Symptom-based ordinal scale fuzzy clustering of functional gastrointestinal disorders

Marjan Mansoruiyan\textsuperscript{1,2}, Hamid Reza Marateb\textsuperscript{3,4,5,*}, Ammar Hassanzadeh Keshteli\textsuperscript{6}, Hamed Daghagh Zadeh\textsuperscript{7}, Miquel Angel Mañanas\textsuperscript{8,9}, Peyman Adibi\textsuperscript{10}

\textsuperscript{1}PhD, Department of Biostatistics and Epidemiology, School of Public Health, Isfahan University of Medical Sciences, Hezar Jerib St., 81745 Isfahan, Iran; email: j_mansourian@hlth.mui.ac.ir

\textsuperscript{2}Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain; email: marjan.mansourian@upc.edu

\textsuperscript{3}PhD, Biomedical Engineering Department, Faculty of Engineering, University of Isfahan, Hezar Jerib St., 81744 Isfahan, Iran; email: h.marateb@eng.ui.ac.ir

\textsuperscript{4}Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain; email: hamid.reza.marateb@upc.edu

\textsuperscript{5}Department of Machine learning, Isfahan Cardiovascular Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran.

\textsuperscript{6}PhD, Department of Medicine, University of Alberta, Edmonton, T6G 2E1 Canada; email: ahassanz@ualberta.ca

\textsuperscript{7}Professor of Medicine, Gastroenterology section, Department of Internal Medicine, Faculty of Medicine, Isfahan University of Medical Sciences, Hezar Jerib St., 81745 Isfahan, Iran; email: daghaghzadeh@mui.ac.ir

\textsuperscript{8}PhD, Department of Automatic Control, Biomedical Engineering Research Center, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain; email: miguel.angel.mananas@upc.edu

\textsuperscript{9}Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain.

\textsuperscript{10}Professor of Medicine, Gastroenterology section, Department of Internal Medicine, Faculty of Medicine, Isfahan University of Medical Sciences, Hezar Jerib St., 81745 Isfahan, Iran; email: adibi@med.mui.ac.ir

*Address for correspondence:
Hamid Reza Marateb
Biomedical Engineering Research Center (CREB)
Building H, Floor 4, Av. Diagonal 647, 08028 Barcelona, Spain
Phone: +34 658085138
Email: hamid.reza.marateb@upc.edu
https://bsoart.upc.edu/en/staff/
Abstract

Background

The validity of Rome III criteria for diagnosing functional gastrointestinal disorders (FGIDs) have been frequently questioned in the literature. In epidemiology, when a disease is diagnosed, the existence of a true cluster must be proven. Thus, clustering the common GI symptoms of individuals and comparing the clusters with FGIDs defined by the Rome III criteria could provide insights about the validity of FGIDs defined by those criteria.

Well-separated compact clusters were detected in responses to questionnaires of the epidemiological features of different FGIDs in Iranian adults using fuzzy ordinal clustering. The representative sample from each cluster i.e. Cluster Representative (CR) was formed whose corresponding FGID was diagnosed with Rome III criteria. Then, FGID diagnosis was performed for all participants in each cluster and the percentage of cases whose FGID was the same as the cluster’s identified FGID (agreement) was reported.

Results

Fourteen valid clusters were detected in 4763 people. The average membership of the objects in each cluster was 77.3%, indicating similarity of the objects in clusters to their corresponding CRs. Eight clusters were assigned to single FGIDs (irritable bowel syndromes: constipation IBS-C, diarrhea IBS-D and un-subtyped IBS-U; functional bloating FB; functional constipation FC; belching disorder BD. The agreement was higher than 50% in single FGID clusters except those whose diagnosis was IBS-U.

Conclusions

IBS-C, IBS-D, FC, BD, and FB defined with Rome III criteria exist in the population, which is not the case for IBS-U.
Key words: Cluster analysis; Data mining; Gastrointestinal diseases; Rome III criteria; Validation studies;
Introduction

Throughout documented history, structural diseases could be identified by pathologists and at times cured by medical technology. However, the nonstructural symptoms described as “functional” remain enigmatic and less amenable for explanation or effective treatment (1).

The research consistency and more reliable diagnosis of FGIDs have been attained utilizing the Rome criteria in conjunction with self-reported questionnaires (1, 2). The Rome classification system is based on symptom clusters assumed to remain consistent through population groups. These criteria have been modified occasionally, and have undergone many revisions since their first presentation (3).

A number of studies have been performed in the literature on the validity of symptom-based FGID diagnosis (4-7). Eslick et al performed the factor analysis of 46 symptom items of 897 patients to identify the underlying latent factors. A $k$-means cluster analysis was then applied on the extracted factors to detect the number of patient groups (4). Alessandro Digesu et al examined whether Rome III criteria are reliable to diagnose constipation in 201 women. They found out that the Rome III questionnaire did not have significant construct validity with patients’ self-report of constipation, stool frequency and stool form. They finally concluded that Rome III is not a valid instrument to diagnose constipation (6). Whitehead and Drossman performed a critical review on the validity of symptom-based criteria for diagnosing irritable bowel syndrome (IBS). They concluded that the sensitivity and specificity of those criteria are 40%-90% and 70%-90%, respectively with clinical diagnoses made by experienced clinicians (7). Dang et al performed a systematic review of symptom-based diagnostic criteria for IBS. They considered different validation studies. The patient
sample size was less than 350 in any studies considered. Based on the manuscripts reviewed by the authors until 2019, they demonstrated poor validity and utilization of Rome III for diagnosing IBS(5).

Symptom-based diagnosis of distinct FGIDs has been criticized on epidemiology due to the overlap between symptoms(4). Traditional clustering methods such as k-means or hierarchical algorithms do not allow the overlap between symptoms or FGIDs. This problem has been solved in data mining by using fuzzy clustering(8). Using advanced Fuzzy data-mining techniques on large sample sizes, could give new insights on the validity of symptom-based FGID diagnosis.

This study aimed to identify whether FGIDs could be identified based on common gastrointestinal symptoms of individuals who formed compact and well-separated epidemiological groups (clusters) in a population.

Methods

The Study on the Epidemiology of Psychological, Alimentary Health and Nutrition (SEPAHAN) project was designed to be a two-phase cross-sectional study. All information was gathered through self-administered questionnaires mainly aimed to investigate the epidemiological features of different FGIDs within a large sample of Iranian adults working in 50 different healthcare centers across Isfahan province, Iran (9). The 4-item rating ordinal-measurement scale including “never or rarely”, “sometimes”, “often”, and “always” labeled as 1-4, were used in our analysis.

Overall, detailed information from 4763 individuals became available for further analysis. The subjects did not have any alarm symptoms. All subjects gave informed consent to the experimental procedure. Ethical approval to conduct the study was provided by the
Medical Research Ethics Committee of IUMS (#189069, #189082, and #189086), conformed to the Declaration of Helsinki.

The main research question in our study was as follows: Do FGIDs diagnosed using Rome III criteria have fuzzy ontology? Also, two research hypotheses were taken into account in our study. 1) It is possible to accurately determine subject groups based on their responses to GI questionnaire; and 2) The groups formed by the clustering method could be assigned to one/more FGIDs, based on the Rome III criteria. Thus, an unsupervised machine learning approach was developed for the accurate identification of different groups (clusters) of FGIDs according to SEPAHAN questionnaire items. The particle swarm optimization-based unsupervised-fuzzy clustering technique was used as the optimum ordinal-to-interval measurement scale conversion.

Data were analyzed using Matlab and Statistics Toolbox Release 2012a (The MathWorks, Inc., Natick, Massachusetts, USA). Quantitative data were expressed as mean (SD) and qualitative as number (percent). Determining different FGID clusters according to SEPAHAN questionnaire using the clustering analysis, 31 significant head questions were selected by two experts in gastroenterology.

First, those whose response to all of the 31 selected head questions was either “never or rarely” or ”sometimes” were regarded as “healthy” and excluded from the clustering analysis. Thus, the following clustering methodology is related to unhealthy subjects. The matrix (“IN”) was then formed for the unhealthy subjects, in which the labels of the answered questions for each person were regarded as a row.

People having similar symptoms could belong to distinct diagnostic groups (FGID categories). Fuzzy clustering could be enrolled in which each subject belongs to all of the FGID groups (clusters) with different membership degrees in [0, 1]. This gives the flexibility
to express that objects can belong to more than one cluster. In fact, the terms ‘cluster’ and ‘cluster membership’ are identical with the epidemiological terms ‘group’ and ‘group membership’ (4). Also, when a disease is diagnosed in epidemiology, the existence of a true disease cluster must be proven(10). The clustering algorithm is briefly illustrated as Algorithm 1 in the appendices. The clustering method used in our study has been presented and then validated on ordinal datasets in the literature and showed promising performance by Brouwer and Groenwold (11) and Marateb et al (12).

A good partition produces a small value for the compactness, while well-separated clusters produce a high value for the separation. Thus, the best partition is related to the minimum of the XB CVI defined as the ratio of compactness to separation (13). After detecting the best clustering structure, similar clusters were merged using single-linkage clustering (using the cluster representatives as the input samples)(14). Each subject was then classified to the cluster for which its membership value was the highest. Like many epidemiologic studies in which a sample object is required from a target population of interest(15), we defined the cluster representative of each group as the representative sample (object) from that cluster. The representative sample was chosen from the objects of corresponding cluster, where the sum of the distance of the other objects in that cluster to the representative is minimal. The representative sample was, in fact, the vector of representative symptoms. The symptoms were then compared with those of ROME III by an expert in Gastroenterology and the corresponding FGIDs were identified.

Results

The socio-demographic and anthropometric factors of the sample population are listed in Table 1. The percentage of healthy people, as defined in the statistical methods section, was 51.15%. They were excluded from the descriptive and clustering analysis.
Table 1: General characteristics of study participants across different categories of variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>categories</th>
<th>Descriptive statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td>Male</td>
<td>2106 (%44.2)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2657 (%55.8)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td>Single</td>
<td>3776 (%81.2)</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>795 (%17.1)</td>
</tr>
<tr>
<td></td>
<td>Divorced or widowed</td>
<td>79 (1.7%)</td>
</tr>
<tr>
<td><strong>Educational level</strong></td>
<td>Under diploma</td>
<td>638 (13.4%)</td>
</tr>
<tr>
<td></td>
<td>Diploma</td>
<td>1348 (28.3%)</td>
</tr>
<tr>
<td></td>
<td>Above diploma but under master</td>
<td>2338 (49.1%)</td>
</tr>
<tr>
<td></td>
<td>Master or above</td>
<td>312 (6.6%)</td>
</tr>
<tr>
<td><strong>Age (year)</strong></td>
<td>-</td>
<td>36.6±3.1</td>
</tr>
<tr>
<td><strong>BMI (kg/m²)</strong></td>
<td>-</td>
<td>24.9±5.1</td>
</tr>
<tr>
<td><strong>Weight (kg)</strong></td>
<td>-</td>
<td>68.89±13.49</td>
</tr>
</tbody>
</table>

All values are mean ± SD for quantitative variables and number (percent for qualitative variables; BMI: Body mass index;

The average membership of the objects in each cluster is shown in Table 2. The membership value for each object was calculated in terms of the relative similarity (16) between the 31 ordinal-scale questionnaire answers of that subject and those of the corresponding cluster representative. It ranged from 0% (complete dissimilarity) to 100% (complete similarity or clustering consistency).

The percentage of people (i.e., the number of people divided by 4763) belonging to identified clusters is also listed in Table 2. The matched FGIDs identified using the major symptoms of the subjects in the identified clusters are also shown in Table 2.
**Table 2:** The identified functional gastrointestinal disorders subtypes and the percentage of people in each cluster, and the average membership of the objects in each cluster identified in the SEPAHAN dataset shown in mean ± SD.

<table>
<thead>
<tr>
<th>Cluster Index</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
<th>C11</th>
<th>C12</th>
<th>C13</th>
<th>C14</th>
</tr>
</thead>
<tbody>
<tr>
<td>% People a</td>
<td>1.46</td>
<td>2.58</td>
<td>1.41</td>
<td>0.61</td>
<td>2.48</td>
<td>0.98</td>
<td>6.12</td>
<td>8.80</td>
<td>2.08</td>
<td>1.14</td>
<td>8.45</td>
<td>8.11</td>
<td>4.41</td>
<td>0.22</td>
</tr>
<tr>
<td>Average member</td>
<td>83±5</td>
<td>82±6</td>
<td>72±6</td>
<td>56±8</td>
<td>79±5</td>
<td>77±5</td>
<td>83±8</td>
<td>82±10</td>
<td>77±6</td>
<td>73±6</td>
<td>82±10</td>
<td>83±7</td>
<td>78±8</td>
<td>75±5</td>
</tr>
<tr>
<td>FGIDs c</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
<td>IBS-C</td>
</tr>
<tr>
<td></td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td></td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
<td>PDS</td>
</tr>
</tbody>
</table>

a: The total percentage of people is 48.9%. They are the “unhealthy” people in the analyzed samples. The rest, i.e. “healthy” people were excluded from the analysis as illustrated in the “statistical methods” section. The percentage of people was calculated as calculated as the number of people in each cluster divided by the total sample size of 4763. b: The membership value for each object was calculated in terms of the relative similarity between 31 the questionnaire answers of that subject and those of the corresponding cluster representative. c: Identification of common FGIDs was done by two gastroenterologists. FGIDs: functional gastrointestinal disorders, IBS: irritable bowel syndrome, IBS-C: constipation predominant IBS, IBS-D: diarrhea predominant IBS, IBS-M: mixed IBS, IBS-U: unsubtyped IBS, FD: functional dyspepsia, FB: functional bloating, FC: functional constipation, PDS: postprandial distress syndrome, UH: uninvestigated heartburn, BD: belching disorder.

Clusters C2, C6, C7, C8, C9, C11, C12, and C13 were assigned to specific FGIDs based on the similarity of the cluster representative with Rome III criteria. The agreement between their symptoms and those of assigned FGIDs were shown as Venn diagrams in Figure 1. In this figure, symptom-based FGID diagnosis was performed for all participants in
each cluster based on the Rome III criteria, and the percentage of cases whose FGID was the same as the cluster’s identified FGID was shown. The agreement was higher than the cut-off similarity level of 50% (17) in all cases except those whose diagnosis was IBS-U (C2, and C12). C2 and C12 were both assigned to IBS-U, although C12 had minor variations with C2. For example, stoppage, enema and defecation problems existed in C12 but not C2. In C2 cluster, people often had bloating which was not the case in C12. C7 and C8 were both assigned to functional bloating. Early satiety, borborygmi and bloating were more frequent in C8 while dyspepsia, belching, and abdominal/non-cardiac chest pain and fullness were more frequent in C7. There were insufficient criteria for the diagnosis of functional dyspepsia, IBS or other FGIDs. They could be considered as functional abdominal bloating based on Rome I and II criteria, although the name was changed to functional bloating in Rome III criteria (18).

Figure 1 about here

Cluster C3 was the combination of IBS-C (52%) and PDS (96%), while clusters C14 included IBS-M (43%) and uninvestigated heartburn (85%), where the agreement percentage was shown in the parenthesis. Cluster C1 was the combination of IBS-C (7%) and FD (non-PDS) (34%). On the other hand, cluster C5 was the combination of IBS-C (1%) and FD (PDS) (74%). Clusters C4 and C10 had the combination IBS-C (24%, 38%), and functional dyspepsia (PDS) (100%, 85%). However, the symptoms of C10 were not as severe as those of C4. For example, anal sore, bloating, mucus, defecation problem, enema, incomplete evacuation, hardness, and abdominal pain were more frequent in C4 in comparison with C10. In the meanwhile, uninvestigated heartburn (21%) was also present in C4. Two identified
clusters C4 and C10, are examples of the concept of severity in IBS introduced by Drossman et al.(19).

Discussion

The clustering algorithms have been widely used in clinical studies to identify subtypes of diseases or human disease etiological factors; this is a well-known methodology(20). In epidemiology, when a disease is first identified, its occurrence as a true cluster (group) must be proven(10). The validity of the identified clusters was assessed based on 1) XB CVI that ensures the clusters are compact and well-separated, and 2) the average membership (similarity) of the objects in each cluster (Table 2; above 77% in single FGID clusters). The cluster representative was selected similar to the epidemiological studies as the sample representative of that cluster(15). The existence of the FGIDs with the minimum similarity level of 50%(17), could be proven in our study. However, since the overlap (similarity) of the single-FGID clusters with Rome-III defined FGIDs are between 0 %-100 %, their degree of existence could be stated in the framework of Fuzzy ontology(21). Thus, the main research question was answered. As a validation of this hypothesis, in total, 21.07% of the population had IBS (Table 2). A recent patient-based analysis of the same database showed that the prevalence of IBS was 21.7% using Rome III diagnostic criteria (22).

Our results concerning that IBS-C, and D could be symptom-based diagnosed are in agreement with those of Whitehead et al (23) and Whitehead and Drossman (24). However, this is in contrast with some studies that claimed poor validity and utilization of Rome III criteria (4, 5, 25). Since the clustering was performed on all of the 31 significant head questions, high agreement with FGIDs diagnosed by few items in Rome III is very hard to reach. Thus, like many microbiological studies(17), the cut-off similarity value of 50% was used in our study. However, tightening the cut-off similarity level to about 70%, which is
suitable for distinguishing unambiguous epidemiological disease, the ontology of the
following FGIDs FC, FB and IBS-C could be verified(17).

In our study, each subject was finally assigned to a cluster. This is one of the limitations
of our study in which concurrent FGIDs could not be entirely identified. Alternatively, the
membership function could be used to identify different clusters to which a subject belongs
that will be focused in the future.

Abbreviations

BD: belching disorder; CR: Cluster Representative; FGIDs: functional gastrointestinal
disorders; FB: functional bloating; FC: functional constipation; IBS: irritable bowel
syndromes; IBS: irritable bowel syndrome, IBS-C: constipation predominant IBS, IBS-D:
diarrhea predominant IBS, IBS-U: unsubtype IBS; PDS: postprandial distress syndrome,
SEPAHAN: Study on the Epidemiology of Psychological, Alimentary Health and Nutrition;
UH: uninvestigated heartburn;

Ethics Approval

Written informed consents were obtained from all participants, and the ethics committee of
Isfahan University of medical sciences approved the study.

Competing Interest

The authors declare that they have no competing interests.

Funding

The research leading to these results has received funding from the European Union’s
Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant
agreement No 712949 (TECNIOspring PLUS) and from the Agency for Business
Competitiveness of the Government of Catalonia and also by the Spanish Ministry of
Economy and Competitiveness- Spain project DPI2017-83989-R.

Authors Contribution
M.M., HR.M., P.A, and A.H.K designed research; M.M. and HR.M. analyzed data; M.A.M
and M.M. wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
The authors are grateful to Kevin McGill for reviewing a draft of this paper. The
authors are thankful to Sobhan Goudarzi for collaboration in implementing part of the
clustering algorithm.

References
4. Eslick GD, Howell SC, Hammer J, Talley NJ. Empirically derived symptom sub-
groups correspond poorly with diagnostic criteria for functional dyspepsia and irritable bowel
syndrome. A factor and cluster analysis of a patient sample. Alimentary pharmacology &
diagnostic criteria for IBS demonstrates poor validity and utilization of Rome III.
Neurogastroenterology and motility : the official journal of the European Gastrointestinal


Algorithm 1: Identifying the clustering structure of the input matrix data “IN”

For $C = 2$ to $\sqrt{4763}$; where $C$ is the possible number of clusters

% Identify the clustering structure assuming $C$ clusters as follows:

% Assign randomly membership degrees to each subject;

% Initialize OIC matrix “F” using the input data “IN”;

% Repeat until the algorithm has converged i.e. membership degrees do not change in two consecutive iterations (max. iterations: =10);

% Estimate “F” matrix so as to minimize the FCM objective function using PSO;

% Update the membership degrees and clusters’ prototype;

END Repeat;

% Calculate the ratio between the compactness of the fuzzy partition and the strength of the separation between clusters (XB Clustering-Validity Index (CVI));

END For;

% Select the optimum number of clusters ($C_{optm}$) based on the minimum value of the XB CVI and save the corresponding clustering structure (“F” matrix, subject’s membership degrees and clusters’ prototypes);
**Figure 1:** The overlap between identified clusters and corresponding functional gastrointestinal disorders (FGIDs) (Based on the Rome III criteria) shown as a Venn diagram. The percentage of the cases whose symptom-based diagnosis was identical to those of their cluster representatives were reported for clusters assigned to a single FGID. IBS: irritable bowel syndrome, IBS-C: constipation predominant IBS, IBS-D: diarrhea predominant IBS, IBS-U: unsubtyped IBS, FB: functional bloating, FC: functional constipation, PDS: postprandial distress syndrome, UH: uninvestigated heartburn, BD: belching disorder.