Journal of Medical Systems (2019) 43: 52
https://doi.org/10.1007/s10916-019-1172-1

TRANSACTIONAL PROCESSING SYSTEMS

Cluster Analysis of Obesity Disease Based on Comorbidities Extracted from Clinical Notes

Ruth Reátegui[1,2] · Sylvie Ratté[1] · Estefanía Bautista-Valarezo[2] · Víctor Duque[2]

Received: 14 June 2018 / Accepted: 16 January 2019 / Published online: 28 January 2019

Abstract

Clinical notes provide a comprehensive and overall impression of the patient's health. However, the automatic extraction of information within these notes is challenging due to their narrative style. In this context, our goal was to identify clusters of patients based on fourteen comorbidities related to obesity, automatically extracted with the cTAKES tool from the i2b2 Obesity Challenge data. Furthermore, results were compared with clusters obtained from experts' annotated data. The sparse K-means algorithms were used in both experiment at two levels: at the first level, three clusters were found, and at the second, new clusters were found by applying the same algorithm to each of the clusters from the former level. The results show that three types of clusters could be identified based on the number of comorbidities and the percentage of patients suffering from them. Diabetes, hypercholesterolemia, atherosclerotic cardiovascular diseases, congestive heart failure, obstructive sleep apnea, and depression were the diseases with the highest weights contributing to the cluster distribution.

Keywords Obesity · Clinical notes · cTAKES · Cluster analysis

Introduction

Patients' information, including diseases, symptoms, treatments, drugs, etc., can be derived from clinical notes such as discharge summaries. These documents have a narrative format, which allows the health professional to write in a flexible manner. The notes contain local dialectal phrases, negations, acronyms, abbreviations, misspellings and typing errors, which combined, make it difficult to automatically extract patients' information from them [1, 2].

Manual extraction of patient information is carried out by experts, and is laborious and time-consuming. Even automatic extraction is extremely difficult because the information sought is hidden within significant amounts of data residing in clinical notes [2]. The process of getting structured medical information requires extracting named entities or concepts and then mapping them to codes according to controlled vocabulary or medical standards [1]. Two standards used to map biomedical concepts are the Unified Medical Language System (UMLS) [3] and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [4]. Clinical features, such as comorbidities (simultaneous presence of two diseases or conditions in a patient) related to a specific disease, are important features and are at the root of other tasks such as cluster analysis.

In the medical field, cluster analysis helps in identifying and tailoring treatment or care delivery, defining boundaries and disease taxonomies, understanding the heterogeneity of the disease, identifying subsets of patients with similar characteristics, identifying relevant pathophysiologies, etc. [5–8]. Many authors have applied cluster analysis to various conditions, such as obstructive sleep apnea [9–11], asthma [6, 12],

This article is part of the Topical Collection on *Transactional Processing Systems*

✉ Víctor Duque
victorduquetf@gmail.com

Ruth Reátegui
rmreategui@utpl.edu.ec

Sylvie Ratté
Sylvie.Ratte@etsmtls.ca

Estefanía Bautista-Valarezo
mebautista@utpl.edu.ec

[1] École de Technologie Supérieure, 1100 Notre-Dame Street West, Montreal, Quebec H3C 1K3, Canada

[2] Universidad Técnica Particular de Loja, San Cayetano Alto, 1101608 Loja, Ecuador

knee osteoarthritis [8], chronic heart failure [13], and chronic obstructive pulmonary diseases [5].

Overweight and obesity are a global health problem that is becoming an epidemic in both children and adults [14]. Obesity is often accompanied by other health risks or comorbidities such as diabetes, dyslipidemia, hypertension, cardiovascular diseases, asthma and osteoarthritis. [15–17].

Cluster analysis studies that take into account diseases related to obesity include that by Sutherland et al. [18], which identified clusters of patients suffering from obesity and asthma simultaneously. Laing et al. [19] analyzed the relationship between obesity and atherosclerosis, while LaGrotte et al. [20] focused on patients with obstructive sleep apnea, obesity, and excessive daytime sleepiness.

Notwithstanding the presence of multiple comorbidities in obesity patients, most of the related works focus on analyzing the relationship between 2 comorbidities instead of 14. Furthermore, all the features in the above works were collected manually or from structured EHR data despite the significant amount of information inside clinical notes. The motivation for this work was therefore to apply cluster analysis to obesity comorbidities in order to gain insights into the different types of obesity patients that can exist according to the number of comorbidities they have.

Based on the above explanation, the goal of our work is to identify clusters of obesity patients based on obesity comorbidities extracted from clinical notes automatically. In addition, a cluster analysis based on the comorbidities annotated by experts from the same dataset was developed in order to allow a comparison with the cluster analysis result from the extracted data. The i2b2 (Informatics for Integrating Biology to the Bedside) Obesity Challenge data was used.

Materials and methods

In this section, we will describe the dataset used and the process for expert annotation and automatic extraction. We will also explain the cluster analysis.

Dataset

We used the i2b2 2008 Obesity dataset. This dataset consists of 1237 discharge summaries of overweight and diabetic patients [21]. The documents in the dataset contain expert annotations that classify 15 obesity comorbidities as present, absent, questionable or unmentioned [21]. Table 1, column 1, shows the 14 comorbidities (known as *diseases* hereinafter) used. Hypertriglyceridemia does not have sufficient samples, and was therefore excluded. Out of 1237 summaries, 412 summaries which had obesity and at least one of the 14 diseases were selected. The last preselection was made to avoid samples with 0 in all the columns, because in the dataset obtained with the automatic extraction, there were some cases where the patients showed obesity without another comorbidity. Also, in this work, we wanted to keep the same patients that were selected in our previous work [22].

Experts' annotation and automatic entity extraction

In our previous work [22], the cTAKES and MetaMap tools were compared in the extraction process of 14 obesity comorbidities. Also, experts' textual annotations were used as a gold standard. The comorbidities were treated as dichotomy variables or features (values of 0 or 1, respectively depicting the non-existence or existence of the disease in discharge summaries). The results showed cTAKES slightly outperforming MetaMap. In this work, therefore, we decided to use the results obtained with cTAKES, together with the expert annotations to identify clusters of obesity patients. These results also consider an aggregation process and some semantics types. Table 1 shows the expert annotations and the cTAKES results. The average for recall, precision and F-score are 0.91, 0.89, and 0.89, respectively. More details of the extraction and aggregation processes are given in [22].

Cluster analysis

Sparse K-means clustering developed by [23] was chosen to conduct two experiments: a cluster analysis using the automatic extracted data and a cluster analysis using the experts' annotated data. The sparse K-means has the advantage of allowing an accurate identification of the groups and providing interpretable results following the identification of the most relevant clustering features [23]. This algorithm assigns a weight to each disease (used as features by the algorithm), with the diseases that contribute the most to a cluster having the highest values. Before using sparse K-means, we applied gap statistics [24] to estimate the number of clusters. Cluster analysis was performed at two levels in both experiments: in the data extracted and in the annotated data. At the first level, sparse K-means was applied to all 412 patients, resulting in three clusters. At the second level, the same algorithm was applied to each of the three clusters of the first level. The second level gave us a total of 11 clusters. Fig. 1 shows their distribution and the cluster equivalence between both experiments.

Results

In this section, we will detail the results of the cluster analysis. Fig. 1 shows the distribution of the clusters of both experiments, along with the correspondence between the clusters according to the highest percentage of patients suffering from a disease.

Table 1 Diseases annotated by experts and extracted with MetaMap

Diseases	Experts' annotation		cTAKES Extraction		Evaluation		
	NP (*)	%(*)	NP	%	Recall	Precision	F-score
Hypertension	325	79	340	83	0.99	0.95	0.97
Diabetes	259	63	266	65	0.92	0.89	0.91
CAD	181	44	205	50	0.92	0.81	0.87
CHF	172	42	183	44	0.92	0.86	0.89
HCL	172	42	146	35	0.81	0.96	0.88
OSA	127	31	102	25	0.76	0.94	0.84
OA	87	21	61	15	0.67	0.95	0.78
Depression	83	20	116	28	0.99	0.71	0.82
Asthma	81	20	92	22	1.00	0.88	0.94
GERD	76	18	85	20	0.99	0.88	0.93
CCY	74	18	68	17	0.89	0.97	0.93
Gout	56	14	58	14	0.98	0.95	0.96
PVD	37	9	32	8	0.84	0.97	0.90
VI	21	5	30	7	1	0.7	0.824
Evaluation Results Average:					**0.91**	**0.89**	**0.89**

Reategui R, Ratte S (2018) Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC medical informatics and decision making 18 (Suppl 3):74. doi:https://doi.org/10.1186/s12911-018-0654-2

(*) Number and percentage of patients with the disease

CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia; OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy; PVD: peripheral vascular disease; VI: venous insufficiency

Cluster analysis with extracted data

The first level had 3 clusters (See Table 2), and the diseases with the highest weights were hypercholesterolemia and diabetes. These clusters had the following characteristics: EC1 had 159 patients, all of them had diabetes, and 79% had hypertension. A moderate percentage had CAD (52%) and CHF (43%). There were no patients with hypercholesterolemia. EC2 had 107 patients, 80% had hypertension, and a moderate percentage had CHF (38%). There were no patients with diabetes and hypercholesterolemia. EC3 had 146 patients, all of them had hypercholesterolemia, 88% had hypertension, 73% had diabetes, 66% had CAD, and a moderate percentage had CHF (51%). In these big groups, it is easy identify obese patients with hypertension, diabetes and without hypercholesterolemia (EC1), obese patients with hypertension and without diabetes and hypercholesterolemia (EC2), and obese patients with hypertension, diabetes, CAD, and hypercholesterolemia (EC3). Other diseases are present in moderate and low rates.

At the second level from EC1, 3 clusters were obtained: EC1.1, EC1.2, and EC1.3. The diseases with the highest weights were CAD and depression. From EC2, 4 clusters were obtained: EC2.1, EC2.2, EC2.3 and EC2.4. The diseases with the highest weights were CHF and OSA. From EC3, 4 clusters were obtained: EC3.1, EC3.2, EC3.3 and EC3.4. The diseases with the highest weights were CAD and CHF. Table 3 shows the results at the second level.

Cluster analysis with annotated data

A cluster analysis using the experts' annotated data was carried out to compare the results obtained with the automatic extracted data. As in the above experiment, the first level had 3 clusters, AC1, AC2, and AC3. The diseases with the highest weights were hypercholesterolemia and diabetes. These new clusters compared with the clusters from the extracted data (at the first level) show a small difference in the percentage of the patients suffering from a specific disease. See Table 2 for more details.

At the second level from AC1, 3 clusters were obtained: AC1.1, AC1.2, and AC1.3. The diseases with the highest weights were CAD and OSA. From AC2, 4 clusters were obtained: AC2.1, AC2.2, AC2.3 and AC2.4. The diseases with the highest weights were hypertension and OSA. From AC3, 4 clusters were obtained: AC3.1, AC3.2, AC3.3 and AC3.4. The diseases with the highest weights were CAD and CHF. The difference between these new clusters and those obtained with the extracted data lies in the fact that with the new ones, we have a new distribution of clusters AC1 based on OSA diseases and of clusters AC2 based on hypertension diseases. Table 4 shows the results at the second level.

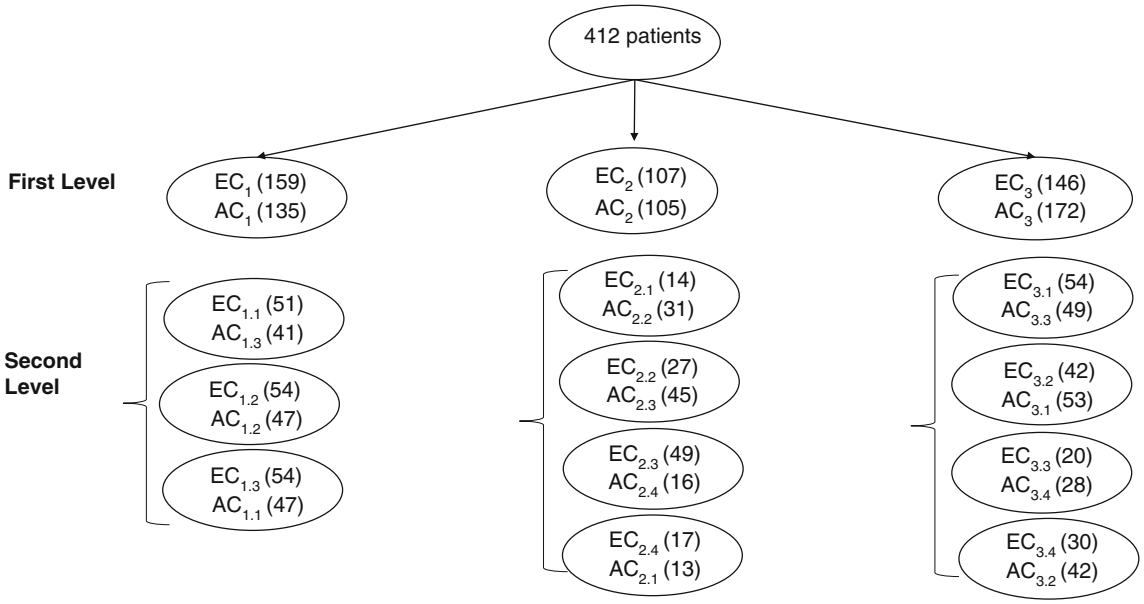

Fig. 1 Cluster analysis by levels. The numbers in parentheses are the patients in each cluster. EC cluster are from the extracted data and AC clusters from annotated data

Cluster classification

Based on an observation of the clusters, and considering the number of diseases (comorbidities) and the percentage of patients with them, three types of comorbidities were identified:

- High comorbidity: Occurs when a cluster has one of the following characteristics: (1) Three or more diseases with a high percentage of patients (67% to 100%); (2) Two diseases with a high percentage of patients (67% to 100%) and one or more diseases, with 33% to 66% of the patients suffering from them.

Table 2 Clusters from the first level

DISEASES	CLUSTERS WITH EXTRACTED DATA					CLUSTERS WITH ANNOTATED DATA				
	NUM PATIENTS	DW	EC1 (HC)	EC2 (MC)	EC3 (HC)	NUM PATIENTS	DW	AC1 (HC)	AC2 (MC)	AC3 (HC)
Hypertension	340	0,00	**79**	**80**	**88**	325	0,00	**73**	**72**	**87**
Diabetes	266	0,23	**100**	0	**73**	259	0,23	**100**	0	**72**
CAD	205	0,00	52	25	**66**	181	0,00	45	17	59
CHF	183	0,00	43	38	51	172	0,00	44	34	45
HCL	146	0,97	0	0	**100**	172	0,97	0	0	**100**
OSA	102	0,00	24	29	23	127	0,00	30	42	24
OA	61	0,00	9	20	17	87	0,00	11	30	23
Depression	116	0,00	32	24	27	83	0,00	28	20	14
Asthma	92	0,00	18	30	22	81	0,00	17	27	17
GERD	85	0,00	17	20	25	76	0,00	14	16	23
CCY	68	0,00	13	21	17	74	0,00	14	22	19
Gout	58	0,00	15	14	13	56	0,00	15	14	12
PVD	32	0,00	11	2	9	37	0,00	13	6	8
VI	30	0,00	6	7	8	21	0,00	3	5	7
Patients per Cluster			159	107	146			135	105	172

The cluster results are represented in percentages according to the number of patients in each group

The Disease Weights (DW) show the weights that sparse K-means assigns to each disease

Bold numbers are the highest values

HC: High Comorbidity; MC: Medium Comorbidity

CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia; OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy; PVD: peripheral vascular disease; VI: venous insufficiency

Table 3 Clusters from the second level with extracted data

DISEASES	DW	EC1.1 (HC)	EC1.2 (HC)	EC1.3 (MC)	DW	EC2.1 (HC)	EC2.2 (HC)	EC2.3 (LC)	EC2.4 (HC)	DW	EC3.1 (HC)	EC3.2 (HC)	EC3.3 (HC)	EC3.4 (HC)
Hypertension	0,01	**86**	**72**	**78**	0,00	**93**	**93**	**73**	**71**	0,00	**80**	**93**	**90**	**97**
Diabetes	0,00	**100**	**100**	**100**	0,00	0	0	0	0	0,00	**76**	**69**	**75**	**73**
CAD	**0,61**	55	**100**	0	0,00	29	37	22	12	**0,46**	**100**	**100**	0	0
CHF	0,05	37	59	31	**0,97**	**100**	**100**	0	0	**0,89**	**100**	0	**100**	0
HCL	0,00	0	0	0	0,00	0	0	0	0	0,00	**100**	**100**	**100**	**100**
OSA	0,01	31	20	20	**0,23**	**100**	0	0	**100**	0,00	26	17	30	20
OA	0,01	8	15	6	0,00	21	19	27	0	0,00	13	17	30	17
Depression	**0,79**	**100**	0	0	0,00	36	33	10	41	0,00	20	43	20	20
Asthma	0,00	18	19	17	0,00	43	33	27	24	0,00	24	21	20	20
GERD	0,02	27	11	13	0,00	29	15	22	12	0,00	26	19	35	27
CCY	0,00	18	9	13	0,00	0	41	14	24	0,00	17	21	10	17
Gout	0,01	22	13	11	0,00	29	19	10	6	0,00	22	5	10	10
PVD	0,01	4	19	9	0,00	0	4	2	0	0,00	11	12	5	3
VI	0,00	2	7	9	0,00	21	7	2	12	0,00	9	2	15	10
Patients per Cluster		51	54	54		14	27	49	17		54	42	20	30

The cluster results are represented in percentages according to the number of patients in each group

The Disease Weights (DW) show the weights that sparse K-means assigns to each disease

Bold numbers are the highest values

HC: High Comorbidity; MC: Medium Comorbidity; LC: Low Comorbidity

CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia; OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy; PVD: peripheral vascular disease; VI: venous insufficiency

- Medium comorbidity: Occurs when a cluster has one of the following characteristics: (1) Two diseases with a high percentage of patients (67% to 100%); (2) One disease with a high percentage of patients (67% to 100%) and one or more diseases, with 33% to 66% of patients suffering from them.
- Low comorbidity: Occurs when a cluster has one of the following characteristics: (1) One disease with a high percentage of patients (67% to 100%); (2) One or more diseases in the 33% to 66% range or between 0% to 32%

Given the above explanations, and considering the clusters obtained with the extracted data, at the first level, EC1 and EC3 have a high comorbidity, and EC2 has a medium comorbidity.

At the second level, EC1.1 and EC1.2 have a high comorbidity, and EC1.3 has a medium comorbidity; EC2.1 has a high comorbidity, EC2.2 and EC2.4 have high comorbidity, and EC2.3 has a low comorbidity, and EC3.1, EC3.2, EC3.3, and EC3.4 have a high comorbidity.

Considering the clusters obtained with the annotated data, at the first level, AC1 and AC3 have a high comorbidity, and AC2 has a medium comorbidity. At the second level, AC1.2 and AC1.3 have a high comorbidity, and AC1.1 has a medium comorbidity; AC2.2 has a high comorbidity, AC2.1 and AC2.3 have medium comorbidity, and AC2.4 has a low comorbidity; AC3.1, AC3.2, AC3.3, and AC3.4 have a high comorbidity.

Discussion

To get a medical interpretation of the clusters obtained with the extracted data, two physicians were asked to express their opinion about Table 2 and Table 3. They provided the comments below. Fig. 2 shows the diseases with a high percentage of patients in each sub-cluster.

Considering the percentage of patients with a comorbidity, hypertension, diabetes, CAD, CHF, HCL and OSA have the highest values. These results agree with previous works, such as [14, 15, 25], which show the common comorbidities related to obesity and overweight people.

In cluster EC1, all patients have diabetes. Also, all EC3 incorporate more than 70% of diabetic patients. Although the experiment does not detail the type of diabetes, we could mention the high association of type 2 diabetes with obesity. More than 80% of cases of type 2 diabetes can be attributed to obesity that causes insulin resistance, and ultimately, hyperglycemia [26]. Although metformin (one of the drugs most commonly used in the treatment of type 2 diabetes) causes weight reduction, some drugs used in type 2 diabetes can cause moderate increases in weight [27].

Table 4 Clusters from the second level with annotated data

DISEASES	DW	AC1.1 (MC)	AC1.2 (HC)	AC1.3 (HC)	DW	AC2.1 (MC)	AC2.2 (HC)	AC2.3 (MC)	AC2.4 (LC)	DW	AC3.1 (HC)	AC3.2 (HC)	AC3.3 (HC)	AC3.4 (HC)
Hypertension	0,04	64	**70**	**88**	**0,23**	0	**100**	**100**	0	0,00	**96**	**90**	**78**	**82**
Diabetes	0,00	**100**	**100**	**100**	0,00	0	0	0	0	0,00	60	64	**90**	**75**
CAD	**0,64**	0	**100**	34	0,00	8	6	22	31	**0,23**	**100**	0	**100**	0
CHF	0,05	32	60	39	0,00	54	26	44	6	**0,97**	0	0	**100**	**100**
HCL	0,00	0	0	0	0,00	0	0	0	0	0,00	**100**	**100**	**100**	**100**
OSA	**0,76**	0	0	100	**0,97**	**100**	**100**	0	0	0,00	19	14	33	36
OA	0,01	6	11	17	0,00	54	26	24	38	0,00	21	26	18	32
Depression	0,02	26	21	39	0,00	31	35	11	6	0,00	11	12	14	21
Asthma	0,03	11	11	32	0,00	15	32	24	31	0,00	6	24	20	25
GERD	0,04	15	2	27	0,00	31	16	13	13	0,00	17	31	24	21
CCY	0,01	19	11	12	0,00	15	13	29	25	0,00	19	19	27	4
Gout	0,00	11	17	17	0,00	23	19	13	0	0,00	6	10	22	11
PVD	0,05	2	28	7	0,00	0	0	7	19	0,00	8	7	10	7
VI	0,00	2	2	5	0,00	15	6	2	0	0,00	2	10	6	14
Patients per Cluster		47	47	41		13	31	45	16		53	42	49	28

The cluster results are represented in percentages according to the number of patients in each group

The Disease Weights (DW) show the weights that sparse K-means assigns to each disease

Bold numbers are the highest values

HC: High Comorbidity; MC: Medium Comorbidity; LC: Low Comorbidity

CAD: atherosclerotic cardiovascular diseases; CHF: congestive heart failure; HCL: hypercholesterolemia; OSA: obstructive sleep apnea; OA: osteoarthritis; GERD: gastroesophageal reflux disease; CCY: cholecystectomy; PVD: peripheral vascular disease; VI: venous insufficiency

In cluster EC2, all the patients do not have diabetes and HCL. This cluster has the highest percentage (29%) of patients with OSA. OSA is a disorder that occurs during sleep, in which the patient experiences repetitive episodes of apnea (stops breathing) or a reduction of airflow due to an obstruction of the upper airway. Obesity is the most potent risk factor in the development of OSA, and its relative risk increases as the body mass index (BMI) increases [28]. In EC2.1 and EC2.4, all the patients have OSA with comorbidities such as hypertension and CHF. These diseases, among others, were identified as comorbidities in OSA patients in the work of [9]. The prevalence of cardiovascular risk factors such as hypertension and type 2 diabetes is substantially higher in patients with OSA. OSA is a clear risk factor for cardiovascular events, although it is not entirely clear whether it is due to its high association with obesity and other coexisting factors [29]. This could explain why not all patients with OSA are present in this cluster, while 24% and 23% of OSA patients are in EC1 and EC3, which have patients with other OSA comorbidities such as diabetes and CAD. Also, EC2 has the highest percentage of hypertensive patients (80%); this can be explained by the fact that hypertension is independently associated with OSA.

In EC3, we find 100% of patients with HCL. Furthermore, this cluster has the highest percentage of patients with hypertension and CAD. The risk of CAD is higher in obese people. Most experts attribute part of this relationship to the coexistence of risk factors, although the American Heart Association considers obesity as an independent risk factor for CAD [14]. In the same cluster, for example, 88% of patients have hypertension, 73% have diabetes, and 100% have HCL. These are precisely the main risk factors for CAD that usually exist in patients with obesity. Previous works [30, 31] showed that diabetes, hypertension, hypercholesterolemia, etc., are related to CAD patients, and these diseases are present in our clusters as well. It is important to note that the sub-clusters EC3.1, EC3.2, and EC1.2 include patients a high percentage of whom have three or more comorbidities. These clusters have a high percentage of patients (over 50%) with at least two of the mentioned risk factors. This reinforces the theory that the relationship between obesity and CAD is mainly due to the coexistence of many acting morbidities as risk factors for cardiovascular events.

This study also allows us to see how a pathology behaves in the different clusters. With respect to CHF and obesity patients, [13] found 4 clusters with phenotypes related to these diseases. Some important comorbidities they identified are hypertension, diabetes, hyperlipidemia, etc. Our clusters (EC2.1, EC2.3, EC3.1) show the presence of the same comorbidities as well.

In EC2.3, no patient has diabetes, CHF, HCL, OSA, and a low percentage have CAD and other diseases. This cluster

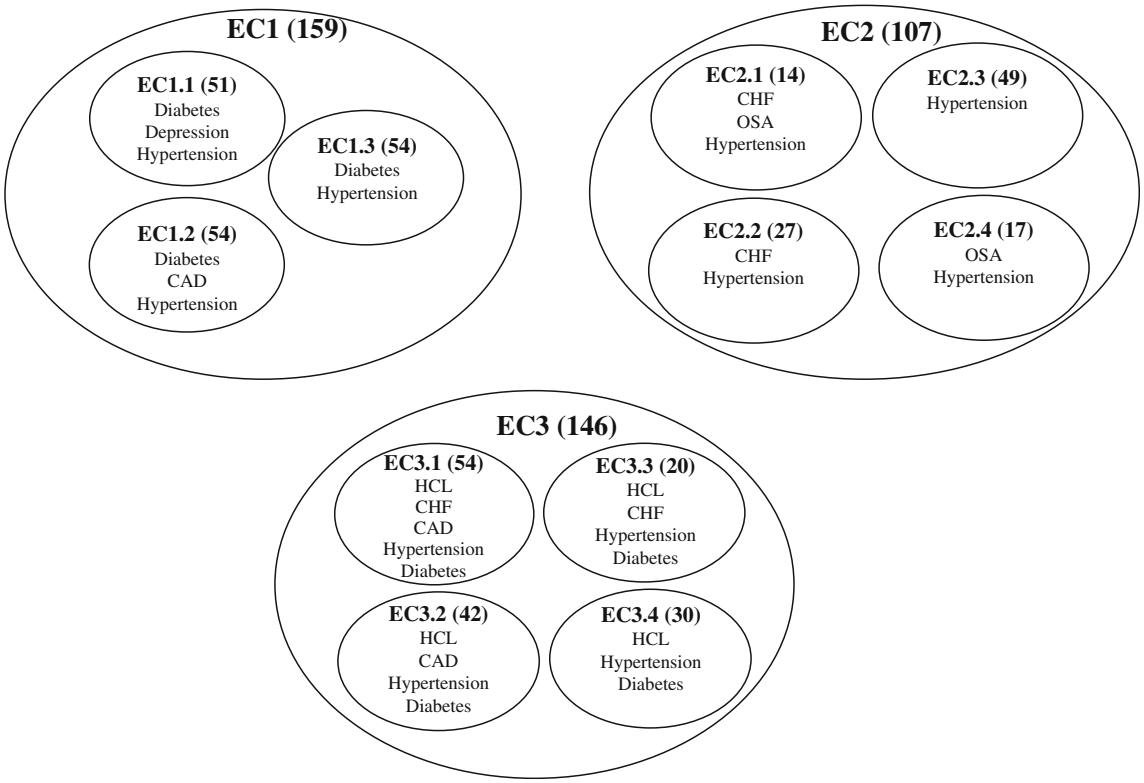

Fig. 2 Diseases with a high percentage (67 to 100%) of patients in each sub-cluster

give us an idea of the multifactorial and multimorbid character generally associated with obesity. We can say this the healthiest group with the lowest mortality risk diseases. In medicine, the term "metabolically healthy obese" is used to refer to obese patients who do not have cardio-metabolic abnormalities associated with adipocytes. Studies have shown that these patients have an increased risk of mortality compared to normal weight and metabolically healthy individuals [32]. As well, the patients in that cluster could have lived with a risk factor for a short time period (e.g., few overweight years); they could also be patients with low BMI. One hypothesis that could be further explored is that this group seems to contain patients who have been suffering from obesity for a few years or whose BMI is not too high (overweight or low-grade obesity). The present study does not have sufficient data to test such a hypothesis.

Another case to analyze is depression. There are diverse opinions respecting the association of obesity with depression. For example, [33] suggests that depression is associated with severe obesity, especially among young women that have a poor body image. A prospective study by [34] indicates that obesity increases the risk of depression, but that the inverse is not true, that is, depression does not appear to increase the risk of gaining weight. On the other hand, a meta-analysis of 15 prospective observational studies published in 2010 similarly shows that the risk of developing depression among obese patients and among depressed patients is the same; in other words, having either one of these pathologies predisposes a patient to the other [35]. This may account for 20% (83 patients in the annotated data) of the 412 patients considered in the present work having depression. In cluster EC1.1, all the patients have depression, while in EC2.1, EC2.2, EC2.4, and EC3.2, only a moderate percentage of them presents this disease.

Asthma is present among a moderate percentage of patients in EC2.1, EC2.2. The relationship between depression and asthma was addressed in a meta-analysis of 8 prospective studies published by [36]. This meta-analysis establishes that the risk of developing asthma increases by 43% among patients who have depression as compared to those who do not, although asthma does not appear to increase the risk of depression.

Conclusion and future work

Considering 14 obesity comorbidities, clustering analysis at 2 levels was applied. The first level provides a general idea of the prevalent diseases afflicting obesity patients, as well as the type of comorbidity (HC and MC) they have. At the second level, groups of patients were identified, with more details provided about their comorbidities. Most of the clusters present a high comorbidity with common diseases mentioned by experts in the literature.

Furthermore, despite the differences in the weights assigned to diseases in the second level, the extracted and annotated data present some equivalence in the clusters found in both experiments. This shows that the automatic extraction of medical entities and cluster analysis allow to discover groups of patients with similar characteristics. These clusters help doctors to gain insights into the variety of patient phenotyping characterizing a disease such as obesity.

The present work has some limitations that should be covered in future studies. For example, 14 obesity comorbidities based on the Obesity Challenge promoted by i2b2 were used, but other diseases present in the discharge summaries could be considered. Moreover, this work did not distinguish between diabetes types, as mentioned above; knowing which patients have type 2 DM can help physicians confirm the relationship between obesity and this disease.

Compliance with Ethical Standards

Conflicts of Interest The authors declare they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Bukhanov, N., Balakhontceva, M., Krikunov, A., Sabirov, A., Semakova, A., Zvartau, N., and Konradi, A., Clustering of comorbidities based on conditional probabilities of diseases in hypertensive patients. *Proc. Comput. Sci.* 108:2478–2487, 2017. https://doi.org/10.1016/j.procs.2017.05.073.
2. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M., A review of approaches to identifying patient phenotype cohorts using electronic health records. *JAMIA* 21(2):221–230, 2014. https://doi.org/10.1136/amiajnl-2013-001935.
3. National Library of Medicine (US), UMLS® Reference Manual, 2009. http://www.ncbi.nlm.nih.gov/books/NBK9676/. Accessed 20 Mar 2018.
4. National Library of Medicine (US), Overview of SNOMED CT, 2016. https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html. Accessed 20 Mar 2018.
5. Chen, C.-Z., Wang, L.-Y., Ou, C.-Y., Lee, C.-H., Lin, C.-C., and Hsiue, T.-R., Using cluster analysis to identify phenotypes and validation of mortality in men with COPD. *Lung* 192(6):889–896, 2014. https://doi.org/10.1007/s00408-014-9646-x.
6. Bourdin, A., Molinari, N., Vachier, I., Varrin, M., Marin, G., Gamez, A.-S., Paganin, F., and Chanez, P., Prognostic value of cluster analysis of severe asthma phenotypes. *J. Allerg. Clin. Immunol.* 134(5):1043–1050, 2014. https://doi.org/10.1016/j.jaci.2014.04.038.
7. Rocha, A., and Rocha, B., Adopting nursing health record standards. *Inform. Health Soc. Care* 39(1):1–14, 2014. https://doi.org/10.3109/17538157.2013.827200.
8. van der Esch, M., Knoop, J., van der Leeden, M., Roorda, L. D., Lems, W. F., Knol, D. L., and Dekker, J., Clinical phenotypes in patients with knee osteoarthritis: A study in the Amsterdam osteoarthritis cohort. *Osteoarthr. Cartil.* 23(4):544–549, 2015. https://doi.org/10.1016/j.joca.2015.01.006.
9. Vavougios, G. D., Natsios, G., Pastaka, C., Zarogiannis, S. G., and Gourgoulianis, K. I., Phenotypes of comorbidity in OSAS patients: Combining categorical principal component analysis with cluster analysis. *J. Sleep Res.* 25(1):31–38, 2016. https://doi.org/10.1111/jsr.12344.
10. Joosten, S. A., Hamza, K., Sands, S., Turton, A., Berger, P., and Hamilton, G., Phenotypes of patients with mild to moderate obstructive sleep apnoea as confirmed by cluster analysis. *Respirology* 17(1):99–107, 2012. https://doi.org/10.1111/j.1440-1843.2011.02037.x.
11. Figueroa, R. L., and Flores, C. A., Extracting information from electronic medical records to identify the obesity status of a patient based on comorbidities and bodyweight measures. *J. Med. Syst.* 40(8):1–9, 2016.
12. Serrano-Pariente, J., Rodrigo, G., Fiz, J. A., Crespo, A., Plaza, V., and High Risk Asthma Res G, Identification and characterization of near-fatal asthma phenotypes by cluster analysis. *Allergy* 70(9):1139–1147, 2015. https://doi.org/10.1111/all.12654.
13. Ahmad, T., Pencina, M. J., Schulte, P. J., O'Brien, E., Whellan, D. J., Pina, I. L., Kitzman, D. W., Lee, K. L., O'Connor, C. M., and Felker, G. M., Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J. Am. Coll. Cardiol.* 64(17):1765–1774, 2014. https://doi.org/10.1016/j.jacc.2014.07.979.
14. Poirier, P., Giles, T. D., Bray, G. A., Hong, Y., Stern, J. S., Pi-Sunyer, F. X., and Eckel, R. H., Obesity and cardiovascular disease: Pathophysiology, evaluation, and effect of weight loss. *Arterioscler. Thromb. Vasc. Biol.* 26(5):968–976, 2006. https://doi.org/10.1161/01.ATV.0000216787.85457.f3.
15. Guh, D. P., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L., and Anis, A. H., The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. *BMC Pub. Health* 9:1–20, 2009. https://doi.org/10.1186/1471-2458-9-88.
16. Foster, M. C., Hwang, S. J., Larson, M. G., Lichtman, J. H., Parikh, N. I., Vasan, R. S., Levy, D., and Fox, C. S., Overweight, obesity, and the development of stage 3 CKD: The Framingham heart study. *Am. J. Kidney Dis. : Off. J. Natl. Kidney Found* 52(1):39–48, 2008. https://doi.org/10.1053/j.ajkd.2008.03.003.
17. Sutherland, E. R., Goleva, E., King, T. S., Lehman, E., Stevens, A. D., Jackson, L. P., Stream, A. R., Fahy, J. V., Leung, D. Y. M., and Asthma Clin Res, N., Cluster analysis of obesity and Asthma phenotypes. *Plos One* 7(5):1–7, 2012. https://doi.org/10.1371/journal.pone.0036631.
18. Laing, S. T., Smulevitz, B., Vatcheva, K. P., Rahbar, M. H., Reininger, B., McPherson, D. D., McCormick, J. B., and Fisher-Hoch, S. P., Subclinical atherosclerosis and obesity phenotypes among Mexican Americans. *J. Am. Heart Assoc.* 4(3):e001540, 2015. https://doi.org/10.1161/jaha.114.001540.
19. LaGrotte, C., Fernandez-Mendoza, J., Calhoun, S. L., Liao, D., Bixler, E. O., and Vgontzas, A. N.., The relative association of obstructive sleep apnea, obesity, and excessive daytime sleepiness with incident depression: A longitudinal, population-based study. *Int. J. Obes.*:1–8, 2016. doi:https://doi.org/10.1038/ijo.2016.87.
20. Uzuner, Ö., Recognizing obesity and comorbidities in sparse data. *JAMIA* 16(4):561–570, 2009.
21. Reategui, R., and Ratte, S., Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med. Inform. Dec. Mak.* 18(Suppl 3):74, 2018. https://doi.org/10.1186/s12911-018-0654-2.
22. Witten, D. M., and Tibshirani, R., A framework for feature selection in clustering. *J. Am. Stat. Assoc.* 105(490):713–726, 2010. https://doi.org/10.1198/jasa.2010.tm09415.

23. Tibshirani, R., Walther, G., and Hastie, T., Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. B* 63:411–423, 2001. https://doi.org/10.1111/1467-9868.00293.
24. Bruce, S. G., Riediger, N. D., Zacharias, J. M., and Young, T. K., Obesity and obesity-related comorbidities in a Canadian first nation population. *Prevent. Chron. Dis.* 8(1):A03, 2011.
25. Willett, W. C., Dietz, W. H., and Colditz, G. A., Guidelines for healthy weight. *N. Engl. J. Med.* 341(6):427–434, 1999. https://doi.org/10.1056/NEJM199908053410607.
26. Leslie, W. S., Hankey, C. R., and Lean, M. E. J., Weight gain as an adverse effect of some commonly prescribed drugs: A systematic review. *Qjm-Int J. Med.* 100(7):395–404, 2007. https://doi.org/10.1093/qjmed/hcm044.
27. Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E., and Hla, K. M., Increased prevalence of sleep-disordered breathing in adults. *Am. J. Epidemiol.* 177(9):1006–1014, 2013. https://doi.org/10.1093/aje/kws342.
28. Wolf, J., Lewicka, J., and Narkiewicz, K., Obstructive sleep apnea: An update on mechanisms and cardiovascular consequences. *Nutr. Metab. Cardiovas.* 17(3):233–240, 2007. https://doi.org/10.1016/j.numecd.2006.12.005.
29. Canto, J. G., Kiefe, C. I., Rogers, W. J., Peterson, E. D., Frederick, P. D., French, W. J., Gibson, C. M., Pollack, C. V., Ornato, J. P., Zalenski, R. J., Penney, J., Tiefenbrunn, A. J., Greenland, P., and Investigators, N., Number of coronary heart disease risk factors and mortality in patients with first myocardial infarction. *Jama J. Am. Med. Assoc.* 306(19):2120–2127, 2011. https://doi.org/10.1001/jama.2011.1654.
30. Mamudu, H. M., Paul, T. K., Wang, L., Veeranki, S. P., Panchal, H. B., Alamian, A., Sarnosky, K., and Budoff, M., The effects of multiple coronary artery disease risk factors on subclinical atherosclerosis in a rural population in the United States. *Prevent. Med.* 88:140–146, 2016. https://doi.org/10.1016/j.ypmed.2016.04.003.
31. Kramer, C. K., Zinman, B., and Retnakaran, R., Are metabolically healthy overweight and obesity benign conditions?: A systematic review and meta-analysis. *Ann. Intern. Med.* 159(11):758–769, 2013. https://doi.org/10.7326/0003-4819-159-11-201312030-00008.
32. Dixon, J. B., Dixon, M. E., and O'Brien, P. E., Depression in association with severe obesity - Changes with weight loss. *Arch. Intern. Med.* 163(17):2058–2065, 2003. https://doi.org/10.1001/archinte.163.17.2058.
33. Roberts, R. E., Deleger, S., Strawbridge, W. J., and Kaplan, G. A., Prospective association between obesity and depression: Evidence from the Alameda County study. *Int. J. Obes.* 27(4):514–521, 2003. https://doi.org/10.1038/sj.ijo.08022204.
34. Luppino, F. S., de Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W., and Zitman, F. G., Overweight, obesity, and depression: A systematic review and meta-analysis of longitudinal studies. *Arch. Gen. Psychiat.* 67(3):220–229, 2010. https://doi.org/10.1001/archgenpsychiatry.2010.2.
35. Gao, Y. H., Zhao, H. S., Zhang, F. R., Gao, Y., Shen, P., Chen, R. C., and Zhang, G. J., The relationship between depression and Asthma: A meta-analysis of prospective studies. *Plos One* 10(7):1–12, 2015. https://doi.org/10.1371/journal.pone.0132424.