not yet been introduced to their diet. So they could also be excluded from the research. Lastly many differences between boys and girls have been found, leading to the assumption that the early diagnosis would benefit from having different questionnaires for boys and girls.

## 4.2 Technical

There were a few technical problems encountered during this research. First of all the data pre-processing proved to be more difficult than expected, as the data was probably entered manually. This lead to the results of the diagnosis being in a different row than the answers to the questionnaire. The goal of the web application was to make an utilizable tool to support the GLUTENSCREEN project by giving an advice according to the presence of different symptoms. Here, to few data was again a problem, but with new data the application could be further expanded and be more representative. For this research Python was used for both the data analysis and the web application, as it is very versatile and often used in data analysis. Because of that it was easy to use Python as well for the web application, as the calculations are the same and could be reused, although Shiny for Python is still relatively new, which made the implementation challenging.

# 5 Further Research

For further research, it would be very interesting to look at other diseases with a similar symptomatic pattern as celiac disease and also test for other diseases to both help early diagnosis of more diseases and to help identify which symptoms are mostly correlated to celiac disease and not other possible diseases. As explained in the introduction there are many genetic influences in the development of celiac disease, which makes heredity an interesting addition to the questionnaire. During the GLUTENSCREEN study, the parents of the participants were asked whether there are family members with celiac disease or similar diseases, but there was not enough data to use in this research. Additionally to the POCT, GLUTENSCREEN could check whether the patients have the necessary HLA genes, so whether the patients are even genetically able to develop the disease before testing for it. Lastly, it would be great to follow up on this research with more data as the screening of young children is still an important tool for early detection of celiac disease and more data could help with discovering more about the disease and its symptoms.

# References

[BG13] M. N. Banu and B. Gomathy. Disease predicting system using data mining techniques. *International Journal of Technical Research and Applications*, 1.5:41–45, 2013.

[C$^+$06] A Carlsson et al. Prevalence of celiac disease: before and after a national change in feeding recommendations. *Scandinavian journal of gastroenterology*, 41.5:553–558, 2006.

[Cam21] M. Camilleri. Diagnosis and treatment of irritable bowel syndrome: a review. *Jama*, 325(9):865–877, 2021.

[CMvB$^+$99] C. G. Csizmadia, M. L. Mearin, B. M. von Blomberg, Brand R., and S. P. Verloove-Vanhorick. An iceberg of childhood coeliac disease in the netherlands. *The Lancet*, 353:813–814, 1999.

[CVS$^+$19] G. Caio, U. Volta, A. Sapone, D. A. Leffler, R. De Giorgio, C. Catassi, and Fasano A. Celiac disease: a comprehensive current review. *BMC Medicine*, 17:142, 2019.

[Fou] Celiac Disease Foundation. Autoimmune disorders. `https://celiac.org/about-celiac-disease/related-conditions/autoimmune-disorders/`. Accessed on 30-09-2022.

[GC07] P. H. Green and C. Cellier. Celiac disease. *The New England Journal of Medicine*, 357:1731–1743, 2007.

[HEAR$^+$10] C.E. Hogen Esch, R. Auricchio, J. Romanos, A. Chmielewska, H. Putter, A. Ivarsson, H. Szajewska, F. Koning, C. Wijmenga, R. Troncone, M.L. Mearin, and PreventCD Study Group. The preventcd study design: towards new strategies for the prevention of coeliac disease. *Eur J Gastroenterol Hepatol*, 22(12):1424–30, 2010.

[KO17] A. K. Kamboj and A. S. Oxentenko. Clinical and histologic mimickers of celiac disease. *Clinical and translational gastroenterology*, 8:e114, 2017.

[Lea] Scikit Learn. sklearn.cluster.kmeans. `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`. Accessed on 16-10-2022.

[M$^+$21a] C. R. Meijer et al. Early diagnosis of coeliac disease in the preventive youth health care centres in the netherlands: study protocol of a case finding study (glutenscreen). *BMJ Paediatrics Open*, 5.1, 2021.

[M$^+$21b] C. R. Meijer et al. Efficient implementation of the 'non-biopsy approach'for the diagnosis of childhood celiac disease in the netherlands: a national prospective evaluation 2010–2013. *European journal of pediatrics*, 180.8:2485–2492, 2021.

[PIO02] L. Å. Persson, A Ivarsson, and Hernell O. Breast-feeding protects against celiac disease in childhood—epidemiological evidence. *Integrating Population Outcomes, Biological Mechanisms and Research Methods in the Study of Human Milk and Lactation*, page 115–123, 2002.

[QPW$^+$14] P. Quitadamo, A. Papadopoulou, T. Wenzl, V. Urbonas, C. M. F. Kneepkens, E. Roman, R. Orel, D. J. Pavkov, J. A. Dias, Y. Vandenplas, A. Kostovski, E. Miele, A. Villani, and A. Staiano. European pediatricians' approach to children with ger symptoms: survey of the implementation of 2009 naspghan-espghan guidelines. *Journal of pediatric gastroenterology and nutrition*, 58.4:505–509, 2014.

[RDLD$^+$21] P. Riznik, L. De Leo, J. Dolinsek, J. Gyimesi, B. Klemenak, M.and Koletzko, and J. Dolinsek. Clinical presentation in children with coeliac disease in central europe. *Journal of Pediatric Gastroenterology and Nutrition*, 72(4):546–551, 2021.

[ZHAS$^+$21] R. Zannat, S. Hossain, S. Al Sakib, S. Akter, and K. T. Tahera. Disease prediction through syndromes by clustering algorithm. *American Journal of Education and Information Technology*, 5.2:93–96, 2021.