BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029594
Article Type:	Research
Date Submitted by the Author:	07-Feb-2019
Complete List of Authors:	Foguet-Boreu, Quintí; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), ; Vic University Hospital, Department of Psychiatry Violan-Fors, Concepción; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Fernández-Bertolín, Sergio; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Guisado-Clavero, Marina; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Cabrera-Bean, Margarita; Universitat Politecnica de Catalunya, Signal Theory and Comunications Department Formiga, F; Hospital Universitari de Bellvitge Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol),
Keywords:	Chronic conditions, Multimorbidity, Cluster analysis, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population

1. Concepción Violán-Fors*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

2. Quintí Foguet-Boreu*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Department of Psychiatry, Vic University Hospital. Francesc Pla el Vigatà, 1, 08500 Vic, Barcelona, Spain. 4. Department of Basic and Methodological Sciences. Faculty of Health Sciences and Welfare. University of Vic- Central University of Catalonia (UVic-UCC)
E-mail: 42292qfb@comb.cat

3. Sergio Fernández-Bertolín. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: sfernandez@idiapjgol.org

4. Marina Guisado-Clavero. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: marina.guisado@gmail.com

5. Margarita Cabrera-Bean. Signal Theory and Comunications Department, Universitat Politècnica de Catalunya, Barcelona Tech. Campus Nord, UPC D5, Jordi Girona 1-2, 08034-Barcelona, Spain.
E-mail: marga.cabrera@upc.edu

6. Francesc Formiga. Internal Medicine Service, Hospital Universitari de Bellvitge, Hospitalet del Llobregat, Barcelona, Catalonia, Spain.
E-mail: fformiga@bellvitgehospital.cat

7. Jose M Valderas. Health Services & Policy Research Group, Academic Collaboration for Primary Care, University of Exeter Medical School, Exeter, EX1 2LU, United Kingdom.
E-mail: J.M.Valderas@exeter.ac.uk

8. Albert Roso-Llorach. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

Corresponding author: Concepción Violán. IDIAPJGol
Quintí Foguet-Boreu. IDIAPJGol
Gran Via Corts Catalanes, 587 àtic.08007 Barcelona. Spain.
Telephone: 0034 93 482 41 24. FAX: 0034 93 482 41 74.
Web page: www.idiapjgol.org.E-mail: cviolan@idiapjgol.org; 42292qfb@comb.cat

Word count: 2 960

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Abstract

Objectives The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied. Fuzzy cluster analysis allows individuals to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.

Design A cross-sectional study was conducted based on data from electronic health records

Setting 284 primary health care centres in Catalonia, Spain (2012).

Participants 916 619 eligible individuals were included (women: 57.7%).

Primary and secondary outcome measures We extracted data on demographics, ICD-10 chronic diagnoses, prescribed drugs, and socioeconomic status for patients aged ≥65. Following principal component analysis of categorical and continuous variables (PCAmix) for dimensionality reduction, machine learning techniques were applied for the identification of disease clusters in a fuzzy c-means analysis. Sensitivity analyses, with different prevalence cut-off points for chronic diseases, were also conducted. Solutions were evaluated from clinical consistency and significance criteria.

Results Multimorbidity was present in 93.1%. Eight clusters were identified with a varying number of disease values: *Nervous and digestive*; *Respiratory, circulatory, and nervous*; *Circulatory, and digestive*; *Mental, nervous, and digestive*; *Mental, digestive, and blood*; *Nervous, musculoskeletal, and circulatory*; *Genitourinary, mental, and musculoskeletal*; and *Non-specified*. Nuclear diseases were identified for each cluster independently of the prevalence cut-off point considered.

Conclusions Multimorbidity patterns were obtained using fuzzy c-means cluster analysis. They are clinically meaningful clusters which support the development of tailored approaches to multimorbidity management and further research.

Keywords: Chronic conditions; Multimorbidity; Epidemiology; Cluster analysis.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Strengths and limitations of this study

- Studies focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns. Hard clustering forces each individual to belong to a single cluster, whereas soft clustering allows elements to be simultaneously classified into multiple cluster.
- Reliable and valid identification of disease clusters is needed for the development of evidence-based clinical practice guidelines and pathways of care for patients with multimorbidity.
- Soft clustering analysis allows for diseases to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.
- The different cut-off points (prevalence filters) applied to obtain multimorbidity patterns permitted the identification of common nuclear diseases which remained independent of their prevalence.
- The literature provides support for the etiopathophysiological and epidemiological associations between conditions forming part of the same cluster.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Introduction

The term multimorbidity widely refers to the existence of numerous medical conditions in a single individual (1). In many regions of the world there is evidence that a substantial, and probably growing, proportion of the adult population is affected by multiple chronic conditions. Moreover, the association of multimorbidity with increasing age leading to a two-fold prevalence in the final decades of life has been proven (2). Multimorbidity has been estimated to be at around 62% between 65 and 74 years, and around 81.5% after 85 years (3). Its true extent is, however, difficult to gauge as there is no agreed definition or classification system (4-7).

Most of the published literature focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns (8). Orienting the analysis of multimorbidity patterns at an individual level, and not of disease, could have crucial implications for patients. In the current context of limited evidence on interventions for unselected patients with multimorbidity, such an approach would allow better understanding of population groups, and facilitate the development and implementation of strategies aimed at prevention, diagnosis, treatment, and prognosis. It would also elicit essential information for the development of clinical guidelines, pathways of care, and lead to better understanding of the nature and range of the required health services (9,10).

Cluster analysis involves assigning individuals so that the items (diseases) in the same cluster are as similar as possible, while individuals belonging to different clusters are as dissimilar as possible. The identification of clusters is based on similarity measures and their choice may depend on the data or the purpose of the analysis (11,12). Hard clustering forces each element to belong to a single cluster, whereas soft clustering (also referred to as fuzzy clustering) allows elements to be simultaneously classified into multiple clusters.

Empirical evidence is needed on how both established and novel techniques influence the identification of multimorbidity patterns. A recent systematic review recommended that future epidemiological studies cover a broad selection of health conditions in order to avoid missing potentially key nosological associations and enhance external validity. When many conditions are

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

considered, the clustering of individuals based on morbidity data will encounter high-dimensional issues. This is particularly important when a clustering-based approach is adopted to assess the impact of multimorbidity on individual health outcomes and health service uses (2, 8, 13-15).

The identification of multimorbidity patterns seems to be implicitly dependent on the prevalence of the included diseases (2,8,16,17). However, to the best of our knowledge no previous study has analysed the identification of multimorbidity patterns explicitly based on the prevalence of the diseases.

The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied.

Methods

Study population

A cross-sectional analysis was carried out in Catalonia (Spain), a Mediterranean region of 7,515,398 inhabitants (2012). The Catalan Health Institute provides universal coverage and operates 284 primary health care centres (PHC).

Data sources

Since 2006 the Information System for Research in Primary Care (SIDIAP) database includes anonymized longitudinal electronic health records from primary and secondary care which gather information on demographics, diagnoses, prescriptions, and socioeconomic status (18). In our study the inclusion criteria were individuals aged 65-99 years on 31st December 2011 with at least one PHC visit since 2012. Only participants that survived until 31st December 2012 (index date) were included in the analysis.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Variables

Diseases were coded in the SIDIAP using the International Classification of Diseases version 10 (ICD-10). An operational definition of multimorbidity was the simultaneous presence of more than one of the selected 60 chronic diseases previously identified by the Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (19).

Additional variables included in the study were sociodemographics (age, sex, socio-economic status (MEDEA index) (20), clinical variables (including number of chronic diseases and invoiced drugs), and use of health services (number of visits to family physicians, nurses, and emergency services).

Patient and Public Involvement

Patients and or public were not involved in the study.

Statistical analysis

Descriptive statistics were used to summarize overall information. Disease prevalence was computed for all the included population. Descriptive analyses were stratified by the presence of multimorbidity. Comparison was performed using t-Student or Mann-Whitney for continuous variables and Chi-Square for categorical ones.

In order to obtain the most representative clusters all patients were included irrespective of whether they presented multimorbidity or not. Sex and age variables, together with chronic diseases selected by prevalence, were included in the analysis. The number of features to be considered varied from the 62 original ones (no prevalence filtering applied) to 54 and 49, for a 1% and 2% prevalence threshold, respectively.

Due to the large number of diseases, a principal component analysis for categorical and continuous data (PCAmix) was implemented to reduce complexity. With this technique both continuous and dichotomous variables were simultaneously processed through the application of Multi Correspondence Analysis to the binary variables and PCA to the continuous ones. Using

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Karlis-Saporta-Spinaki criterion to select the optimal number of dimensions to retain, the dataset of 49 features per individual per 2% prevalence cut-off was transformed to a new dimensionally reduced dataset of 13 continuous features per individual, which concentrated most of the variability of the newly transformed dataset (21).

Once the transformed dataset was obtained, clusters of chronic conditions at baseline were identified using the fuzzy c-means clustering algorithm (22). This machine learning technique forces every individual to belong to every cluster in accordance with its characteristics and by assigning a membership degree factor in (0,1) to each individual with respect to each pattern. This provides the flexibility enabling patients to belong to more than one multimorbidity pattern (23).

The main parameters in this clustering procedure were the number of clusters and a fuzziness parameter, denoted *m*, that ranged from just above 1 to infinity. High *m* values produce a fuzzy set of clusters, so that individuals are equally distributed across clusters, whereas lower ones generate non-overlapped clusters. Further details on the stability and validation techniques applied to obtain the best fuzzy c-means parameters and the set of centroids, are presented in Additional File 1.

To describe the multimorbidity patterns, frequencies and percentages of diseases (P) in each cluster were calculated. Observed/expected ratios (O/E-ratios) were calculated by dividing disease prevalence in the cluster by disease prevalence in the overall population. As the membership of each individual to any of the clusters was given by a membership degree factor, and not as a binary variable, the observed disease prevalence (O) in a cluster was computed as the sum of the disease membership degree factors corresponding to all individuals suffering the disease. Exclusivity, defined as the proportion of patients with the disease included in the cluster over the total number of patients with the disease, was also calculated. Further details on how these ratios were computed using the membership factors are given in Additional File 1. A disease was considered to be part of a multimorbidity cluster when O/E-ratio was ≥2 or exclusivity value ≥25% (24).

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

We conducted a sensitivity analysis by modifying the prevalence threshold for disease inclusion in the cluster analysis. For chronic diseases we considered as alternatives no filtering, and ≥1% and ≥2% filters among the included population. The content of each cluster was compared across filtering approaches in terms of diseases associated with that cluster, characteristics of the included population, and cluster size. Clinical evaluation of the consistency and significance of these solutions was also conducted.

The analyses were carried out using R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). The significance level was set at 0.05.

Results

In this study 916,619 individuals were included (women: 57.7%; mean age: 75.4 (standard deviation, SD: 7.4), and 853,085 (93.1%) of them met multimorbidity criteria (Figure 1).

Participants' characteristics are summarized in Table 1. Statistically significant differences were present between the multimorbidity and non-multimorbidity groups for all the variables included in the analysis (Table 1).

Among the 60 SNAC-K chronic diseases, the most prevalent were: hypertension (71.0%), dyslipidaemia (50.9%), osteoarthritis and other degenerative joint diseases (32.8%), obesity (28.7%), diabetes (25.1%), and anaemia (18.3%) (Table 2).

Eight multimorbidity patterns were identified using fuzzy c-means algorithm with fuzziness parameter of m=1.1, after computing different validation indices to obtain the optimal number of clusters (Additional File 1). This number was the same for the three different prevalence thresholds: no filtering, and ≥1% and ≥2% filters. The cluster formed by the most prevalent diseases was designated *Non-specified* (O/E ratio < 2 and exclusivity < 20). The remaining 7 clusters were specific: *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive; Mental, nervous, and digestive*; *Mental, digestive, and blood; Nervous,*

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

musculoskeletal, and circulatory; and *Genitourinary, mental, and musculoskeletal* (Table 3). Table 3 shows the results, considering a 2% prevalence filter, for each pattern based on the fifteen diseases with the higher O/E-ratios.

Women were more represented than men in almost all clusters, from 52.7% for *Respiratory, circulatory, and neurological* to 83.6% for *Mental, nervous, and digestive.* The exception was *Genitourinary, mental, and musculoskeletal* in which men made up 90.9% due to the presence of male reproductive system diseases (Table 4).

The highest O/E ratio and exclusivity value were observed in *Nervous and digestive* for Parkinson, parkinsonism, and other neurological diseases (17.0% and 74.3%; and 15.9% and 69.4%, respectively). The lowest values were found in *Non-specified*. Clusters 1 to 3 presented the highest median number of visits with *Circulatory and digestive* being associated with the greatest number of visits over a one-year period (median 18 visits), and the *Non-specified* pattern presenting the lowest median number of visits which was equal to 5 (Table 4).

Multimorbidity patterns varied according to requirements for minimal prevalence of selected conditions in the population. As an example, Figure 2 depicts the composition of Cluster 1 according to prevalence levels of disease, and the other clusters are shown in Additional file 2. Disease prevalence varied more greatly in the less populated patterns (e.g. *Non-specified*) (Additional File 2). Nevertheless, there was a group that remained in some clusters across all prevalence levels, for instance, some in *Neurological and digestive* (Parkinson and parkinsonism, other neurological diseases, chronic liver diseases, chronic pancreas, biliary tract, and gallbladder diseases) formed part of the cluster regardless of changes in cut-off prevalence (Additional File 2). The selected level of prevalence resulted in changes in O/E ratios, with some of them doubling their values.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Discussion

The soft clustering method we employed identified eight multimorbidity patterns, regardless of the prevalence selected. The *Non-specified* cluster included not only the largest number of individuals, but also those who presented the smallest multimorbidity prevalence. In this pattern diseases did not exhibit an association higher than chance because values of the O/E ratio and exclusivity were less than 2% and 20%, respectively. This suggests that such patients during their lives could change group. Two clusters presenting gender dominance were observed: *Nervous, musculoskeletal and circulatory* was predominately made up of women >70 years, while *Genitourinary, mental and musculoskeletal* was mostly formed of men of the same age. Such patterns represent 61% of the elderly participants included in the study. The rest had fewer individuals and some diseases were over-represented such as Parkinson and parkinsonism in *Nervous and digestive,* and asthma in *Respiratory, circulatory, and nervous.*

We observed that some diseases with O/E ratios ≥ 2 were consistently associated with each other as part of the same clusters (for instance, *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive*; and *Mental, nervous, and digestive*) regardless of the prevalence threshold that had been set. They can be considered core components of those clusters. Further research is needed to establish the role of these conditions from a longitudinal perspective.

Comparison with the literature

Comparison with other studies is hindered by variations in methods, data sources and structures, populations, and diseases studied. Nevertheless, there are similarities with other authors. The non-specified pattern is the one most replicated in the literature, for example Prados et al who employed an exploratory factor analysis (25) and our group with k-means (24).

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Recent research has provided support for physio-pathological and genetic associations that explain the observed multimorbidity patterns. For instance, *Neurological and digestive* included chronic liver disease which has been linked to Parkinson through the accumulation of toxic substances in the brain (ammonia and manganese) and neuroinflammation (26). A higher risk of Parkinson among patients with chronic hepatitis C virus has also been reported (OR: 1.35) (27), in addition to associations between digestive diseases and neurodegenerative ones (e.g. Parkinson and Alzheimer) through the microbiome-gut-brain axis (27). A possible link between microbiota and digestive diseases such as chronic pancreatitis and pancreatic cancer has also been suggested (28,29). For the *Respiratory, circulatory, and neurological* cluster there is evidence of an association between chronic bronchial pathology, particularly asthma and obstructive pulmonary disease (COPD), and the risk of cardiovascular events (30). Longitudinal studies have observed an increased risk of developing Parkinson among individuals suffering from asthma and/or COPD (31,32). The association between asthma and allergy is known, and its coexistence defines a specific phenotype. For the *Circulatory and digestive c*luster, non-alcoholic fatty liver disease has been associated with the development of atrial fibrillation (33), and hepatitis C infection with an increase in the risk of developing cardio- and cerebrovascular events (34). In addition, anaemia has been associated with advanced stages of chronic renal diseases and erythropoietin deficiency (35). Iron-deficiency anaemia has been associated with an increased risk of stroke (36) through thromboembolic phenomena secondary to reactive thrombocytosis. Chronic kidney disease produces auricle injuries (dilatation, fibrosis) and systemic inflammation, both of which can favour the onset and maintenance of atrial fibrillation (37).

Strengths and limitations

A major strength of this study is that it has employed a large, high-quality database made up of primary care records representative of the Catalan population aged ≥ 65 years (18). Patterns of multimorbidity have been studied based on the whole eligible sample. This approach is epidemiologically robust as the prevalence of diseases has been estimated on the whole sample rather than limited to patients with multimorbidity (2). Another strength is that individuals rather

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

than diseases have been considered as the unit of analysis (8, 24). Such an approach permits a more realistic and rational monitoring of participants than cohort studies in order to analyse multimorbidity patterns along time. Moreover, the use of different prevalence cut-offs to obtain multimorbidity patterns has allowed the identification of nuclear diseases. We selected the higher prevalence (2%) because the patterns obtained had more clinical representativeness. The inclusion of all the potential diagnoses may have signified a greater complexity that would have hindered both the interpretation of findings and comparison with other studies.

Compared to hierarchical clustering, fuzzy c-means cluster analysis is less susceptible to: outliers in the data, choice of distance measure, and the inclusion of inappropriate or irrelevant variables (38). Nevertheless, some disadvantages of the method are that different solutions for each set of seed points can occur and there is no guarantee of optimal clustering (11). To minimize this shortcoming, we carried out 100 cluster realizations with different seeds to finally use the average result of all of them. In addition, the method is not efficient when a large number of potential cluster solutions are to be considered (38). To address this limitation, we computed the optimal number of clusters using analytical indexes (Additional File 1).

Other limitations need to be taken into account. The dimensional reduction method performed in this work to reduce data complexity was PCAmix. Such methods can produce low percentages of variation on principal axes and make it difficult to choose the number of dimensions to retain. In order to decide on the most suitable number of dimensions we applied Karlis-Saporta-Spinaki rule (27) which resulted in a 13-dimensional space for the 2% prevalence cut-off.

Implications for practice, policy, and research

Soft clustering methods offer a new methodological approach to understanding the relationships between specific diseases in individuals. This is an essential step in improving the care of patients and health systems. Analysing multimorbidity patterns permits the identification of patient subgroups with different associated diseases.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

The inclusion of varying cut-off points (prevalence filters) of the diseases that form the multimorbidity patterns allowed us to identify common nuclear diseases that remained independent from the prevalence that build such patterns.

It is noteworthy that 60% of the population ≥65 years was included in multimorbidity patterns made up of the most prevalent diseases. The rest of the population was grouped into five more specific patterns which permitted their better management.

Whilst clinical guidelines are currently aimed at covering the management of the diseases found in the *Non-specified* cluster, there is a lack of information regarding the associated diseases in the other patterns. The challenge will be to refocus healthcare policy from that based on individual diseases, with the accompanying consequences (increased risk of functional decline, poorer quality of life, greater use of services, polypharmacy, and increased mortality), to a multimorbidity orientation (39).

Further investigation on this topic is called for with particular focus on four major issues. First, the genetic study of these patterns will help the identification of risk subgroups. Second, research is needed on the life style and environmental factors (diet, physical exercise, toxics) associated with such patterns. Third, longitudinal studies should be performed to establish the onset order of the core diseases. Fourth, the characteristics of the diseases in the same cluster and their potential implication on the quality of primary care should be ascertained in greater detail.

Our findings suggest non-hierarchical cluster analysis identified multimorbidity patterns and phenotypes of certain sub-groups of patients that were more consistent with clinical practice.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Supplementary Data

Additional File 1. Extracting and validating multimorbidity patterns by applying the fuzzy c-means clustering algorithm and Computation of the observed/expected ratio and the exclusivity ratio.

Additional File 2. Composition of multimorbidity patterns according to disease levels of prevalence.

Footnotes

Contributors: All authors contributed to the design of the study, revised the article and approved the final version. CV, ARL and SFB obtained the funding. CV, QFB and SFB drafted the article. CV, QFB, SFB, MGC, MCB, ARL contributed to the analysis and interpretation of data. CV, QFB and SFB and CV-F wrote the first draft, and all authors contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript.

Funding

This work was supported by a research grant from the Carlos III Institute of Health, Ministry of Economy and Competitiveness (Spain), awarded on the 2016 call under the Health Strategy Action 2013-2016, within the National Research Program oriented to Societal Challenges, within the Technical, Scientific and Innovation Research National Plan 2013-2016 '[grant number PI16/00639]', co-funded with European Union ERDF funds (European Regional Development Fund) and Department of Health of the Catalan Government, in the call corresponding to 2017 for the granting of subsidies from the Strategic Plan for Research in Health (*Pla Estratègic de Recerca i Innovació en Salut*, PERIS) 2016-2020, modality research oriented to Primary care '[grant number SLT002/16/00058]' and from the Catalan Government '[grant number AGAUR 2017 SGR 578].

Disclaimer: The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service, the National Institute for Health Research or the National Department of Health.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Competing interests None declared.

Patient consent: Detail has been removed from this case description/these case descriptions to ensure anonymity. The editors and reviewers have seen the detailed information available and are satisfied that the information backs up the case the authors are making.

Ethics approval

The protocol of the study was approved by the Committee on the Ethics of Clinical Research, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) (P16/151). All data were anonymized and the confidentiality of EHR was respected at all times in accordance with national and international law.

Data sharing statement: The datasets are not available because researchers have signed an agreement with the Information System for the Development of Research in Primary Care (SIDIAP) concerning confidentiality and security of the dataset that forbids providing data to third parties. This organisation is subject to periodic audits to ensure the validity and quality of the data.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

References

1. Valderas Starfield B, Sibbald B, Salisbuty C, Roland M JM. Defining Comorbidity: Implications for Understanding Health and Health Services. Ann Fam Med 2009; 7:357–63.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, Salisbury C, Blom J, Freitag M, et al. Prevalence, determinants and patterns of multimorbidity in Primary Care: a systematic review of observational studies. PLOS One 2014; 21;9(7): e102149.

3. Salive ME. Multimorbidity in Older Adults. Epidemiol Rev 2013; 35:75-83.

4. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012; 380(9836):37-43.

5. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015; 386 (9995):743-800.

6. Gruneir A, Bronskill SE, Maxwell CJ, Bai YQ, Kone AJ, Thavorn K, et al. The association between multimorbidity and hospitalization is modified by individual demographics and physician continuity of care: a retrospective cohort study. BMC Health Serv Res 2016; 16:154.

7. Rocca WA, Boyd CM, Grossardt BR, Bobo WV, Finney Rutten LJ, Roger VL, et al. Prevalence of multimorbidity in a geographically defined American population: patterns by age, sex, and race/ethnicity. Mayo Clin Proc 2014; 89(10):1336-49.

8. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. J Clin Epidemiol 2014; 67(3):254-66.

9. Muth C, Blom JW, Smith SM, Johnell K, Gonzalez-Gonzalez AI, Nguyen TS, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. J Intern Med 2018; [Epub ahead of print]

10. Palmer K, Marengoni A, Forjaz MJ, Jureviciene E, Laatikainen T, Mammarella F, et al. Multimorbidity care model: Recommendations from the consensus meeting of the Joint Action on Chronic Diseases and Promoting Healthy Ageing across the Life Cycle (JA-CHRODIS). Health Policy 2018;122(1):4-11.

11.Wolfram. Fuzzy Clustering [Internet]. Available from: https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html

12. MathWorks. Fuzzy Clustering [Internet]. Available from: https://www.mathworks.com/help/fuzzy/fuzzy-clustering.html

13. France EF, Wyke S, Gunn JM, Mair FS, McLean G, Mercer SW. Multimorbidity in primary care: a systematic review of prospective cohort studies. Br J Gen Pract 2012; 62 (597): e297-307.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

14. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018; 47(5):1687-1704.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, Rodriguez-Blanco T, Pons-Vigués M, Pujol-Ribera E, et al. Burden of multimorbidity, socioeconomic status and use of health services across stages of life in urban areas: a cross-sectional study. BMC Public Health 2014;14(1):530.

16. Willadsen TG, Bebe A, Køster-Rasmussen R, Jarbøl DE, Guassora AD, Waldorff FB, et al. The role of diseases, risk factors and symptoms in the definition of multimorbidity – a systematic review. Scand J Prim Health Care 2016;34(2):112–21.

17. Xu X, Mishra GD, Jones M. Evidence on multimorbidity from definition to intervention: An overview of systematic reviews. Ageing Res Rev 2017; 7:53-68.

18. Del Mar García-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care 2012;19(3):135–45.

19. Calderón-Larrañaga A, Vetrano DL, Onder G, Gimeno-Feliu LA, Coscollar-Santaliestra C, Carfí A, et al. Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. J Gerontol A Biol Sci Med Sci 2017; 72 (10):1417-1423.

20. Domínguez-Berjón MF, Borrell C, Cano-Serral G, Esnaola S, Nolasco A, Pasarín MI, et al. Constructing a deprivation index based on census data in large Spanish cities (the MEDEA project)]. Gac Sanit 2008; 22(3):179-87.

21. Karlis D, Saporta G, Spinakis A. A simple rule for the selection of principal components. Commun Stat- Theory Methods 2003;32(3):643–66.

22. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2):191–203.

23. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

24. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018;19(1): 108.

25. Prados-Torres A, Poblador-Plou B, Calderón-Larrañaga A, Gimeno-Feliu LA, González-Rubio F, Poncel-Falcó A, et al. Multimorbidity Patterns in Primary Care: Interactions among Chronic Diseases Using Factor Analysis. PLoS One 2012; 7 (2): e32190.

26. Shin HW, Park HK. Recent Updates on Acquired Hepatocerebral Degeneration. Tremor Other Hyperkinet Mov (N Y) 2017;7:463.

27. Wijarnpreecha K, Chesdachai S, Jaruvongvanich V, Ungprasert P. Hepatitis C virus infection and risk of Parkinson's disease: A systematic review and meta-analysis. Eur J Gastroenterol Hepatol 2018;30(1):9–13.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

28. Westfall S, Lomis N, Kahouli I, Dia SY, Singh SP, Prakash S. Microbiome, probiotics and neurodegenerative diseases: deciphering the gut brain axis. Cell Mol Life Sci 2017; 74(20):3769–87.

29. Memba R, Duggan SN, Ni Chonchubhair HM, Griffin OM, Bashir Y, O'Connor DB, et al. The potential role of gut microbiota in pancreatic disease: A systematic review. Pancreatology 2017;17(6):867–74.

30. Xu M, Xu J, Yang X. Asthma and risk of cardiovascular disease or all-cause mortality: A meta-analysis. Ann Saudi Med 2017;37(2):99–105.

31. Cheng CM, Wu YH, Tsai SJ, Bai YM, Hsu JW, Huang KL, et al. Risk of developing Parkinson's disease among patients with asthma: A nationwide longitudinal study. Allergy 2015;70(12):1605–12.

32. Li CH, Chen WC, Liao WC, Tu CY, Lin CL, Sung FC, et al. The association between chronic obstructive pulmonary disease and Parkinson's disease: A nationwide population-based retrospective cohort study. Qjm. 2015;108(1):39–45.

33. Wijarnpreecha K, Boonpheng B, Thongprayoon C, Jaruvongvanich V, Ungprasert P. The association between non-alcoholic fatty liver disease and atrial fibrillation: A meta-analysis. Clin Res Hepatol Gastroenterol 2017 Oct;41(5):525-532.

34. Ambrosino P, Lupoli R, Di Minno A, Tarantino L, Spadarella G, Tarantino P, et al. The risk of coronary artery disease and cerebrovascular disease in patients with hepatitis C: A systematic review and meta-analysis. Int J Cardiol 2016;221:746-54.

35. Kepez A, Mutlu B, Degertekin M, Erol C. Association between left ventricular dysfunction, anemia, and chronic renal failure. Analysis of the Heart Failure Prevalence and Predictors in Turkey (HAPPY) cohort. Herz 2015;40(4):616–23.

36. Chang YL, Hung SH, Ling W, Lin HC, Li HC, Chung SD. Association between ischemic stroke and iron-deficiency anemia: a population-based study. PLoS One 2013;8(12):e82952.

37. Turakhia MP, Blankestijn PJ, Carrero JJ, Clase CM, Deo R, Herzog CA, et al. Chronic kidney disease and arrhythmias: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Eur Heart J 2018;39(24):2314–2325e.

38. Badsha MB, Mollah MN, Jahan N, Kurata H. Robust complementary hierarchical clustering for gene expression data analysis by β-divergence. J Biosci Bioeng 2013;116(3):397-407.

39. Yarnall AJ, Sayer AA, Clegg A, Rockwood K, Parker S, Hindle J V. New horizons in multimorbidity in older adults. Age Ageing 2017;46(6):882–8.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 1. Characteristics of study participants aged 65-94 years stratified by multimorbidity and non-multimorbidity (N= 916 619, Catalonia, 2012)

Variables*	Multimorbidity (n= 853 085)	Non-multimorbidity (n= 63 534)	All (N=916 619)
Sex, women, n (%)	496 294 (58.2)	32 837 (51.7)	529 131 (57.7)
Age, mean (SD)	75.6 (7.4)	73.2 (7.3)	75.4 (7.4)
Age (categories), n (%)			
[65,70)	225 514 (26.4)	26 664 (42.0)	252 178 (27.5)
[70,80)	370 356 (43.4)	24 230 (38.1)	394 586 (43.0)
[80,90)	224 143 (26.3)	10 601 (16.7)	234 744 (25.6)
≥90	33 072 (3.9)	2039 (3.2)	35 111 (3.8)
MEDEA index†			
Q1	130 894 (16.5)	13 897 (23.4)	144 791 (17.0)
Q2	126 537 (16.0)	9894 (16.6)	136 431 (16.0)
Q3	129 246 (16.3)	8976 (15.1)	138 222 (16.2)
Q4	125 322 (15.8)	7666 (12.9)	132 988 (15.6)
Q5	110 916 (14.0)	5967 (10.0)	116 883 (13.7)
Rural	169 190 (21.4)	13 059 (22.0)	182 249 (21.4)
Number of chronic diseases, median [IQR]	6.0 [4.0;8.0]	1.0 [0.0;1.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)			
0	0 (0.0)	25 380 (39.9)	25 380 (2.8)
1	0 (0.0)	38 154 (60.1)	38 154 (4.2)
[2, 5)	268 836 (31.5)	0 (0.0)	268 836 (29.3)
[5,10)	463 709 (54.4)	0 (0.0)	463 709 (50.6)
≥10	120 540 (14.1)	0 (0.0)	120 540 (13.2)
Number of drugs, median [IQR]	5.0 [3.0;8.0]	0.0 [0.0;1.0]	5.0 [2.0;8.0]
Number of drugs (categories):			
0	72 557 (8.5)	40 811 (64.2)	113 368 (12.4)
1	48 704 (5.7)	8378 (13.2)	57 082 (6.2)
[2, 5)	247 095 (29.0)	11 572 (18.2)	258 667 (28.2)
[5,10)	360 030 (42.2)	2651 (4.2)	362 681 (39.6)
≥10	124 699 (14.6)	122 (0.2)	124 821 (13.6)
Number of visits, median [IQR]	10.0 [6.0;17.0]	1.0 [0.0;4.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)			
0	24 543 (2.9)	23,402 (36.8)	47 945 (5.2)
1	24 281 (2.8%)	9603 (15.1%)	33 884 (3.7)
[2, 5)	114 198 (13.4%)	16 241 (25.6%)	130 439 (14.2%)
[5, 10)	239 181 (28.0%)	10 168 (16.0%)	249 349 (27.2%)
≥10	450 882 (52.9%)	4120 (6.5%)	455 002 (49.6%)

All comparisons between variables in multimorbidity and non-multimorbidity showed P<0.001
†MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 2. Prevalence of the 60 chronic diseases included in the study in individuals aged 65-94 years (N= 916 619, Catalonia, 2012). In three last columns, list of diseases included by prevalence cut off (1%, 2%, All)

Rank	*Chronic conditions*	*Frequency*	*Percentage (%)*	*All diseases included*	*1%*	*2%*
1	Hypertension	650 899	71.0			
2	Dyslipidaemia	466 585	50.9			
3	Osteoarthritis and other degenerative joint diseases	300 803	32.8			
4	Obesity	262 888	28.7			
5	Diabetes	230 460	25.1			
6	Anaemia	167 577	18.3			
7	Cataract and other lens diseases	156 622	17.1			
8	Chronic kidney diseases	153 756	16.8			
9	Prostate diseases	153 635	16.8			
10	Osteoporosis	151 847	16.6			
11	Depression and mood diseases	148 751	16.2			
12	Solid neoplasms	137 045	15.0			
13	Colitis and related diseases	131 512	14.4			
14	Venous and lymphatic diseases	126 997	13.9			
15	Other musculoskeletal and joint diseases	124 765	13.6			
16	Dorsopathies	124 603	13.6			
17	Neurotic, stress-related and somatoform diseases	123 395	13.5			
18	COPD, emphysema, chronic bronchitis	109 603	12.0			
19	Ischemic heart disease	95 434	10.4			
20	Deafness, hearing impairment	90 261	9.9			
21	Sleep disorders	88 739	9.7			
22	Thyroid diseases	88 445	9.7			
23	Other genitourinary diseases	85 468	9.3			
24	Cerebrovascular disease	80 264	8.8			
25	Atrial fibrillation	80 247	8.8			
26	Esophagus, stomach and duodenum diseases	80 043	8.7			
27	Heart failure	74 077	8.1			
28	Other eye diseases	68 939	7.5			
29	Glaucoma	66 162	7.2			
30	Inflammatory arthropathies	62 450	6.8			
31	Dementia	59 213	6.5			
32	Cardiac valve diseases	52 100	5.7			
33	Peripheral neuropathy	49 127	5.4			
34	Other psychiatric and behavioural diseases	46 841	5.1			
35	Asthma	43 663	4.8			
36	Allergy	40 394	4.4			
37	Autoimmune diseases	39 350	4.3			
38	Ear, nose, throat diseases	38 752	4.2			
39	Peripheral vascular disease	30 674	3.4			
40	Other neurological diseases	28 541	3.1			
41	Chronic pancreas, biliary tract and gallbladder diseases	27 321	3.0			
42	Migraine and facial pain syndromes	25 999	2.8			
43	Bradycardias and conduction diseases	25 476	2.8			
44	Chronic liver diseases	22 633	2.5			
45	Other digestive diseases	22 022	2.4			
46	Parkinson and parkinsonism	20 833	2.3			
47	Other metabolic diseases	18 997	2.1			
48	Other cardiovascular diseases	16 833	1.8			
49	Other skin diseases	15 363	1.7			
50	Chronic ulcer of the skin	13 869	1.5			
51	Blood and blood forming organ diseases	13 575	1.5			
52	Other respiratory diseases	9974	1.1			
53	Epilepsy	8981	1.0			
54	Haematological neoplasms	8174	0.9			
55	Chronic infectious diseases	6647	0.7			
56	Inflammatory bowel diseases	5549	0.6			
57	Schizophrenia and delusional diseases	4792	0.5			
58	Blindness, visual impairment	4772	0.5			
59	Multiple sclerosis	576	0.1			
60	Chromosomal abnormalities	77	0.0			

Abbreviations: COPD: Chronic obstructive Pulmonary Disease.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 3. Most frequent 15 diseases found in multimorbidity patterns in individuals aged 65-94 years (N= 916 619,

Pattern	Disease	O	O/E ratio	EX	Pattern	Disease	O	O/E ratio	EX
1 Nervous and digestive (n= 40 037)	Parkinson and parkinsonism	38.7	17.0	74.3	2 Respiratory, circulatory and nervous (n= 50 639)	Asthma	34.5	7.2	40.0
	Other neurological diseases	49.5	15.9	69.4		Peripheral vascular disease	13.9	4.2	22.9
	Chronic liver diseases	13.2	5.4	23.4		Parkinson and parkinsonism	8.5	3.8	20.8
	Chronic pancreas, biliary tract and gallbladder diseases	7.9	2.7	11.6		Other neurological diseases	11.7	3.7	20.7
	Dementia	14.7	2.3	9.9		COPD, emphysema, chronic bronchitis	31.0	2.6	14.3
	Other digestive diseases	4.8	2.0	8.7		Allergy	10.8	2.4	13.5
	Cerebrovascular disease	16.9	1.9	8.4		Heart failure	16.6	2.0	11.3
	Colitis and related diseases	24.1	1.7	7.3		Ischemic heart disease	21.1	2.0	11.2
	Other metabolic diseases	3.4	1.7	7.2		Other eye diseases	14.0	1.9	10.3
	Depression and mood diseases	25.0	1.5	6.7		Autoimmune diseases	7.2	1.7	9.3
	Anaemia	26.1	1.4	6.2		Other psychiatric and behavioural diseases	8.5	1.7	9.2
	Esophagus, stomach and duodenum diseases	11.3	1.3	5.6		Ear. nose. throat diseases	7.1	1.7	9.2
	Sleep disorders	12.4	1.3	5.6		Anaemia	30.4	1.7	9.2
	Other eye diseases	9.6	1.3	5.6		Peripheral neuropathy	8.8	1.6	9.1
	Dorsopathies	17.0	1.2	5.4		Cerebrovascular disease	14.3	1.6	9.0
3 Circulatory and digestive (n= 67 492)	Heart failure	51.4	6.4	46.9	4 Mental, nervous and digestive (n= 94 453)	Neurotic, stress-related and somatoform diseases	64.9	4.8	49.7
	Cardiac valve diseases	34.2	6.0	44.3		Depression and mood diseases	66.4	4.1	42.1
	Atrial fibrillation	47.3	5.4	39.8		Migraine and facial pain syndromes	8.2	2.9	29.6
	Bradycardias and conduction diseases	13.5	4.9	35.9		Sleep disorders	19.0	2.0	20.2
	Ischemic heart disease	33.7	3.2	23.8		Esophagus. stomach and duodenum diseases	14.9	1.7	17.6
	Chronic pancreas, biliary tract and gallbladder diseases	8.0	2.7	19.7		Osteoporosis	28.0	1.7	17.4
	Chronic liver diseases	6.1	2.5	18.2		Thyroid diseases	16.0	1.7	17.1
	Chronic kidney diseases	35.9	2.1	15.8		Colitis and related diseases	23.7	1.7	17.0
	Anemia	38.6	2.1	15.5		Other genitourinary diseases	14.4	1.5	15.9
	Cerebrovascular disease	18.3	2.1	15.4		Ear, nose, throat diseases	6.2	1.5	15.2
	COPD, emphysema, chronic bronchitis	23.6	2.0	14.5		Venous and lymphatic diseases	19.9	1.4	14.8
	Other digestive diseases	4.6	1.9	14.0		Allergy	6.1	1.4	14.3
	Peripheral vascular disease	6.1	1.8	13.3		Osteoarthritis and other degenerative joint diseases	45.0	1.4	14.1
	Other metabolic diseases	3.2	1.5	11.3		Dorsopathies	18.0	1.3	13.7
	Dementia	9.5	1.5	10.9		Cardiac valve diseases	7.4	1.3	13.5
5 Mental, digestive and blood (n= 106 845)	Dementia	21.8	3.4	39.4	6 Nervous, musculoskeletal and circulatory (n= 145 074)	Peripheral neuropathy	12.4	2.3	36.6
	Other digestive diseases	5.8	2.4	28.1		Other musculoskeletal and joint diseases	26.0	1.9	30.2
	Anemia	38.5	2.1	24.6		Venous and lymphatic diseases	26.4	1.9	30.2
	Chronic kidney diseases	33.3	2.0	23.1		Dorsopathies	25.3	1.9	29.4
	Colitis and related diseases	26.2	1.8	21.3		Obesity	51.0	1.8	28.2
	Cerebrovascular disease	14.8	1.7	19.7		Other genitourinary diseases	16.0	1.7	27.2
	Osteoporosis	26.0	1.6	18.3		Osteoarthritis and other degenerative joint diseases	55.0	1.7	26.5
	Cataract and other lens diseases	25.9	1.5	17.7		Osteoporosis	24.8	1.5	23.7
	Deafness. hearing impairment	14.0	1.4	16.5		Other eye diseases	10.7	1.4	22.4
	Venous and lymphatic diseases	19.5	1.4	16.4		Cataract and other lens diseases	22.5	1.3	20.8
	Osteoarthritis and other degenerative joint diseases	45.5	1.4	16.2		Thyroid diseases	12.6	1.3	20.7
	Depression and mood diseases	22.5	1.4	16.1		Glaucoma	9.2	1.3	20.1
	Other genitourinary diseases	12.3	1.3	15.4		Diabetes	31.3	1.2	19.7
	Other eye diseases	9.9	1.3	15.4		Ear, nose, throat diseases	5.2	1.2	19.5
	Sleep disorders	12.4	1.3	14.9		Dyslipidemia	62.7	1.2	19.5
7 Genitourinary, mental and musculoskeletal (n=173 746)	Prostate diseases	54.7	3.3	61.8	8 Non-specified (n=238 333)	Dyslipidemia	38.4	0.8	19.6
	Other psychiatric and behavioural diseases	11.1	2.2	41.2		Thyroid diseases	7.3	0.8	19.6
	Inflammatory arthropathies	12.4	1.8	34.5		Osteoporosis	12.2	0.7	19.2
	COPD, emphysema, chronic bronchitis	20.5	1.7	32.5		Hypertension	47.6	0.7	17.4
	Solid neoplasms	21.8	1.5	27.7		Glaucoma	4.4	0.6	16.0
	Peripheral vascular disease	4.7	1.4	26.7		Solid neoplasms	9.1	0.6	15.7
	Ischemic heart disease	13.7	1.3	25.0		Migraine and facial pain syndromes	1.7	0.6	15.7
	Diabetes	31.8	1.3	24.0		Autoimmune diseases	2.2	0.5	13.4
	Ear, nose, throat diseases	5.3	1.3	23.7		Other metabolic diseases	1.1	0.5	13.3
	Deafness, hearing impairment	11.6	1.2	22.3		Allergy	2.2	0.5	13.0
	Allergy	4.8	1.1	20.5		Chronic liver diseases	1.2	0.5	12.8
	Hypertension	75.8	1.1	20.2		Other genitourinary diseases	4.5	0.5	12.7
	Glaucoma	7.5	1.0	19.6		Esophagus, stomach and duodenum diseases	4.1	0.5	12.2
	Autoimmune diseases	4.4	1.0	19.4		Other psychiatric and behavioral diseases	2.4	0.5	12.0
	Obesity	29.0	1.0	19.2		Diabetes	10.8	0.4	11.2

Abbreviations: O: Disease prevalence in the cluster; O/E ratio: observed/expected ratio; Ex: exclusivity; COPD: Chronic obstructive Pulmonary Disease.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 4. Variables characterizing each cluster in baseline study (N= 916 619)

	1.Nervous and digestive	2. Respiratory, circulator and nervous	3. Circulatory and digestive	4. Mental, nervous and digestive	5. Mental, digestive and blood	6. Nervous, musculoskeletal and circulatory	7. Genitourinary, mental and musculoskeletal	8. Non-specified
Multimorbidity, n (%)	39 776 (99.3)	50 513 (99.8)	67 443 (99.9)	94 442 (100.0)	106 696 (99.9)	144 869 (99.9)	171 983 (99.0)	177 363 (74.4)
Polypharmacy, n (%)	28 484 (71.1)	38 869 (76.8)	54 658 (81.0)	64 154 (67.9)	71 830 (67.2)	86 317 (59.5)	90 603 (52.1)	52 588 (22.1)
Women, n (%)	22 628 (56.5)	26 690 (52.7)	38 023 (56.3)	78 922 (83.6)	85 735 (80.2)	113 629 (78.3)	15 730 (9.1)	147 773 (62.0)
Men, n (%)	17 409 (43.5)	23 949 (47.3)	29 469 (43.7)	15 531 (16.4)	21 110 (19.8)	31 445 (21.7)	158 016 (90.9)	90 560 (38.0)
Age (categories), n (%)								
[65,70)	7188 (18.0)	10 400 (20.5)	7233 (10.7)	28 305 (30.0)	12 036 (11.3)	38 829 (26.8)	52 003 (29.9)	96 184 (40.4)
[70,80)	17 804 (44.5)	22 743 (44.9)	24 724 (36.6)	40 577 (43.0)	33 624 (31.5)	70 643 (48.7)	84 037 (48.4)	100 435 (42.1)
[80,90)	13 460 (33.6)	15 568 (30.7)	29 908 (44.3)	22 638 (24.0)	48 453 (45.3)	32 714 (22.5)	34 785 (20.0)	37 217 (15.6)
[90,99]	1587 (4.0)	1927 (3.8)	5628 (8.3)	2934 (3.1)	12 732 (11.9)	2888 (2.0)	2920 (1.7)	4497 (1.9)
MEDEA* index								
R	7831 (21.8)	9300 (20.2)	13 718 (23.2)	17 266 (19.7)	22 183 (23.0)	27 401 (21.1)	35 145 (21.5)	49 405 (21.9)
U1	6010 (16.7)	6890 (15.0)	9537 (16.1)	15 027 (17.2)	16 556 (17.2)	19 599 (15.1)	25 656 (15.7)	45 516 (20.2)
U2	5690 (15.8)	7134 (15.5)	9140 (15.4)	14 335 (16.4)	15 272 (15.8)	21 379 (16.5)	25 951 (15.9)	37 530 (16.6)
U3	5941 (16.5)	7520 (16.4)	9187 (15.5)	14 223 (16.3)	15 421 (16.0)	23 261 (17.9)	26 908 (16.5)	35 761 (15.8)
U4	5540 (15.4)	7686 (16.7)	9016 (15.2)	14 012 (16.0)	14 272 (14.8)	23 780 (18.3)	26 526 (16.2)	32 157 (14.2)
U5	4982 (13.8)	7421 (16.2)	8638 (14.6)	12 652 (14.5)	12 699 (13.2)	21 923 (16.9)	23 064 (14.1)	25 506 (11.3)
Number of chronic diseases, median [IQR]	8.0 [6.0;10.0]	8.0 [6.0;10.0]	8.0 [7.0;11.0]	7.0 [6.0;9.0]	7.0 [5.0;9.0]	6.0 [5.0;8.0]	5.0 [4.0;7.0]	3.0 [3.0;4.0]
Number of chronic diseases (categories), n (%)								
0	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.0%)	0 (0.0%)	235 (0.1)	25 144 (10.5)
1	262 (0.7)	125 (0.2)	49 (0.1)	11.0 (0.0)	149 (0.1)	204 (0.1)	1528 (0.9)	35 826 (15.0)
[2, 5)	5409 (13.5	4507 (8.9)	4275 (6.3)	8781 (9.3)	14 601 (13.7)	22 400 (15.4)	57 561 (33.1)	151 302 (63.5)
[5,10)	23 502 (58.7)	30 257 (59.8)	37 910 (56.2)	62 490 (66.2)	73 427 (68.7)	105 620 (72.8)	104 915 (60.4)	25 588 (10.7)
≥10	10 864 (27.1)	15 749 (31.1)	25 259 (37.4)	231 715 (24.5)	18 668 (17.5)	16 850 (11.6)	9506 (5.5)	473 (0.2)
Number of drugs, median [IQR]	7.0 [4.0;9.0]	7.0 [5.0;10.0]	8.0 [5.0;11.0]	6.0 [4.0;9.0]	6.0 [4.0;9.0]	5.0 [3.0;8.0]	5.0 [3.0;7.0]	2.0 [0.0;4.0]
Number of drugs (categories)								
0	2576 (6.4)	2491 (4.9)	3349 (5.0)	5636 (6.0)	7,037 (6.6)	8330 (5.8)	13 389 (7.7)	70 561 (29.6)
1	1212 (3.0)	1072 (2.1)	1015 (1.5)	2939 (3.1)	3390 (3.2)	6772 (4.7)	11 440 (6.6)	29 242 (12.3)
[2, 5)	7766 (19.4)	8207 (16.2)	8471 (12.6)	21 725 (23.0)	24 587 (23.0)	43 656 (30.1)	58 314 (33.6)	85 942 (36.1)
[5,10)	18 510 (46.2)	23 597 (46.6)	31 850 (47.2)	46 022 (48.7)	52 653 (49.3)	68 193 (47.0)	73 694 (42.4)	48 161 (20.2)
≥10	9973 (24.9)	15 272 (30.2)	22 808 (33.8)	18 132 (19.2)	19 177 (17.9)	18 123 (12.5)	16 909 (9.7)	4427 (1.9)
Number of visits 2012, median [IQR]	12.0 [7.0;20.0]	14.0 [8.0;22.0]	18.0 [9.0;30.0]	11.0 [6.0;19.0]	12.0 [7.0;19.0]	11.0 [7.0;17.0]	9.0 [5.0;15.0]	5.0 [2.0;9.0]
Number of visits 2012 (categories), n (%)								
0	976 (2.4)	871 (1.7)	1143 (1.7)	2219 (2.3)	2515 (2.4)	2410.3 (1.7)	4137 (2.4)	33 673 (14.1)
1	874 (2.2)	754 (1.5)	929 (1.4)	2055 (2.2)	2238 (2.1)	2412.4 (1.7)	4685 (2.7)	19 938 (8.4)
[2, 5)	4000 (10.0)	3918 (7.7)	4329 (6.4)	10 589 (11.2)	11 018 (10.3)	14943.7 (10.3)	24 319 (14.0)	57 322 (24.1)
[5, 10)	9158 (22.9)	10 774 (21.3)	10 883 (16.1)	24 504 (25.9)	27 003 (25.3)	42180.7 (29.1)	54 212 (31.2)	70 634 (29.6)
≥10	25 030 (62.5)	34 322 (67.8)	50 209 (74.4)	55 085 (58.3)	64 071 (60.0)	83126.5 (57.3)	86 393 (49.7)	56 766 (23.8)

*MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l
d by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Supérieur (ABES) .
For peer review only

Figure 1. Study population flow chart

*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

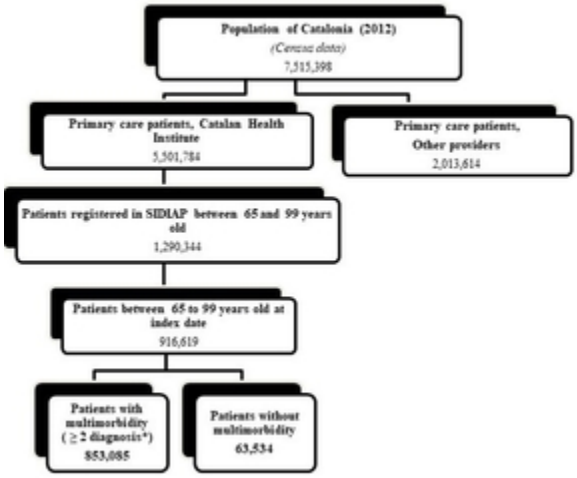


Figure 1. Study population flow chart
*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

25x22mm (300 x 300 DPI)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

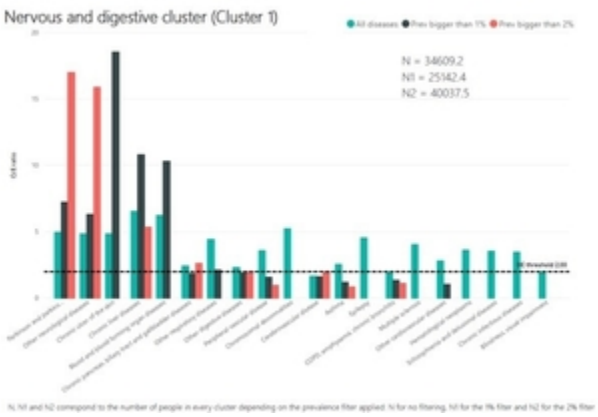


Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

26x17mm (300 x 300 DPI)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Additional File 1

A) Extracting and Validating Multimorbidity Patterns by applying the Fuzzy C Means Clustering algorithm.

In this annex we present a description of the procedure followed to obtain a set of multimorbidity patterns characterizing a patient population aged 65 or more in Catalonia (Spain).

Dataset dimension reduction.

The initial dataset was composed on 31st December, 2012, of a registered active diagnosis with a certain prevalence value, out of 60 possible diseases for the N=916,619 patients included in the study. Additionally, considering age and the gender, each patient was initially characterized by a vector of 62 features, most of which were binary variables indicating the presence/absence of a disease at the end of 2012. For most of the study, diseases with prevalence ≥2% were filtered, resulting in 47 diseases and the corresponding 49 features (adding age and gender). Since most of the selected features were categorical instead of quantitative, the dataset was a mixture of numerical and categorical variables. We processed this dataset by applying a mixture of the well-known Principal Component Analysis (PCA) to the numeric original features and a Multiple Correspondence Analysis (MCA) to the binary ones, in order to obtain a new dataset of reduced dimension. We selected the PCAmix algorithm, as described by Chavent et al, to perform the dimensionality reduction. It follows the criterion based on concentrating most of the variability of the new transformed features, that is to say, variance of the data in the low-dimensional representation were maximized. The Karlis-Saporta-Spinaki rule was followed to select the first 13 dimensions out of the 49 for the 2% prevalence filtering, according to the eigenvalues of the PCAmix and the number of features and individuals in the dataset. As a result, after the PCAmix transformation and the extraction of the optimal number of dimensions, the new dataset was composed of N=916,619 vectors of $d = 13$ features each one. In the following we denote this new dataset as $\boldsymbol{\mathcal{Y}} \coloneqq \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$, denoting by $\boldsymbol{y}_2 \in \mathbb{R}^{13}$ for $n = 1, \dots, N$ the new vector representing patient n.

Soft clustering algorithm

Once the transformed dataset $\boldsymbol{\mathcal{Y}}$ was computed, a soft clustering algorithm was applied to fuzzily distribute the population into a set of clusters, corresponding to the different multimorbidity patterns. In a traditional clustering procedure patients are grouped in an exclusive way, so that if a certain patient belongs to a definite cluster then s/he cannot be included in another one. In contrast, an overlapping clustering, such as the Fuzzy C Means

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

(FCM) algorithm, uses fuzzy sets to cluster patients, so that each patient belongs to all clusters with different degrees of membership. The choice between a hard or a soft clustering algorithm is traditionally made based on the application and the performance obtained. In our case, the use of the FCM algorithm presented performance results similar to those of the hard clustering algorithm Kmeans, but clinically more solid. It was, therefore, chosen as the most appropriate method for the description of the multimorbidity patterns.

FCM was originally introduced by Bezdek and yields an unsupervised form of grouping in which individuals can belong to more than one cluster. To do so, they are associated with an appropriate set of K membership values, where K denotes the number of clusters. The parameters that determine the clustering process are a set of K centroids $\mathbf{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_K\}$ where $v_k \in \mathbb{R}^{13}$ for $k = 1, \dots, K$ and a set of membership factors $\mathbf{U} = \{u_{jn}; j = 1, \dots, K; n = 1, \dots, N\}$ with $0 \leq u_{jn} \leq 1$. Factor u_{jn} indicates the degree to which individual n^{th} belongs to cluster j^{th}. Both centroids $\mathbf{V}$ and membership factors $\mathbf{U}$ are obtained by iteratively minimizing the objective function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$, which is the weighted sum of squared errors within clusters

$$J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) = \sum_{n=1}^{N} \sum_{j=1}^{K} (u_{jn})^m \left\| \boldsymbol{y}_n - \boldsymbol{v}_j \right\|^2; \quad 1 < m < \infty \quad (1)$$

Thus, the similarity between an individual and a cluster centroid is measured through the squared error between the vector associated with the patient and the centroid prototyping the cluster. The fuzziness weighting parameter m, is selected to adjust the blending of the different clusters and it is any real number greater than 1. High m values would produce a fuzzy set of clusters so that individuals would tend to be equally distributed across clusters, whereas lower ones would generate a non-overlapped set of clusters. The FCM method iteratively alternates between computing the centroids in $\mathbf{V}$ as the average of the individual's features in $\boldsymbol{\mathcal{Y}}$ previously weighted by the correspondent membership factors and estimating the membership factors in $\mathbf{U}$ in order to maximize the cost function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$ given the updated centroids in $\mathbf{V}$. In our work, we randomly initialized the set of centroids $\mathbf{V}$ and halted the iterative process when $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) < \epsilon$, where $0 < \epsilon \ll 1$. This procedure converges to a local minimum or saddle point of $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$.

Cluster stability validation.

Stable clusters are required in order to characterize multimorbidity patterns, consequently we applied 100 FCM independent runs to the transformed dataset $\boldsymbol{\mathcal{Y}}$ and averaged both the membership factors and the centroid vectors, after ordering the clusters in descending order in terms of the summation of memberships to clusters, measured as $\sum_{n=1}^{N} (u_{jn})^m$. This is equivalent to selecting the centroid and membership factors associated with the cluster with

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

more population in each run and averaging them. Then after removing the selected cluster from each set, the procedure is repeated until a final set of clusters, composed of the K averaged centroids and the corresponding averaged membership factors, is obtained. In this averaging process we previously verified the similarity between the averaged parameters by a heuristic inspection of some randomly selected run results

Number of clusters and fuzziness parameter validation.

Since clustering algorithms are unsupervised, machine-learning techniques, the model fitting the dataset is traditionally computed through cost functions that depend on both the dataset and the clustering parameters and are denoted as validation indices. We computed three different well-known validation indices to obtain the optimal number of clusters K and the optimal value of the fuzziness parameter m: the partition coefficient validation index whose cost function is maximum for the optimal model, the Xie-Beni, and the partition entropy validation indices whose cost functions are minimum for the optimal models. A cross-validation technique was applied using a split sample approach, by randomly dividing the individuals into two different datasets, a first (50%) training dataset used for obtaining the averaged FCM clusters, and a second (50%) test dataset used to verify the model fitting the data.

This validation procedure was applied to the set of clusters obtained after the previously explained averaging process, with the 2% prevalence filtering and considering 49 features before PCAmix reduction. We checked $m = 1.1, 1.2,$ and 1.5 and $K = 5,..,20$. In Figure1 the performance obtained through the three validation indices is depicted. The behaviour for m=1.1 is shown in Figure 2 and from Figure 3 we can conclude that the optimal number of clusters for m=1.1 ranges from 6 to 12, validated with both the training dataset and the test dataset (more details in figures).

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

B) Computation of the observed/expected ratio and the exclusivity ratio.

The observed/expected $(O/E)_{dj}$ ratio and the exclusivity ratio EX_{dj} have been used in this work in order to decide whether a disease d is overrepresented or not in any given cluster j.

The $(O/E)_{dj}$ ratio was calculated by dividing disease prevalence in the cluster O_{dj} by disease prevalence in the overall population E_d. As membership of an individual n in a cluster j was denoted by a membership degree factor u_{nj}, and not as a binary variable, the observed disease prevalence O_{dj} in a cluster j was computed as the ratio between the summation of the membership degree factors corresponding to all individuals suffering the disease d and the summation of all the membership degree factors corresponding to the cluster j. Let us assume that there are n_d individuals suffering the disease d and that they are grouped in the set I_d, then the observed prevalence was computed as

$$O_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{\sum_{n=1}^{N} u_{nj}}$$

while the expected prevalence was computed as

$$E_d = \frac{n_d}{N}$$

Exclusivity ratio EX_{dj}, defined as the proportion of individuals with the disease d included in the cluster j over the total number of individuals with the disease n_d, was computed as

$$EX_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{n_d}$$

References

1. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate analysis of mixed data: The PCAmixdata R package. 2014; eprint arXiv:1411.4911.

2. Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press; 1981.

3. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

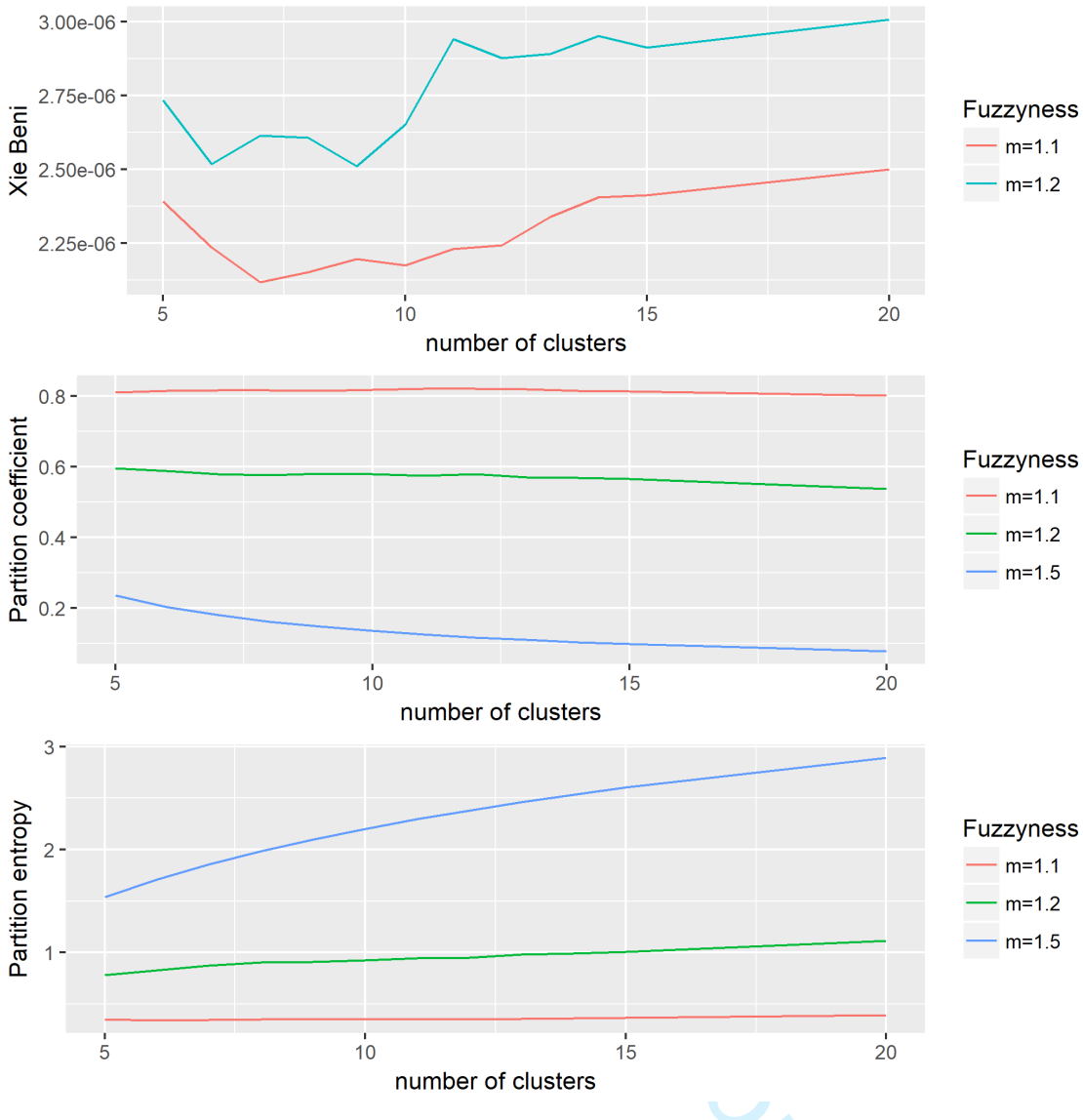
4. Pal NR, Bezdek JC. On Cluster Validity for the Fuzzy c-Means Model. IEEE Trans Fuzzy Syst 1995;3(3):370–9.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Figure 1. Selection of the optimal m parameter



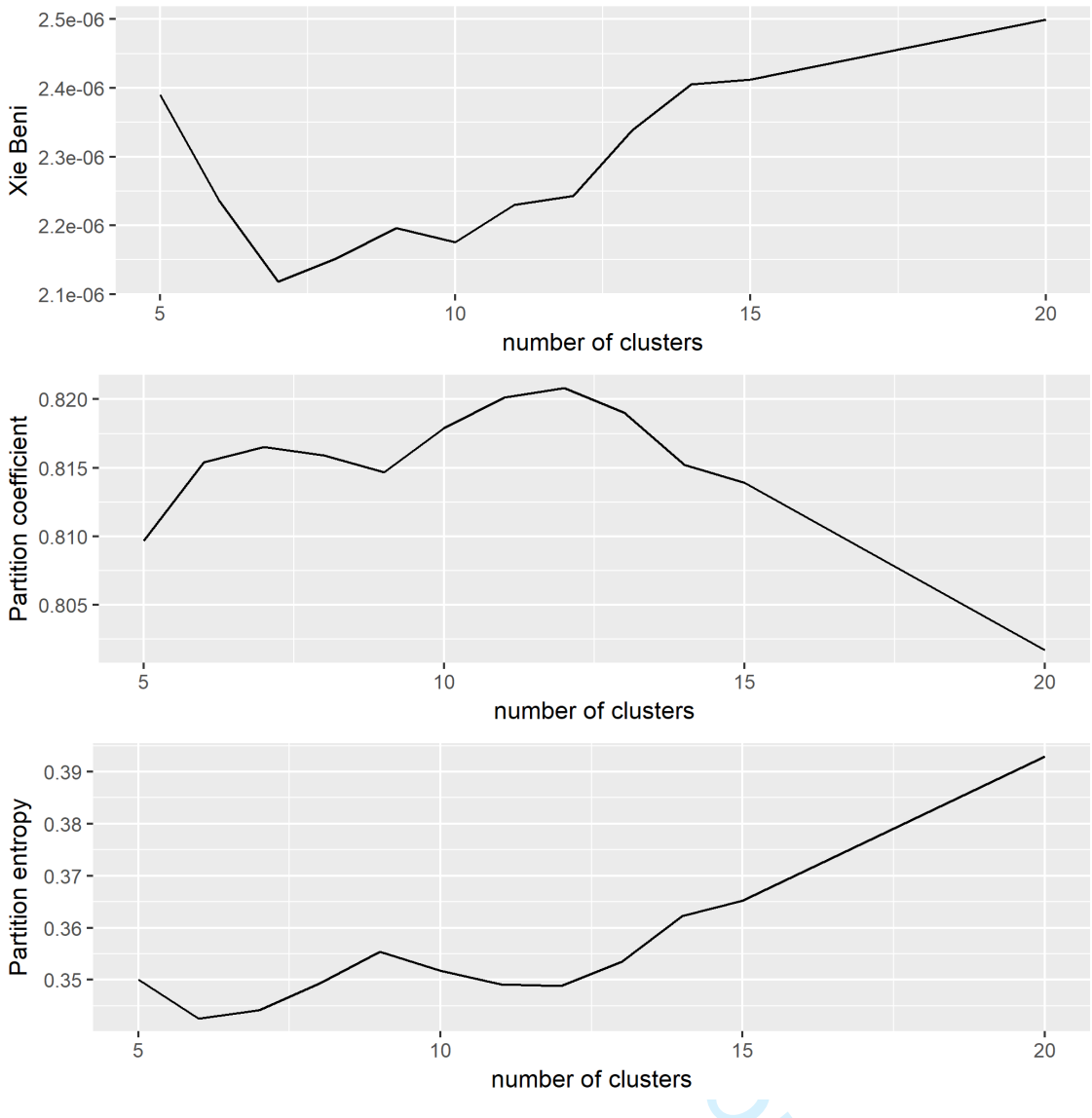
Index m = 1.5 was also computed for Xie-Beni indices, but not included in the graph because the curve is significantly higher than the other two in the plot.

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. For this reason, all plots are showing that m = 1.1 is the best parameter to optimize all the computed indices.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

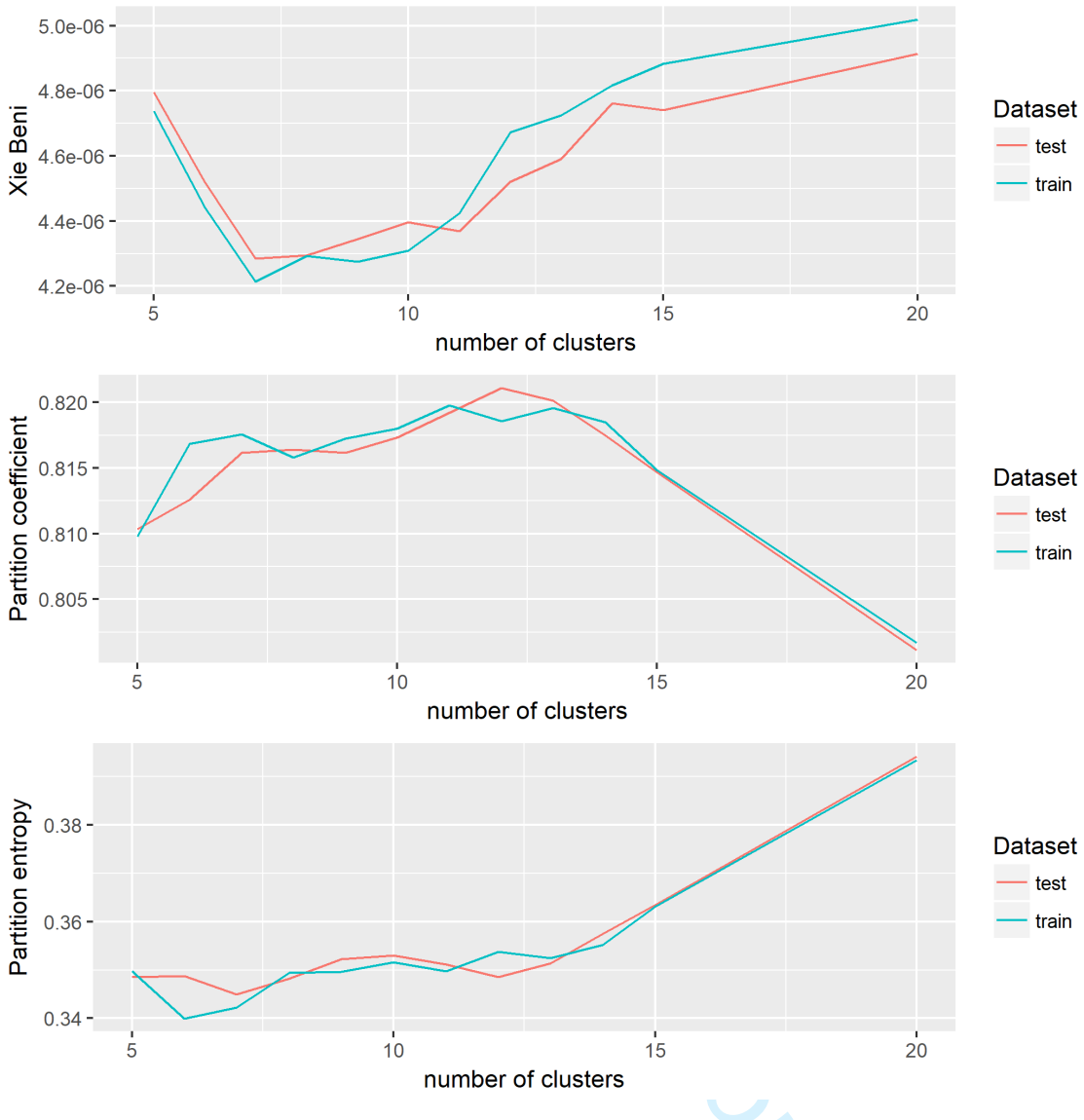
Figure 2. Selection of the optimal number of clusters (m = 1.1)



Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. Within the plots above, optimal values are located in the range from 6 to 12 clusters.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

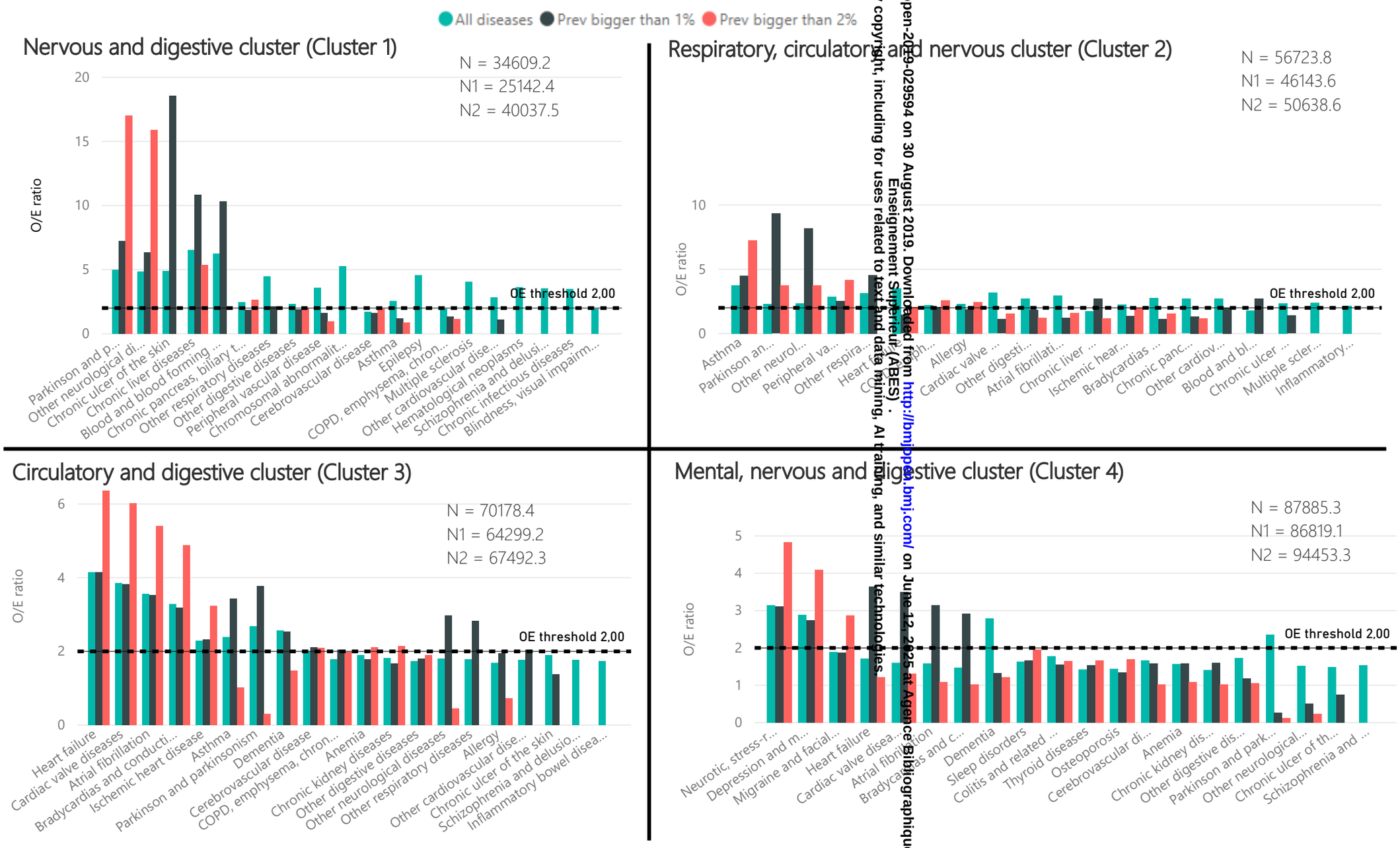
Figure 3. Cross-validation of the clustering with m = 1.1



Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. In the plots above we can find the optimal values in the range from 6 to 12 clusters. Additionally, no significant variation is registered in the indices regardless of the dataset selection.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

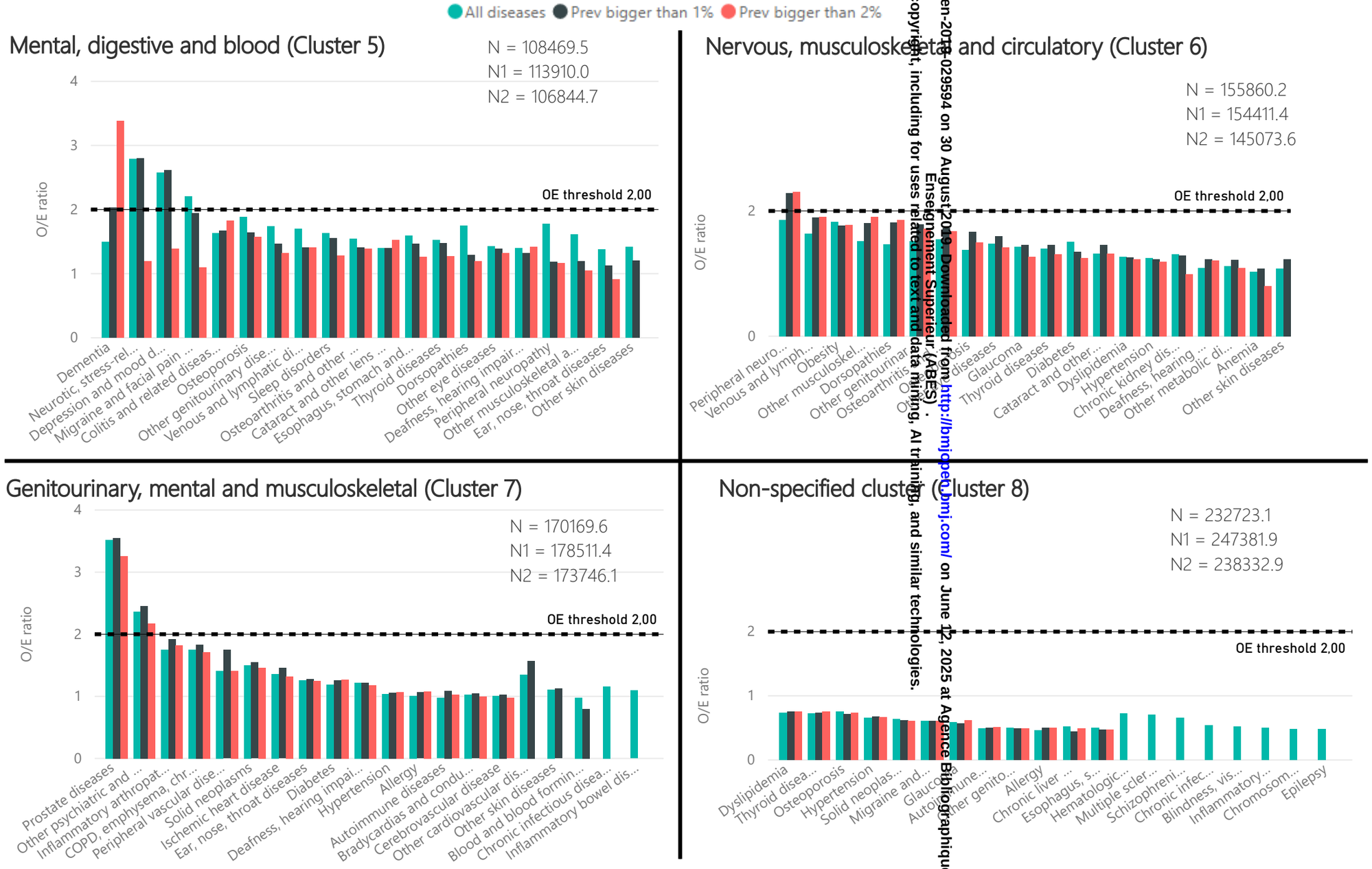
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .



N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied: N for no filtering, N1 for the 1% filter and N2 for the 2% filter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

36/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES) . cted by copyright, including for uses related to text and data mining, AI training, and similar technologies.



N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied: N for no filtering, N1 for the 1% filter and N2 for the 2% filter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

36/bmjopen-2018-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES) .
cted by copyright, including for uses related to text and data mining, AI training, and similar technologies.

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	5
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*6*
Bias	9	Describe any efforts to address potential sources of bias	7
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	6
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	6
		(*b*) Describe any methods used to examine subgroups and interactions	6
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	
		(*e*) Describe any sensitivity analyses	7
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	Figure 1
		(c) Consider use of a flow diagram	Figure 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 1
		(b) Indicate number of participants with missing data for each variable of interest	Tables
Outcome data	15*	Report numbers of outcome events or summary measures	8-9

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	8-9 Tables
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Additional File 1
Discussion			
Key results	18	Summarise key results with reference to study objectives	10
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	12
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	12
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	14

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029594.R1
Article Type:	Research
Date Submitted by the Author:	18-Apr-2019
Complete List of Authors:	Violan-Fors, Concepción; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Foguet-Boreu, Quintí; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol) Universitat Autònoma de Barcelona, ; Hospital de Campdevànol, Emergency room Fernández-Bertolín, Sergio; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Guisado-Clavero, Marina; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Cabrera-Bean, Margarita; Universitat Politecnica de Catalunya, Signal Theory and Comunications Department Formiga, F; Hospital Universitari de Bellvitge Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol),
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, General practice / Family practice
Keywords:	Chronic conditions, Multimorbidity, Cluster analysis, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population

1. Concepción Violán-Fors*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

2. Quintí Foguet-Boreu*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Department of Psychiatry, Vic University Hospital. Francesc Pla el Vigatà, 1, 08500 Vic, Barcelona, Spain. 4. Department of Basic and Methodological Sciences. Faculty of Health Sciences and Welfare. University of Vic-Central University of Catalonia (UVic-UCC)
E-mail: 42292qfb@comb.cat

3. Sergio Fernández-Bertolín. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: sfernandez@idiapjgol.org

4. Marina Guisado-Clavero. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: marina.guisado@gmail.com

5. Margarita Cabrera-Bean. Signal Theory and Comunications Department, Universitat Politècnica de Catalunya, Barcelona Tech. Campus Nord, UPC D5, Jordi Girona 1-2, 08034-Barcelona, Spain.
E-mail: marga.cabrera@upc.edu

6. Francesc Formiga. Internal Medicine Service, Hospital Universitari de Bellvitge, Hospitalet del Llobregat, Barcelona, Catalonia, Spain.
E-mail: fformiga@bellvitgehospital.cat

7. Jose M Valderas. Health Services & Policy Research Group, Academic Collaboration for Primary Care, University of Exeter Medical School, Exeter, EX1 2LU, United Kingdom.
E-mail: J.M.Valderas@exeter.ac.uk

8. Albert Roso-Llorach. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

Corresponding author: Concepción Violán. IDIAPJGol
Quintí Foguet-Boreu. IDIAPJGol
Gran Via Corts Catalanes, 587 àtic.08007 Barcelona. Spain.
Telephone: 0034 93 482 41 24. FAX: 0034 93 482 41 74.
Web page: www.idiapjgol.org.E-mail: cviolan@idiapjgol.org; 42292qfb@comb.cat
Word count: 3 164

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Abstract

Objectives The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied. Fuzzy cluster analysis allows individuals to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.

Design A cross-sectional study was conducted based on data from electronic health records

Setting 284 primary health care centres in Catalonia, Spain (2012).

Participants 916 619 eligible individuals were included (women: 57.7%).

Primary and secondary outcome measures We extracted data on demographics, ICD-10 chronic diagnoses, prescribed drugs, and socioeconomic status for patients aged ≥65. Following principal component analysis of categorical and continuous variables (PCAmix) for dimensionality reduction, machine learning techniques were applied for the identification of disease clusters in a fuzzy c-means analysis. Sensitivity analyses, with different prevalence cut-off points for chronic diseases, were also conducted. Solutions were evaluated from clinical consistency and significance criteria.

Results Multimorbidity was present in 93.1%. Eight clusters were identified with a varying number of disease values: *Nervous and digestive*; *Respiratory, circulatory, and nervous*; *Circulatory, and digestive*; *Mental, nervous, and digestive*; *Mental, digestive, and blood*; *Nervous, musculoskeletal, and circulatory*; *Genitourinary, mental, and musculoskeletal*; and *Non-specified*. Nuclear diseases were identified for each cluster independently of the prevalence cut-off point considered.

Conclusions Multimorbidity patterns were obtained using fuzzy c-means cluster analysis. They are clinically meaningful clusters which support the development of tailored approaches to multimorbidity management and further research.

Keywords: Chronic conditions; Multimorbidity; Epidemiology; Cluster analysis.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

Strengths and limitations of this study

- Studies focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns (hard clustering forces each individual to belong to a single cluster, whereas soft clustering allows elements to be simultaneously classified into multiple cluster).
- Reliable and valid identification of disease clusters is needed for the development of evidence-based clinical practice guidelines and pathways of care for patients that correspond to the wide spectrum of diseases in patients with multimorbidity.
- Soft clustering analysis allows for diseases to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.
- The different cut-off points (prevalence filters) applied to obtain multimorbidity patterns permitted the identification of common nuclear diseases which remained independent of their prevalence.
- The literature provides support for the etiopathophysiological and epidemiological associations between conditions forming part of the same cluster.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Introduction

The term multimorbidity widely refers to the existence of numerous medical conditions in a single individual (1). In many regions of the world there is evidence that a substantial, and probably growing, proportion of the adult population is affected by multiple chronic conditions. Moreover, the association of multimorbidity with increasing age leading to a two-fold prevalence in the final decades of life has been proven (2). Multimorbidity has been estimated to be at around 62% between 65 and 74 years, and around 81.5% after 85 years (3). Its true extent is, however, difficult to gauge as there is no agreed definition or classification system (4-7).

Most of the published literature focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns (8). Orienting the analysis of multimorbidity patterns at an individual level, and not of disease, could have crucial implications for patients. In the current context of limited evidence on interventions for unselected patients with multimorbidity, such an approach would allow better understanding of population groups, and facilitate the development and implementation of strategies aimed at prevention, diagnosis, treatment, and prognosis. It would also elicit essential information for the development of clinical guidelines, pathways of care, and lead to better understanding of the nature and range of the required health services (9,10).

Cluster analysis involves assigning individuals so that the items (diseases) in the same cluster are as similar as possible, while individuals belonging to different clusters are as dissimilar as possible. The identification of clusters is based on similarity measures and their choice may depend on the data or the purpose of the analysis (11,12). Hard clustering forces each element to belong to a single cluster, whereas soft clustering (also referred to as fuzzy clustering) allows elements to be simultaneously classified into multiple clusters.

Empirical evidence is needed on how both established and novel techniques influence the identification of multimorbidity patterns. A recent systematic review recommended that future epidemiological studies cover a broad selection of health conditions in order to avoid missing

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

potentially key nosological associations and enhance external validity. When many conditions are considered, the clustering of individuals based on morbidity data will encounter high-dimensional issues. This is particularly important when a clustering-based approach is adopted to assess the impact of multimorbidity on individual health outcomes and health service uses (2, 8, 13-15).

The identification of multimorbidity patterns seems to be implicitly dependent on the prevalence of the included diseases (2,8,16,17). However, to the best of our knowledge no previous study has analysed the identification of multimorbidity patterns explicitly based on the prevalence of the diseases.

The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied.

Methods

Study population

A cross-sectional analysis was carried out in Catalonia (Spain), a Mediterranean region of 7,515,398 inhabitants (2012). The Catalan Health Institute provides universal coverage and operates 284 primary health care centres (PHC).

Data sources

Since 2006 the Information System for Research in Primary Care (SIDIAP) database includes anonymized longitudinal electronic health records from primary and secondary care which gather information on demographics, diagnoses, prescriptions, and socioeconomic status (18). In our study the inclusion criteria were individuals aged 65-99 years on 31st December 2011

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

with at least one PHC visit since 2012. Only participants that survived until 31st December 2012 (index date) were included in the analysis.

Variables

Diseases were coded in the SIDIAP using the International Classification of Diseases version 10 (ICD-10). An operational definition of multimorbidity was the simultaneous presence of more than one of the selected 60 chronic diseases previously identified by the Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (19).

Additional variables included in the study were sociodemographics (age, sex, socio-economic status (MEDEA index) (20), clinical variables (including number of chronic diseases and invoiced drugs), and use of health services (number of visits to family physicians, nurses, and emergency services).

Statistical analysis

Descriptive statistics were used to summarize overall information. Disease prevalence was computed for all the included population. Descriptive analyses were stratified by the presence of multimorbidity. Comparison was performed using t-Student or Mann-Whitney for continuous variables and Chi-Square for categorical ones.

In order to obtain the most representative clusters all patients were included irrespective of whether they presented multimorbidity or not. Sex and age variables, together with chronic diseases selected by prevalence, were included in the analysis. The number of features to be considered varied from the 62 original ones (no prevalence filtering applied) to 54 and 49, for a 1% and 2% prevalence threshold, respectively.

Due to the large number of diseases, a principal component analysis for categorical and continuous data (PCAmix) was implemented to reduce complexity. With this technique both continuous and dichotomous variables were simultaneously processed through the application of Multi Correspondence Analysis to the binary variables and PCA to the continuous ones.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Using Karlis-Saporta-Spinaki criterion to select the optimal number of dimensions to retain, the dataset of 49 features per individual per 2% prevalence cut-off was transformed to a new dimensionally reduced dataset of 13 continuous features per individual, which concentrated most of the variability of the newly transformed dataset (21).

Once the transformed dataset was obtained, clusters of chronic conditions at baseline were identified using the fuzzy c-means clustering algorithm (22). This machine learning technique forces every individual to belong to every cluster in accordance with its characteristics and by assigning a membership degree factor in (0,1) to each individual with respect to each pattern. This provides the flexibility enabling patients to belong to more than one multimorbidity pattern (23).

The main parameters in this clustering procedure were the number of clusters and a fuzziness parameter, denoted *m*, that ranged from just above 1 to infinity. High *m* values produce a fuzzy set of clusters, so that individuals are equally distributed across clusters, whereas lower ones generate non-overlapped clusters. Further details on the stability and validation techniques applied to obtain the best fuzzy c-means parameters and the set of centroids, are presented in Additional File 1.

To describe the multimorbidity patterns, frequencies and percentages of diseases (P) in each cluster were calculated. Observed/expected ratios (O/E-ratios) were calculated by dividing disease prevalence in the cluster by disease prevalence in the overall population. As the membership of each individual to any of the clusters was given by a membership degree factor, and not as a binary variable, the observed disease prevalence (O) in a cluster was computed as the sum of the disease membership degree factors corresponding to all individuals suffering the disease. Exclusivity, defined as the proportion of patients with the disease included in the cluster over the total number of patients with the disease, was also calculated. Further details on how these ratios were computed using the membership factors are given in Additional File 1. A

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

disease was considered to be part of a multimorbidity cluster when O/E-ratio was ≥2 or exclusivity value ≥25% (24).

We conducted a sensitivity analysis by modifying the prevalence threshold for disease inclusion in the cluster analysis. For chronic diseases we considered as alternatives no filtering, and ≥1% and ≥2% filters among the included population. In order to conform to the Karlis-Saporta-Spinaki rule, a different number of dimensions of the transformed dataset were retained to construct the clusters for every prevalence cut-off: 13 dimensions for the 2% prevalence, 14 dimensions for the 1% prevalence, and 17 dimensions with no filtering. The content of each cluster was compared across filtering approaches in terms of diseases associated with that cluster, characteristics of the included population, and cluster size. Clinical evaluation of the consistency and significance of these solutions was also conducted.

The analyses were carried out using R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). The significance level was set at 0.05.

Patient and public involvement

Patients were not involved in the study based on anonymised data.

Results

In this study 916,619 individuals were included (women: 57.7%; mean age: 75.4 (standard deviation, SD: 7.4), and 853,085 (93.1%) of them met multimorbidity criteria (Figure 1).

Participants' characteristics are summarized in Table 1. Statistically significant differences were present between the multimorbidity and non-multimorbidity groups for all the variables included in the analysis (Table 1).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Among the 60 SNAC-K chronic diseases, the most prevalent were: hypertension (71.0%), dyslipidaemia (50.9%), osteoarthritis and other degenerative joint diseases (32.8%), obesity (28.7%), diabetes (25.1%), and anaemia (18.3%) (Table 2).

Eight multimorbidity patterns were identified using fuzzy c-means algorithm with fuzziness parameter of m=1.1, after computing different validation indices to obtain the optimal number of clusters (Additional File 1). This number was the same for the three different prevalence thresholds: no filtering, and ≥1% and ≥2% filters. The cluster formed by the most prevalent diseases was designated *Non-specified* (O/E ratio < 2 and exclusivity < 20). The remaining 7 clusters were specific: *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive; Mental, nervous, and digestive*; *Mental, digestive, and blood; Nervous, musculoskeletal, and circulatory*; and *Genitourinary, mental, and musculoskeletal* (Table 3). Table 3 shows the results, considering a 2% prevalence filter, for each pattern based on the fifteen diseases with the higher O/E-ratios.

Women were more represented than men in almost all clusters, from 52.7% for *Respiratory, circulatory, and neurological* to 83.6% for *Mental, nervous, and digestive.* The exception was *Genitourinary, mental, and musculoskeletal* in which men made up 90.9% due to the presence of male reproductive system diseases (Table 4).

The highest O/E ratio and exclusivity value were observed in *Nervous and digestive* for Parkinson, parkinsonism, and other neurological diseases (17.0% and 74.3%; and 15.9% and 69.4%, respectively). The lowest values were found in *Non-specified*. Clusters 1 to 3 presented the highest median number of visits with *Circulatory and digestive* being associated with the greatest number of visits over a one-year period (median 18 visits), and the *Non-specified* pattern presenting the lowest median number of visits which was equal to 5 (Table 4).

Multimorbidity patterns varied according to requirements for minimal prevalence of selected conditions in the population. As an example, Figure 2 depicts the composition of Cluster 1 according to prevalence levels of disease, and the other clusters are shown in Additional file 2.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Disease prevalence varied more greatly in the less populated patterns (e.g. *Non-specified*) (Additional File 2). Nevertheless, there was a group that remained in some clusters across all prevalence levels, for instance, some in *Neurological and digestive* (Parkinson and parkinsonism, other neurological diseases, chronic liver diseases, chronic pancreas, biliary tract, and gallbladder diseases) formed part of the cluster regardless of changes in cut-off prevalence (Additional File 2). The selected level of prevalence resulted in changes in O/E ratios, with some of them doubling their values.

Discussion

The soft clustering method we employed identified eight multimorbidity patterns, regardless of the prevalence selected. The *Non-specified* cluster included not only the largest number of individuals, but also those who presented the smallest multimorbidity prevalence. In this pattern diseases did not exhibit an association higher than chance because values of the O/E ratio and exclusivity were less than 2% and 20%, respectively. This suggests that such patients during their lives could change group. Two clusters presenting gender dominance were observed: *Nervous, musculoskeletal and circulatory* was predominately made up of women >70 years, while *Genitourinary, mental and musculoskeletal* was mostly formed of men of the same age. Such patterns represent 61% of the elderly participants included in the study. The rest had fewer individuals and some diseases were over-represented such as Parkinson and parkinsonism in *Nervous and digestive,* and asthma in *Respiratory, circulatory, and nervous.*

We observed that some diseases with O/E ratios ≥ 2 were consistently associated with each other as part of the same clusters (for instance, *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive*; and *Mental, nervous, and digestive*) regardless of the prevalence threshold that had been set. They can be considered core components of those

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

clusters. Further research is needed to establish the role of these conditions from a longitudinal perspective.

Comparison with the literature

Comparison with other studies is hindered by variations in methods, data sources and structures, populations, and diseases studied. Nevertheless, there are similarities with other authors. The non-specified pattern is the one most replicated in the literature, for example Prados et al who employed an exploratory factor analysis (25) and our group with k-means (24). Specifically, although the age range and the exclusivity threshold in our previous study were different, the hard clustering method provided clusters that overlap with some of the patterns obtained in this study, since both clustering results were predominantly defined by the O/E ratio (≥2) criteria. However, the soft approach allows a more flexible distribution of the individual and diseases.

Recent research has provided support for physio-pathological and genetic associations that explain the observed multimorbidity patterns. For instance, *Neurological and digestive* included chronic liver disease which has been linked to Parkinson through the accumulation of toxic substances in the brain (ammonia and manganese) and neuroinflammation (26). A higher risk of Parkinson among patients with chronic hepatitis C virus has also been reported (OR: 1.35) (27), in addition to associations between digestive diseases and neurodegenerative ones (e.g. Parkinson and Alzheimer) through the microbiome-gut-brain axis (27). A possible link between microbiota and digestive diseases such as chronic pancreatitis and pancreatic cancer has also been suggested (28,29). For the *Respiratory, circulatory, and neurological* cluster there is evidence of an association between chronic bronchial pathology, particularly asthma and obstructive pulmonary disease (COPD), and the risk of cardiovascular events (30). Longitudinal studies have observed an increased risk of developing Parkinson among individuals suffering from asthma and/or COPD (31,32). The association between asthma and allergy is known, and its coexistence defines a specific phenotype. For the *Circulatory and digestive c*luster, non-alcoholic fatty liver disease has been associated with the development of atrial fibrillation (33),

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

and hepatitis C infection with an increase in the risk of developing cardio- and cerebrovascular events (34). In addition, anaemia has been associated with advanced stages of chronic renal diseases and erythropoietin deficiency (35). Iron-deficiency anaemia has been associated with an increased risk of stroke (36) through thromboembolic phenomena secondary to reactive thrombocytosis. Chronic kidney disease produces auricle injuries (dilatation, fibrosis) and systemic inflammation, both of which can favour the onset and maintenance of atrial fibrillation (37).

Strengths and limitations

A major strength of this study is that it has employed a large, high-quality database made up of primary care records representative of the Catalan population aged ≥ 65 years (18). Patterns of multimorbidity have been studied based on the whole eligible sample. This approach is epidemiologically robust as the prevalence of diseases has been estimated on the whole sample rather than limited to patients with multimorbidity (2). Another strength is that individuals rather than diseases have been considered as the unit of analysis (8, 24). Such an approach permits a more realistic and rational monitoring of participants than cohort studies in order to analyse multimorbidity patterns along time. Moreover, the use of different prevalence cut-offs to obtain multimorbidity patterns has allowed the identification of nuclear diseases. We selected the higher prevalence (2%) because the patterns obtained had more clinical representativeness. The inclusion of all the potential diagnoses may have signified a greater complexity that would have hindered both the interpretation of findings and comparison with other studies.

Compared to hierarchical clustering, fuzzy c-means cluster analysis is less susceptible to: outliers in the data, choice of distance measure, and the inclusion of inappropriate or irrelevant variables (38). Nevertheless, some disadvantages of the method are that different solutions for each set of seed points can occur and there is no guarantee of optimal clustering (11). To minimize this shortcoming, we carried out 100 cluster realizations with different seeds to finally use the average result of all of them. In addition, the method is not efficient when a large

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

number of potential cluster solutions are to be considered (38). To address this limitation, we computed the optimal number of clusters using analytical indexes (Additional File 1).

Other limitations need to be taken into account. The dimensional reduction method performed in this work to reduce data complexity was PCAmix. Such methods can produce low percentages of variation on principal axes and make it difficult to choose the number of dimensions to retain. In order to decide on the most suitable number of dimensions we applied the Karlis-Saporta-Spinaki rule (27) which resulted in a 13-dimensional space for the 2% prevalence cut-off. Furthermore, the feasibility of developing clinical practice guidelines in accordance with these patterns might prove difficult due to the dimension of the diseases included in each pattern. Nonetheless, new clinical practice guidelines should consider the diseases that are overrepresented (O/E ratio≥2).

Implications for practice, policy, and research

Soft clustering methods offer a new methodological approach to understanding the relationships between specific diseases in individuals. This is an essential step in improving the care of patients and health systems. Analysing multimorbidity patterns permits the identification of patient subgroups with different associated diseases. Our analysis focuses on groups of patients as opposed to diseases. In this case, a disease is present in all patterns (clusters), but in different degrees. In this context, the observed/expected ratios (O/E-ratios) are used to measure which diseases are overrepresented in each cluster and to lead the clinical practice guidelines. The inclusion of varying cut-off points (prevalence filters) of the diseases that form the multimorbidity patterns allowed us to identify common nuclear diseases that remained independent from the prevalence that build such patterns.

It is noteworthy that 60% of the population ≥65 years was included in multimorbidity patterns made up of the most prevalent diseases. The rest of the population was grouped into five more specific patterns which permitted their better management.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Whilst clinical guidelines are currently aimed at covering the management of the diseases found in the *Non-specified* cluster, there is a lack of information regarding the associated diseases in the other patterns. The challenge will be to refocus healthcare policy from that based on individual diseases, with the accompanying consequences (increased risk of functional decline, poorer quality of life, greater use of services, polypharmacy, and increased mortality), to a multimorbidity orientation (39).

Further investigation on this topic is called for with particular focus on four major issues. First, the genetic study of these patterns will help the identification of risk subgroups. Second, research is needed on the life style and environmental factors (diet, physical exercise, toxics) associated with such patterns. Third, longitudinal studies should be performed to establish the onset order of the core diseases. Fourth, the characteristics of the diseases in the same cluster and their potential implication on the quality of primary care should be ascertained in greater detail.

Our findings suggest non-hierarchical cluster analysis identified multimorbidity patterns and phenotypes of certain sub-groups of patients that were more consistent with clinical practice.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Supplementary Data

Additional File 1. Extracting and validating multimorbidity patterns by applying the fuzzy c-means clustering algorithm and Computation of the observed/expected ratio and the exclusivity ratio.

Additional File 2. Composition of multimorbidity patterns according to disease levels of prevalence.

Footnotes

CVF and QFB contributed equally.

Contributors: All authors contributed to the design of the study, revised the article and approved the final version. CV, ARL and SFB obtained the funding. CV, QFB and SFB drafted the article. CV, QFB, SFB, MGC, MCB, FF, JMV and ARL contributed to the analysis and interpretation of data. CV, QFB and SFB wrote the first draft, and all authors contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript.

Funding: This work was supported by a research grant from the Carlos III Institute of Health, Ministry of Economy and Competitiveness (Spain), awarded on the 2016 call under the Health Strategy Action 2013-2016, within the National Research Program oriented to Societal Challenges, within the Technical, Scientific and Innovation Research National Plan 2013-2016 '[grant number PI16/00639]', co-funded with European Union ERDF funds (European Regional Development Fund) and Department of Health of the Catalan Government, in the call corresponding to 2017 for the granting of subsidies from the Strategic Plan for Research in Health (*Pla Estratègic de Recerca i Innovació en Salut*, PERIS) 2016-2020, modality research oriented to Primary care '[grant number SLT002/16/00058]' and from the Catalan Government '[grant number AGAUR 2017 SGR 578].

Disclaimer: The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service, the National Institute for Health Research or the National Department of Health.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Competing interests None declared.

Ethics approval: The protocol of the study was approved by the Committee on the Ethics of Clinical Research, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) (P16/151). All data were anonymized and the confidentiality of EHR was respected at all times in accordance with national and international law.

Data sharing statement: The datasets are not available because researchers have signed an agreement with the Information System for the Development of Research in Primary Care (SIDIAP) concerning confidentiality and security of the dataset that forbids providing data to third parties. This organisation is subject to periodic audits to ensure the validity and quality of the data.

Patient consent: Not required.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

References

1. Valderas Starfield B, Sibbald B, Salisbuty C, Roland M JM. Defining Comorbidity: Implications for Understanding Health and Health Services. Ann Fam Med 2009; 7:357–63.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, Salisbury C, Blom J, Freitag M, et al. Prevalence, determinants and patterns of multimorbidity in Primary Care: a systematic review of observational studies. PLOS One 2014; 21;9(7): e102149.

3. Salive ME. Multimorbidity in Older Adults. Epidemiol Rev 2013; 35:75-83.

4. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012; 380(9836):37-43.

5. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015; 386 (9995):743-800.

6. Gruneir A, Bronskill SE, Maxwell CJ, Bai YQ, Kone AJ, Thavorn K, et al. The association between multimorbidity and hospitalization is modified by individual demographics and physician continuity of care: a retrospective cohort study. BMC Health Serv Res 2016; 16:154.

7. Rocca WA, Boyd CM, Grossardt BR, Bobo WV, Finney Rutten LJ, Roger VL, et al. Prevalence of multimorbidity in a geographically defined American population: patterns by age, sex, and race/ethnicity. Mayo Clin Proc 2014; 89(10):1336-49.

8. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. J Clin Epidemiol 2014; 67(3):254-66.

9. Muth C, Blom JW, Smith SM, Johnell K, Gonzalez-Gonzalez AI, Nguyen TS, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. J Intern Med 2018; [Epub ahead of print]

10. Palmer K, Marengoni A, Forjaz MJ, Jureviciene E, Laatikainen T, Mammarella F, et al. Multimorbidity care model: Recommendations from the consensus meeting of the Joint Action on Chronic Diseases and Promoting Healthy Ageing across the Life Cycle (JA-CHRODIS). Health Policy 2018;122(1):4-11.

11.Wolfram. Fuzzy Clustering [Internet]. Available from: https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html

12. MathWorks. Fuzzy Clustering [Internet]. Available from: https://www.mathworks.com/help/fuzzy/fuzzy-clustering.html

13. France EF, Wyke S, Gunn JM, Mair FS, McLean G, Mercer SW. Multimorbidity in primary care: a systematic review of prospective cohort studies. Br J Gen Pract 2012; 62 (597): e297-307.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

14. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018; 47(5):1687-1704.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, Rodriguez-Blanco T, Pons-Vigués M, Pujol-Ribera E, et al. Burden of multimorbidity, socioeconomic status and use of health services across stages of life in urban areas: a cross-sectional study. BMC Public Health 2014;14(1):530.

16. Willadsen TG, Bebe A, Køster-Rasmussen R, Jarbøl DE, Guassora AD, Waldorff FB, et al. The role of diseases, risk factors and symptoms in the definition of multimorbidity – a systematic review. Scand J Prim Health Care 2016;34(2):112–21.

17. Xu X, Mishra GD, Jones M. Evidence on multimorbidity from definition to intervention: An overview of systematic reviews. Ageing Res Rev 2017; 7:53-68.

18. Del Mar García-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care 2012;19(3):135–45.

19. Calderón-Larrañaga A, Vetrano DL, Onder G, Gimeno-Feliu LA, Coscollar-Santaliestra C, Carfí A, et al. Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. J Gerontol A Biol Sci Med Sci 2017; 72 (10):1417-1423.

20. Domínguez-Berjón MF, Borrell C, Cano-Serral G, Esnaola S, Nolasco A, Pasarín MI, et al. Constructing a deprivation index based on census data in large Spanish cities (the MEDEA project)]. Gac Sanit 2008; 22(3):179-87.

21. Karlis D, Saporta G, Spinakis A. A simple rule for the selection of principal components. Commun Stat- Theory Methods 2003;32(3):643–66.

22. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2):191–203.

23. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

24. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018;19(1): 108.

25. Prados-Torres A, Poblador-Plou B, Calderón-Larrañaga A, Gimeno-Feliu LA, González-Rubio F, Poncel-Falcó A, et al. Multimorbidity Patterns in Primary Care: Interactions among Chronic Diseases Using Factor Analysis. PLoS One 2012; 7 (2): e32190.

26. Shin HW, Park HK. Recent Updates on Acquired Hepatocerebral Degeneration. Tremor Other Hyperkinet Mov (N Y) 2017;7:463.

27. Wijarnpreecha K, Chesdachai S, Jaruvongvanich V, Ungprasert P. Hepatitis C virus infection and risk of Parkinson's disease: A systematic review and meta-analysis. Eur J Gastroenterol Hepatol 2018;30(1):9–13.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

28. Westfall S, Lomis N, Kahouli I, Dia SY, Singh SP, Prakash S. Microbiome, probiotics and neurodegenerative diseases: deciphering the gut brain axis. Cell Mol Life Sci 2017; 74(20):3769–87.

29. Memba R, Duggan SN, Ni Chonchubhair HM, Griffin OM, Bashir Y, O'Connor DB, et al. The potential role of gut microbiota in pancreatic disease: A systematic review. Pancreatology 2017;17(6):867–74.

30. Xu M, Xu J, Yang X. Asthma and risk of cardiovascular disease or all-cause mortality: A meta-analysis. Ann Saudi Med 2017;37(2):99–105.

31. Cheng CM, Wu YH, Tsai SJ, Bai YM, Hsu JW, Huang KL, et al. Risk of developing Parkinson's disease among patients with asthma: A nationwide longitudinal study. Allergy 2015;70(12):1605–12.

32. Li CH, Chen WC, Liao WC, Tu CY, Lin CL, Sung FC, et al. The association between chronic obstructive pulmonary disease and Parkinson's disease: A nationwide population-based retrospective cohort study. Qjm. 2015;108(1):39–45.

33. Wijarnpreecha K, Boonpheng B, Thongprayoon C, Jaruvongvanich V, Ungprasert P. The association between non-alcoholic fatty liver disease and atrial fibrillation: A meta-analysis. Clin Res Hepatol Gastroenterol 2017 Oct;41(5):525-532.

34. Ambrosino P, Lupoli R, Di Minno A, Tarantino L, Spadarella G, Tarantino P, et al. The risk of coronary artery disease and cerebrovascular disease in patients with hepatitis C: A systematic review and meta-analysis. Int J Cardiol 2016;221:746-54.

35. Kepez A, Mutlu B, Degertekin M, Erol C. Association between left ventricular dysfunction, anemia, and chronic renal failure. Analysis of the Heart Failure Prevalence and Predictors in Turkey (HAPPY) cohort. Herz 2015;40(4):616–23.

36. Chang YL, Hung SH, Ling W, Lin HC, Li HC, Chung SD. Association between ischemic stroke and iron-deficiency anemia: a population-based study. PLoS One 2013;8(12):e82952.

37. Turakhia MP, Blankestijn PJ, Carrero JJ, Clase CM, Deo R, Herzog CA, et al. Chronic kidney disease and arrhythmias: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Eur Heart J 2018;39(24):2314–2325e.

38. Badsha MB, Mollah MN, Jahan N, Kurata H. Robust complementary hierarchical clustering for gene expression data analysis by β-divergence. J Biosci Bioeng 2013;116(3):397-407.

39. Yarnall AJ, Sayer AA, Clegg A, Rockwood K, Parker S, Hindle J V. New horizons in multimorbidity in older adults. Age Ageing 2017;46(6):882–8.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Table 1. Characteristics of study participants aged 65-94 years stratified by multimorbidity and non-multimorbidity (N= 916 619, Catalonia, 2012)

Variables*	Multimorbidity (n= 853 085)	Non-multimorbidity (n= 63 534)	All (N=916 619)
Sex, women, n (%)	496 294 (58.2)	32 837 (51.7)	529 131 (57.7)
Age, mean (SD)	75.6 (7.4)	73.2 (7.3)	75.4 (7.4)
Age (categories), n (%)			
[65,70)	225 514 (26.4)	26 664 (42.0)	252 178 (27.5)
[70,80)	370 356 (43.4)	24 230 (38.1)	394 586 (43.0)
[80,90)	224 143 (26.3)	10 601 (16.7)	234 744 (25.6)
≥90	33 072 (3.9)	2039 (3.2)	35 111 (3.8)
MEDEA index†			
Q1	130 894 (16.5)	13 897 (23.4)	144 791 (17.0)
Q2	126 537 (16.0)	9894 (16.6)	136 431 (16.0)
Q3	129 246 (16.3)	8976 (15.1)	138 222 (16.2)
Q4	125 322 (15.8)	7666 (12.9)	132 988 (15.6)
Q5	110 916 (14.0)	5967 (10.0)	116 883 (13.7)
Rural	169 190 (21.4)	13 059 (22.0)	182 249 (21.4)
Number of chronic diseases, median [IQR]	6.0 [4.0;8.0]	1.0 [0.0;1.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)			
0	0 (0.0)	25 380 (39.9)	25 380 (2.8)
1	0 (0.0)	38 154 (60.1)	38 154 (4.2)
[2, 5)	268 836 (31.5)	0 (0.0)	268 836 (29.3)
[5,10)	463 709 (54.4)	0 (0.0)	463 709 (50.6)
≥10	120 540 (14.1)	0 (0.0)	120 540 (13.2)
Number of drugs, median [IQR]	5.0 [3.0;8.0]	0.0 [0.0;1.0]	5.0 [2.0;8.0]
Number of drugs (categories):			
0	72 557 (8.5)	40 811 (64.2)	113 368 (12.4)
1	48 704 (5.7)	8378 (13.2)	57 082 (6.2)
[2, 5)	247 095 (29.0)	11 572 (18.2)	258 667 (28.2)
[5,10)	360 030 (42.2)	2651 (4.2)	362 681 (39.6)
≥10	124 699 (14.6)	122 (0.2)	124 821 (13.6)
Number of visits, median [IQR]	10.0 [6.0;17.0]	1.0 [0.0;4.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)			
0	24 543 (2.9)	23,402 (36.8)	47 945 (5.2)
1	24 281 (2.8%)	9603 (15.1%)	33 884 (3.7)
[2, 5)	114 198 (13.4%)	16 241 (25.6%)	130 439 (14.2%)
[5, 10)	239 181 (28.0%)	10 168 (16.0%)	249 349 (27.2%)
≥10	450 882 (52.9%)	4120 (6.5%)	455 002 (49.6%)

All comparisons between variables in multimorbidity and non-multimorbidity showed P<0.001
†MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 2. Prevalence of the 60 chronic diseases included in the study in individuals aged 65-94 years (N= 916 619, Catalonia, 2012). In three last columns, list of diseases included by prevalence cut off (1%, 2%, All)

Rank	*Chronic conditions*	*Frequency*	*Percentage (%)*	*All diseases included*	*1%*	*2%*
1	Hypertension	650 899	71.0			
2	Dyslipidaemia	466 585	50.9			
3	Osteoarthritis and other degenerative joint diseases	300 803	32.8			
4	Obesity	262 888	28.7			
5	Diabetes	230 460	25.1			
6	Anaemia	167 577	18.3			
7	Cataract and other lens diseases	156 622	17.1			
8	Chronic kidney diseases	153 756	16.8			
9	Prostate diseases	153 635	16.8			
10	Osteoporosis	151 847	16.6			
11	Depression and mood diseases	148 751	16.2			
12	Solid neoplasms	137 045	15.0			
13	Colitis and related diseases	131 512	14.4			
14	Venous and lymphatic diseases	126 997	13.9			
15	Other musculoskeletal and joint diseases	124 765	13.6			
16	Dorsopathies	124 603	13.6			
17	Neurotic, stress-related and somatoform diseases	123 395	13.5			
18	COPD, emphysema, chronic bronchitis	109 603	12.0			
19	Ischemic heart disease	95 434	10.4			
20	Deafness, hearing impairment	90 261	9.9			
21	Sleep disorders	88 739	9.7			
22	Thyroid diseases	88 445	9.7			
23	Other genitourinary diseases	85 468	9.3			
24	Cerebrovascular disease	80 264	8.8			
25	Atrial fibrillation	80 247	8.8			
26	Esophagus, stomach and duodenum diseases	80 043	8.7			
27	Heart failure	74 077	8.1			
28	Other eye diseases	68 939	7.5			
29	Glaucoma	66 162	7.2			
30	Inflammatory arthropathies	62 450	6.8			
31	Dementia	59 213	6.5			
32	Cardiac valve diseases	52 100	5.7			
33	Peripheral neuropathy	49 127	5.4			
34	Other psychiatric and behavioural diseases	46 841	5.1			
35	Asthma	43 663	4.8			
36	Allergy	40 394	4.4			
37	Autoimmune diseases	39 350	4.3			
38	Ear, nose, throat diseases	38 752	4.2			
39	Peripheral vascular disease	30 674	3.4			
40	Other neurological diseases	28 541	3.1			
41	Chronic pancreas, biliary tract and gallbladder diseases	27 321	3.0			
42	Migraine and facial pain syndromes	25 999	2.8			
43	Bradycardias and conduction diseases	25 476	2.8			
44	Chronic liver diseases	22 633	2.5			
45	Other digestive diseases	22 022	2.4			
46	Parkinson and parkinsonism	20 833	2.3			
47	Other metabolic diseases	18 997	2.1			
48	Other cardiovascular diseases	16 833	1.8			
49	Other skin diseases	15 363	1.7			
50	Chronic ulcer of the skin	13 869	1.5			
51	Blood and blood forming organ diseases	13 575	1.5			
52	Other respiratory diseases	9974	1.1			
53	Epilepsy	8981	1.0			
54	Haematological neoplasms	8174	0.9			
55	Chronic infectious diseases	6647	0.7			
56	Inflammatory bowel diseases	5549	0.6			
57	Schizophrenia and delusional diseases	4792	0.5			
58	Blindness, visual impairment	4772	0.5			
59	Multiple sclerosis	576	0.1			
60	Chromosomal abnormalities	77	0.0			

Abbreviations: COPD: Chronic obstructive Pulmonary Disease.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Table 3. Most frequent 15 diseases found in multimorbidity patterns in individuals aged 65-94 years (N= 916 619,

Pattern	Disease	O	O/E ratio	EX	Pattern	Disease	O	O/E ratio	EX
1 Nervous and digestive (n= 40 037)	Parkinson and parkinsonism	38.7	17.0	74.3	2 Respiratory, circulatory and nervous (n= 50 639)	Asthma	34.5	7.2	40.0
	Other neurological diseases	49.5	15.9	69.4		Peripheral vascular disease	13.9	4.2	22.9
	Chronic liver diseases	13.2	5.4	23.4		Parkinson and parkinsonism	8.5	3.8	20.8
	Chronic pancreas, biliary tract and gallbladder diseases	7.9	2.7	11.6		Other neurological diseases	11.7	3.7	20.7
	Dementia	14.7	2.3	9.9		COPD, emphysema, chronic bronchitis	31.0	2.6	14.3
	Other digestive diseases	4.8	2.0	8.7		Allergy	10.8	2.4	13.5
	Cerebrovascular disease	16.9	1.9	8.4		Heart failure	16.6	2.0	11.3
	Colitis and related diseases	24.1	1.7	7.3		Ischemic heart disease	21.1	2.0	11.2
	Other metabolic diseases	3.4	1.7	7.2		Other eye diseases	14.0	1.9	10.3
	Depression and mood diseases	25.0	1.5	6.7		Autoimmune diseases	7.2	1.7	9.3
	Anaemia	26.1	1.4	6.2		Other psychiatric and behavioural diseases	8.5	1.7	9.2
	Esophagus, stomach and duodenum diseases	11.3	1.3	5.6		Ear. nose. throat diseases	7.1	1.7	9.2
	Sleep disorders	12.4	1.3	5.6		Anaemia	30.4	1.7	9.2
	Other eye diseases	9.6	1.3	5.6		Peripheral neuropathy	8.8	1.6	9.1
	Dorsopathies	17.0	1.2	5.4		Cerebrovascular disease	14.3	1.6	9.0
3 Circulatory and digestive (n= 67 492)	Heart failure	51.4	6.4	46.9	4 Mental, nervous and digestive (n= 94 453)	Neurotic, stress-related and somatoform diseases	64.9	4.8	49.7
	Cardiac valve diseases	34.2	6.0	44.3		Depression and mood diseases	66.4	4.1	42.1
	Atrial fibrillation	47.3	5.4	39.8		Migraine and facial pain syndromes	8.2	2.9	29.6
	Bradycardias and conduction diseases	13.5	4.9	35.9		Sleep disorders	19.0	2.0	20.2
	Ischemic heart disease	33.7	3.2	23.8		Esophagus. stomach and duodenum diseases	14.9	1.7	17.6
	Chronic pancreas, biliary tract and gallbladder diseases	8.0	2.7	19.7		Osteoporosis	28.0	1.7	17.4
	Chronic liver diseases	6.1	2.5	18.2		Thyroid diseases	16.0	1.7	17.1
	Chronic kidney diseases	35.9	2.1	15.8		Colitis and related diseases	23.7	1.7	17.0
	Anemia	38.6	2.1	15.5		Other genitourinary diseases	14.4	1.5	15.9
	Cerebrovascular disease	18.3	2.1	15.4		Ear, nose, throat diseases	6.2	1.5	15.2
	COPD, emphysema, chronic bronchitis	23.6	2.0	14.5		Venous and lymphatic diseases	19.9	1.4	14.8
	Other digestive diseases	4.6	1.9	14.0		Allergy	6.1	1.4	14.3
	Peripheral vascular disease	6.1	1.8	13.3		Osteoarthritis and other degenerative joint diseases	45.0	1.4	14.1
	Other metabolic diseases	3.2	1.5	11.3		Dorsopathies	18.0	1.3	13.7
	Dementia	9.5	1.5	10.9		Cardiac valve diseases	7.4	1.3	13.5
5 Mental, digestive and blood (n= 106 845)	Dementia	21.8	3.4	39.4	6 Nervous, musculoskeletal and circulatory (n= 145 074)	Peripheral neuropathy	12.4	2.3	36.6
	Other digestive diseases	5.8	2.4	28.1		Other musculoskeletal and joint diseases	26.0	1.9	30.2
	Anemia	38.5	2.1	24.6		Venous and lymphatic diseases	26.4	1.9	30.2
	Chronic kidney diseases	33.3	2.0	23.1		Dorsopathies	25.3	1.9	29.4
	Colitis and related diseases	26.2	1.8	21.3		Obesity	51.0	1.8	28.2
	Cerebrovascular disease	14.8	1.7	19.7		Other genitourinary diseases	16.0	1.7	27.2
	Osteoporosis	26.0	1.6	18.3		Osteoarthritis and other degenerative joint diseases	55.0	1.7	26.5
	Cataract and other lens diseases	25.9	1.5	17.7		Osteoporosis	24.8	1.5	23.7
	Deafness. hearing impairment	14.0	1.4	16.5		Other eye diseases	10.7	1.4	22.4
	Venous and lymphatic diseases	19.5	1.4	16.4		Cataract and other lens diseases	22.5	1.3	20.8
	Osteoarthritis and other degenerative joint diseases	45.5	1.4	16.2		Thyroid diseases	12.6	1.3	20.7
	Depression and mood diseases	22.5	1.4	16.1		Glaucoma	9.2	1.3	20.1
	Other genitourinary diseases	12.3	1.3	15.4		Diabetes	31.3	1.2	19.7
	Other eye diseases	9.9	1.3	15.4		Ear, nose, throat diseases	5.2	1.2	19.5
	Sleep disorders	12.4	1.3	14.9		Dyslipidemia	62.7	1.2	19.5
7 Genitourinary, mental and musculoskeletal (n=173 746)	Prostate diseases	54.7	3.3	61.8	8 Non-specified (n=238 333)	Dyslipidemia	38.4	0.8	19.6
	Other psychiatric and behavioural diseases	11.1	2.2	41.2		Thyroid diseases	7.3	0.8	19.6
	Inflammatory arthropathies	12.4	1.8	34.5		Osteoporosis	12.2	0.7	19.2
	COPD, emphysema, chronic bronchitis	20.5	1.7	32.5		Hypertension	47.6	0.7	17.4
	Solid neoplasms	21.8	1.5	27.7		Glaucoma	4.4	0.6	16.0
	Peripheral vascular disease	4.7	1.4	26.7		Solid neoplasms	9.1	0.6	15.7
	Ischemic heart disease	13.7	1.3	25.0		Migraine and facial pain syndromes	1.7	0.6	15.7
	Diabetes	31.8	1.3	24.0		Autoimmune diseases	2.2	0.5	13.4
	Ear, nose, throat diseases	5.3	1.3	23.7		Other metabolic diseases	1.1	0.5	13.3
	Deafness, hearing impairment	11.6	1.2	22.3		Allergy	2.2	0.5	13.0
	Allergy	4.8	1.1	20.5		Chronic liver diseases	1.2	0.5	12.8
	Hypertension	75.8	1.1	20.2		Other genitourinary diseases	4.5	0.5	12.7
	Glaucoma	7.5	1.0	19.6		Esophagus, stomach and duodenum diseases	4.1	0.5	12.2
	Autoimmune diseases	4.4	1.0	19.4		Other psychiatric and behavioral diseases	2.4	0.5	12.0
	Obesity	29.0	1.0	19.2		Diabetes	10.8	0.4	11.2

Abbreviations: O: Disease prevalence in the cluster; O/E ratio: observed/expected ratio; Ex: exclusivity; COPD: Chronic obstructive Pulmonary Disease.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 4. Variables characterizing each cluster in baseline study for 2% prevalence cut-off point (N= 916 619)

	1.Nervous and digestive	2. Respiratory, circulator and nervous	3. Circulatory and digestive	4. Mental, nervous and digestive	5. Mental, digestive and blood	6. Nervous, musculoskeletal and circulatory	7. Genitourinary, mental and musculoskeletal	8. Non-specified	All
Number of people, n	40 037	50 639	67 492	94 453	106 845	145 07	173 746	238 333	916 619
Multimorbidity, n (%)	39 776 (99.3)	50 513 (99.8)	67 443 (99.9)	94 442 (100.0)	106 696 (99.9)	144 86 (99)	171 983 (99.0)	177 363 (74.4)	853 085 (93.1)
Polypharmacy, n (%)	28 484 (71.1)	38 869 (76.8)	54 658 (81.0)	64 154 (67.9)	71 830 (67.2)	86 317 (59.)	90 603 (52.1)	52 588 (22.1)	487 502 (53.1)
Women, n (%)	22 628 (56.5)	26 690 (52.7)	38 023 (56.3)	78 922 (83.6)	85 735 (80.2)	113 62 (0)	15 730 (9.1)	147 773 (62.0)	529 131 (57.7)
Men, n (%)	17 409 (43.5)	23 949 (47.3)	29 469 (43.7)	15 531 (16.4)	21 110 (19.8)	31 445 ( )	158 016 (90.9)	90 560 (38.0)	387 488 (42.3)
Age (categories), n (%)									
[65,70)	7188 (18.0)	10 400 (20.5)	7233 (10.7)	28 305 (30.0)	12 036 (11.3)	38 829 ( )	52 003 (29.9)	96 184 (40.4)	252 178 (27.5)
[70,80)	17 804 (44.5)	22 743 (44.9)	24 724 (36.6)	40 577 (43.0)	33 624 (31.5)	70 643 ( )	84 037 (48.4)	100 435 (42.1)	394 586 (43.0)
[80,90)	13 460 (33.6)	15 568 (30.7)	29 908 (44.3)	22 638 (24.0)	48 453 (45.3)	32 714 ( )	34 785 (20.0)	37 217 (15.6)	234 744 (25.6)
[90,99]	1587 (4.0)	1927 (3.8)	5628 (8.3)	2934 (3.1)	12 732 (11.9)	2888 ( )	2920 (1.7)	4497 (1.9)	35 111 (3.8)
MEDEA* index									
R	7831 (21.8)	9300 (20.2)	13 718 (23.2)	17 266 (19.7)	22 183 (23.0)	27 40 ( )	35 145 (21.5)	49 405 (21.9)	182249 (21.4)
U1	6010 (16.7)	6890 (15.0)	9537 (16.1)	15 027 (17.2)	16 556 (17.2)	19 599 ( )	25 656 (15.7)	45 516 (20.2)	144791 (17.0)
U2	5690 (15.8)	7134 (15.5)	9140 (15.4)	14 335 (16.4)	15 272 (15.8)	21 379 ( )	25 951 (15.9)	37 530 (16.6)	136431 (16.0)
U3	5941 (16.5)	7520 (16.4)	9187 (15.5)	14 223 (16.3)	15 421 (16.0)	23 26 ( )	26 908 (16.5)	35 761 (15.8)	138222 (16.2)
U4	5540 (15.4)	7686 (16.7)	9016 (15.2)	14 012 (16.0)	14 272 (14.8)	23 780 ( )	26 526 (16.2)	32 157 (14.2)	132988 (15.6)
U5	4982 (13.8)	7421 (16.2)	8638 (14.6)	12 652 (14.5)	12 699 (13.2)	21 923 ( )	23 064 (14.1)	25 506 (11.3)	116883 (13.7)
Number of chronic diseases, median [IQR]	8.0 [6.0;10.0]	8.0 [6.0;10.0]	8.0 [7.0;11.0]	7.0 [6.0;9.0]	7.0 [5.0;9.0]	6.0 [5]	5.0 [4.0;7.0]	3.0 [3.0;4.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)									
0	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.0%)	0 (0.0)	235 (0.1)	25 144 (10.5)	25 380 (2.8)
1	262 (0.7)	125 (0.2)	49 (0.1)	11.0 (0.0)	149 (0.1)	204 (0.)	1528 (0.9)	35 826 (15.0)	38 154 (4.2)
[2, 5)	5409 (13.5	4507 (8.9)	4275 (6.3)	8781 (9.3)	14 601 (13.7)	22 400 (15.)	57 561 (33.1)	151 302 (63.5)	268 836 (29.3)
[5,10)	23 502 (58.7)	30 257 (59.8)	37 910 (56.2)	62 490 (66.2)	73 427 (68.7)	105 62 (72.)	104 915 (60.4)	25 588 (10.7)	463 709 (50.6)
≥10	10 864 (27.1)	15 749 (31.1)	25 259 (37.4)	231 715 (24.5)	18 668 (17.5)	16 850 (11.6)	9506 (5.5)	473 (0.2)	120 540 (13.2)
Number of drugs, median [IQR]	7.0 [4.0;9.0]	7.0 [5.0;10.0]	8.0 [5.0;11.0]	6.0 [4.0;9.0]	6.0 [4.0;9.0]	5.0 [3. ;8.0]	5.0 [3.0;7.0]	2.0 [0.0;4.0]	5.0 [2.0;8.0]
Number of drugs (categories)									
0	2576 (6.4)	2491 (4.9)	3349 (5.0)	5636 (6.0)	7,037 (6.6)	8330 (.7)	13 389 (7.7)	70 561 (29.6)	113 368 (12.4)
1	1212 (3.0)	1072 (2.1)	1015 (1.5)	2939 (3.1)	3390 (3.2)	6772 (.7)	11 440 (6.6)	29 242 (12.3)	57 082 (6.2)
[2, 5)	7766 (19.4)	8207 (16.2)	8471 (12.6)	21 725 (23.0)	24 587 (23.0)	43 656 (30.)	58 314 (33.6)	85 942 (36.1)	258 667 (28.2)
[5,10)	18 510 (46.2)	23 597 (46.6)	31 850 (47.2)	46 022 (48.7)	52 653 (49.3)	68 193 (47.)	73 694 (42.4)	48 161 (20.2)	362 681 (39.6)
≥10	9973 (24.9)	15 272 (30.2)	22 808 (33.8)	18 132 (19.2)	19 177 (17.9)	18 123 (12.)	16 909 (9.7)	4427 (1.9)	124 821 (13.6)
Number of visits 2012, median [IQR]	12.0 [7.0;20.0]	14.0 [8.0;22.0]	18.0 [9.0;30.0]	11.0 [6.0;19.0]	12.0 [7.0;19.0]	11.0 [.0;17.]	9.0 [5.0;15.0]	5.0 [2.0;9.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)									
0	976 (2.4)	871 (1.7)	1143 (1.7)	2219 (2.3)	2515 (2.4)	2410.3 (1.7)	4137 (2.4)	33 673 (14.1)	47 945 (5.2)
1	874 (2.2)	754 (1.5)	929 (1.4)	2055 (2.2)	2238 (2.1)	2412.4 (1.7)	4685 (2.7)	19 938 (8.4)	33 884 (3.7)
[2, 5)	4000 (10.0)	3918 (7.7)	4329 (6.4)	10 589 (11.2)	11 018 (10.3)	14943.7 (10.)	24 319 (14.0)	57 322 (24.1)	130 439 (14.2)
[5, 10)	9158 (22.9)	10 774 (21.3)	10 883 (16.1)	24 504 (25.9)	27 003 (25.3)	42180.7 (29.)	54 212 (31.2)	70 634 (29.6)	249 349 (27.2)
≥10	25 030 (62.5)	34 322 (67.8)	50 209 (74.4)	55 085 (58.3)	64 071 (60.0)	83126.5 (57.)	86 393 (49.7)	56 766 (23.8)	455 002 (49.6)

For the sake of simplicity, all numbers in the table were rounded to its closest natural number. *MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Figure 1. Study population flow chart

*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

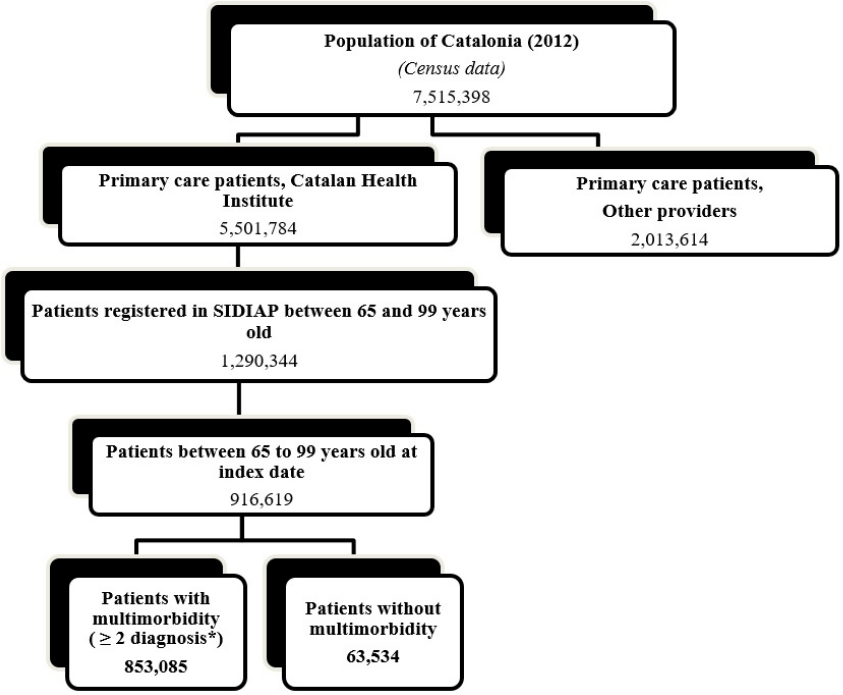


Figure 1. Study population flow chart
*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

217x161mm (115 x 115 DPI)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

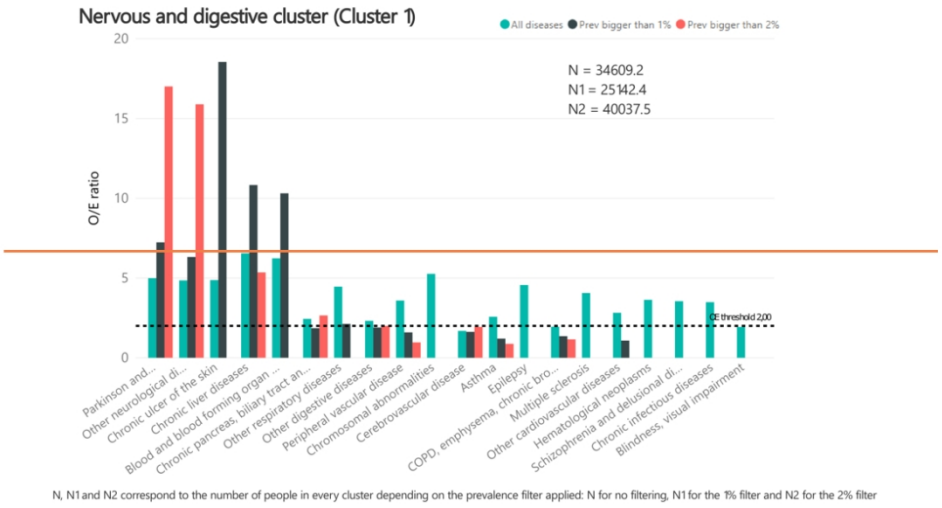


Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Additional File 1

A) Extracting and Validating Multimorbidity Patterns by applying the Fuzzy C Means Clustering algorithm.

In this annex we present a description of the procedure followed to obtain a set of multimorbidity patterns characterizing a patient population aged 65 or more in Catalonia (Spain).

Dataset dimension reduction.

The initial dataset was composed on 31st December, 2012, of a registered active diagnosis with a certain prevalence value, out of 60 possible diseases for the N=916,619 patients included in the study. Additionally, considering age and the gender, each patient was initially characterized by a vector of 62 features, most of which were binary variables indicating the presence/absence of a disease at the end of 2012. For most of the study, diseases with prevalence ≥2% were filtered, resulting in 47 diseases and the corresponding 49 features (adding age and gender). Since most of the selected features were categorical instead of quantitative, the dataset was a mixture of numerical and categorical variables. We processed this dataset by applying a mixture of the well-known Principal Component Analysis (PCA) to the numeric original features and a Multiple Correspondence Analysis (MCA) to the binary ones, in order to obtain a new dataset of reduced dimension. We selected the PCAmix algorithm, as described by Chavent et al, to perform the dimensionality reduction. It follows the criterion based on concentrating most of the variability of the new transformed features, that is to say, variance of the data in the low-dimensional representation were maximized. The Karlis-Saporta-Spinaki rule was followed to select the first 13 dimensions out of the 49 for the 2% prevalence filtering, according to the eigenvalues of the PCAmix and the number of features and individuals in the dataset. As a result, after the PCAmix transformation and the extraction of the optimal number of dimensions, the new dataset was composed of N=916,619 vectors of $d = 13$ features each one. In the following we denote this new dataset as $\boldsymbol{\mathcal{Y}} \coloneqq \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$, denoting by $\boldsymbol{y}_2 \in \mathbb{R}^{13}$ for $n = 1, \dots, N$ the new vector representing patient n.

Soft clustering algorithm

Once the transformed dataset $\boldsymbol{\mathcal{Y}}$ was computed, a soft clustering algorithm was applied to fuzzily distribute the population into a set of clusters, corresponding to the different multimorbidity patterns. In a traditional clustering procedure patients are grouped in an exclusive way, so that if a certain patient belongs to a definite cluster then s/he cannot be included in another one. In contrast, an overlapping clustering, such as the Fuzzy C Means (FCM) algorithm, uses fuzzy sets to cluster patients, so that each patient belongs to all clusters with different degrees of

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

membership. The choice between a hard or a soft clustering algorithm is traditionally made based on the application and the performance obtained. In our case, the use of the FCM algorithm presented performance results similar to those of the hard clustering algorithm Kmeans, but clinically more solid. It was, therefore, chosen as the most appropriate method for the description of the multimorbidity patterns.

FCM was originally introduced by Bezdek and yields an unsupervised form of grouping in which individuals can belong to more than one cluster. To do so, they are associated with an appropriate set of K membership values, where K denotes the number of clusters. The parameters that determine the clustering process are a set of K centroids $\mathbf{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_K\}$ where $v_k \in \mathbb{R}^{13}$ for $k = 1, \dots, K$ and a set of membership factors $\mathbf{U} = \{u_{jn}; j = 1, \dots, K; n = 1, \dots, N\}$ with $0 \leq u_{jn} \leq 1$. Factor u_{jn} indicates the degree to which individual n^{th} belongs to cluster j^{th}. Both centroids $\mathbf{V}$ and membership factors $\mathbf{U}$ are obtained by iteratively minimizing the objective function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$, which is the weighted sum of squared errors within clusters

$$J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) = \sum_{n=1}^{N} \sum_{j=1}^{K} (u_{jn})^m \left\| \boldsymbol{y}_n - \boldsymbol{v}_j \right\|^2; \quad 1 < m < \infty \quad (1)$$

Thus, the similarity between an individual and a cluster centroid is measured through the squared error between the vector associated with the patient and the centroid prototyping the cluster. The fuzziness weighting parameter m, is selected to adjust the blending of the different clusters and it is any real number greater than 1. High m values would produce a fuzzy set of clusters so that individuals would tend to be equally distributed across clusters, whereas lower ones would generate a non-overlapped set of clusters. The FCM method iteratively alternates between computing the centroids in $\mathbf{V}$ as the average of the individual's features in $\boldsymbol{\mathcal{Y}}$ previously weighted by the correspondent membership factors and estimating the membership factors in $\mathbf{U}$ in order to maximize the cost function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$ given the updated centroids in $\mathbf{V}$. In our work, we randomly initialized the set of centroids $\mathbf{V}$ and halted the iterative process when $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) < \epsilon$, where $0 < \epsilon \ll 1$. This procedure converges to a local minimum or saddle point of $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$.

Cluster stability validation.

Stable clusters are required in order to characterize multimorbidity patterns, consequently we applied 100 FCM independent runs to the transformed dataset $\boldsymbol{\mathcal{Y}}$ and averaged both the membership factors and the centroid vectors, after ordering the clusters in descending order in terms of the summation of memberships to clusters, measured as $\sum_{n=1}^{N} (u_{jn})^m$. This is equivalent to selecting the centroid and membership factors associated with the cluster with more population in each run and averaging them. Then after removing the selected cluster from each set, the procedure is repeated until a final set of clusters, composed of the K averaged

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

centroids and the corresponding averaged membership factors, is obtained. In this averaging process we previously verified the similarity between the averaged parameters by a heuristic inspection of some randomly selected run results

Number of clusters and fuzziness parameter validation.

Since clustering algorithms are unsupervised, machine-learning techniques, the model fitting the dataset is traditionally computed through cost functions that depend on both the dataset and the clustering parameters and are denoted as validation indices. We computed three different well-known validation indices to obtain the optimal number of clusters K and the optimal value of the fuzziness parameter m: the partition coefficient validation index whose cost function is maximum for the optimal model, the Xie-Beni, and the partition entropy validation indices whose cost functions are minimum for the optimal models. A cross-validation technique was applied using a split sample approach, by randomly dividing the individuals into two different datasets, a first (50%) training dataset used for obtaining the averaged FCM clusters, and a second (50%) test dataset used to verify the model fitting the data.

This validation procedure was applied to the set of clusters obtained after the previously explained averaging process, with the 2% prevalence filtering and considering 49 features before PCAmix reduction. We checked $m = 1.1, 1.2,$ and 1.5 and $K = 5,..,20$. In Figure1 the performance obtained through the three validation indices is depicted. The best behaviour is obtained for m=1.1 and as is shown in Figure 2 and Figure 3 we can conclude that the optimal number of clusters for m=1.1 ranges from 6 to 12, validated with both the training dataset and the test dataset (more details are given in figures).

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

B) Computation of the observed/expected ratio and the exclusivity ratio.

The observed/expected $(O/E)_{dj}$ ratio and the exclusivity ratio EX_{dj} have been used in this work in order to decide whether a disease d is overrepresented or not in any given cluster j.

The $(O/E)_{dj}$ ratio was calculated by dividing disease prevalence in the cluster O_{dj} by disease prevalence in the overall population E_d. As membership of an individual n in a cluster j was denoted by a membership degree factor u_{nj}, and not as a binary variable, the observed disease prevalence O_{dj} in a cluster j was computed as the ratio between the summation of the membership degree factors corresponding to all individuals suffering the disease d and the summation of all the membership degree factors corresponding to the cluster j. Let us assume that there are n_d individuals suffering the disease d and that they are grouped in the set I_d, then the observed prevalence was computed as

$$O_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{\sum_{n=1}^{N} u_{nj}}$$

while the expected prevalence was computed as

$$E_d = \frac{n_d}{N}$$

Exclusivity ratio EX_{dj}, defined as the proportion of individuals with the disease d included in the cluster j over the total number of individuals with the disease n_d, was computed as

$$EX_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{n_d}$$

References

1. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate analysis of mixed data: The PCAmixdata R package. 2014; eprint arXiv:1411.4911.

2. Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press; 1981.

3. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

4. Pal NR, Bezdek JC. On Cluster Validity for the Fuzzy c-Means Model. IEEE Trans Fuzzy Syst 1995;3(3):370–9.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

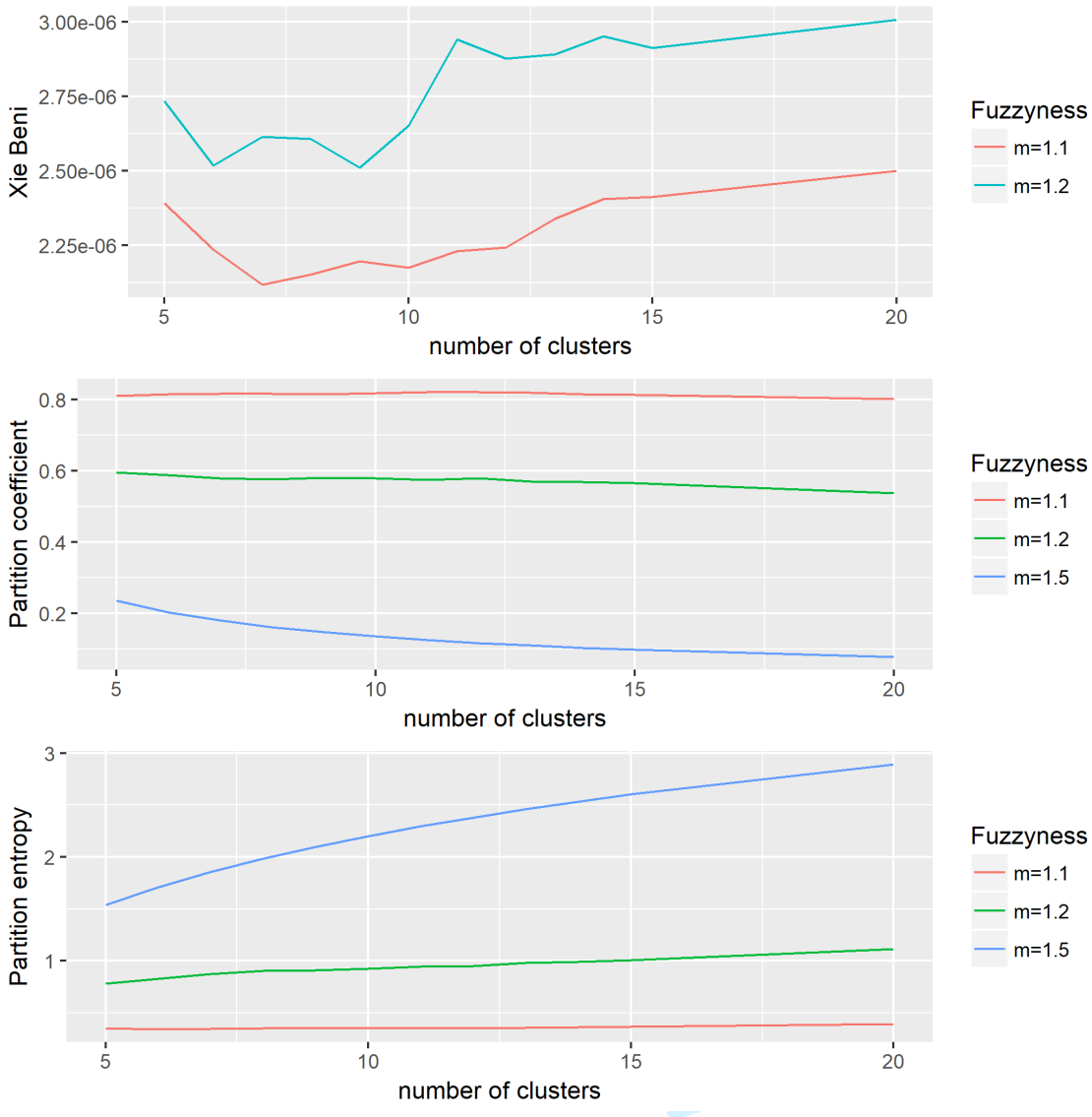


Figure 1. Selection of the optimal m parameter

Index m = 1.5 was also computed for Xie-Beni indices, but not included in the graph because the curve is significantly higher than the other two in the plot.

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. For this reason, all plots are showing that m = 1.1 is the best parameter to optimize all the computed indices.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

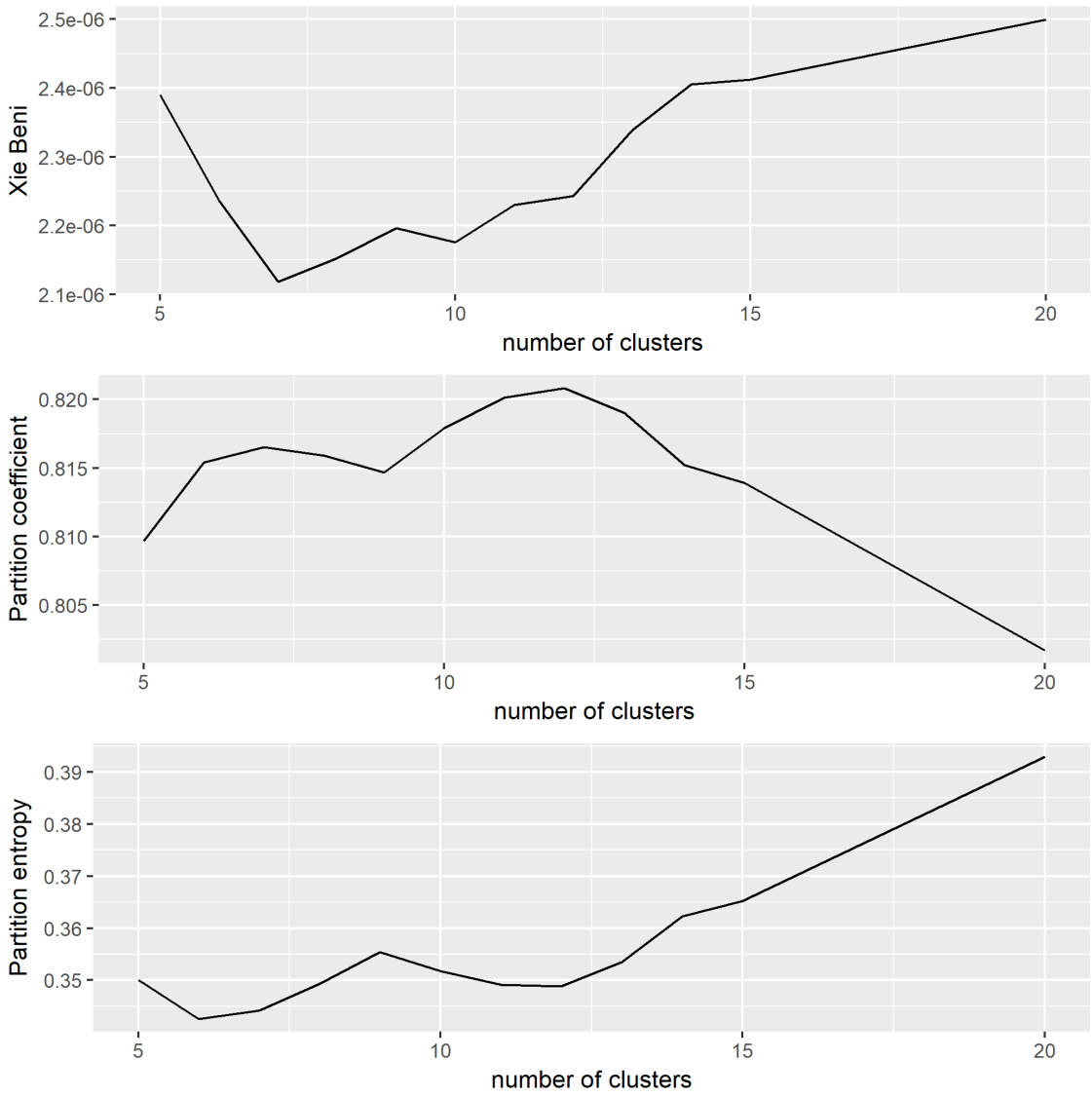


Figure 2. Selection of the optimal number of clusters (m = 1.1)

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. Within the plots above, optimal values are located in the range from 6 to 12 clusters.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

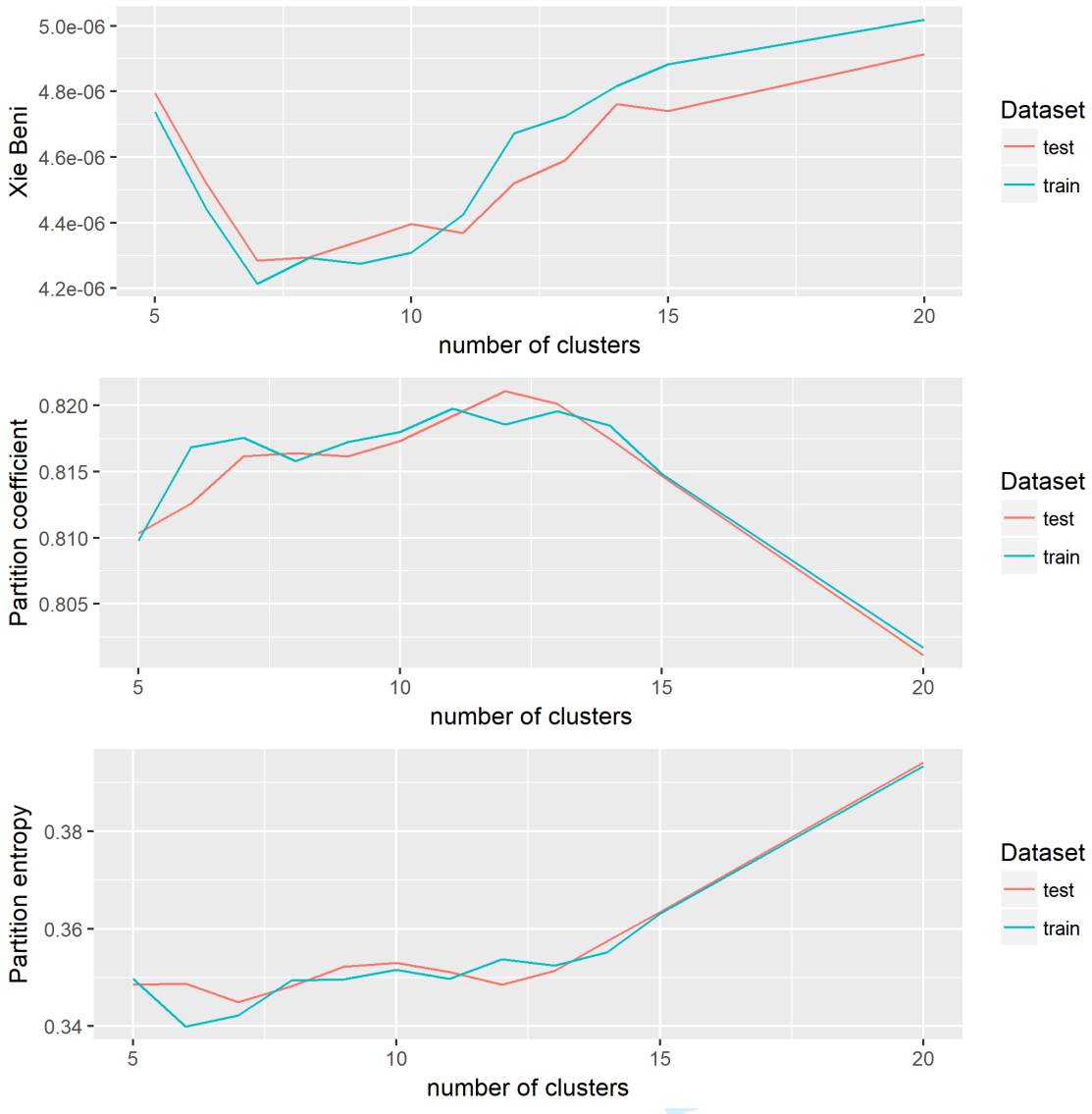
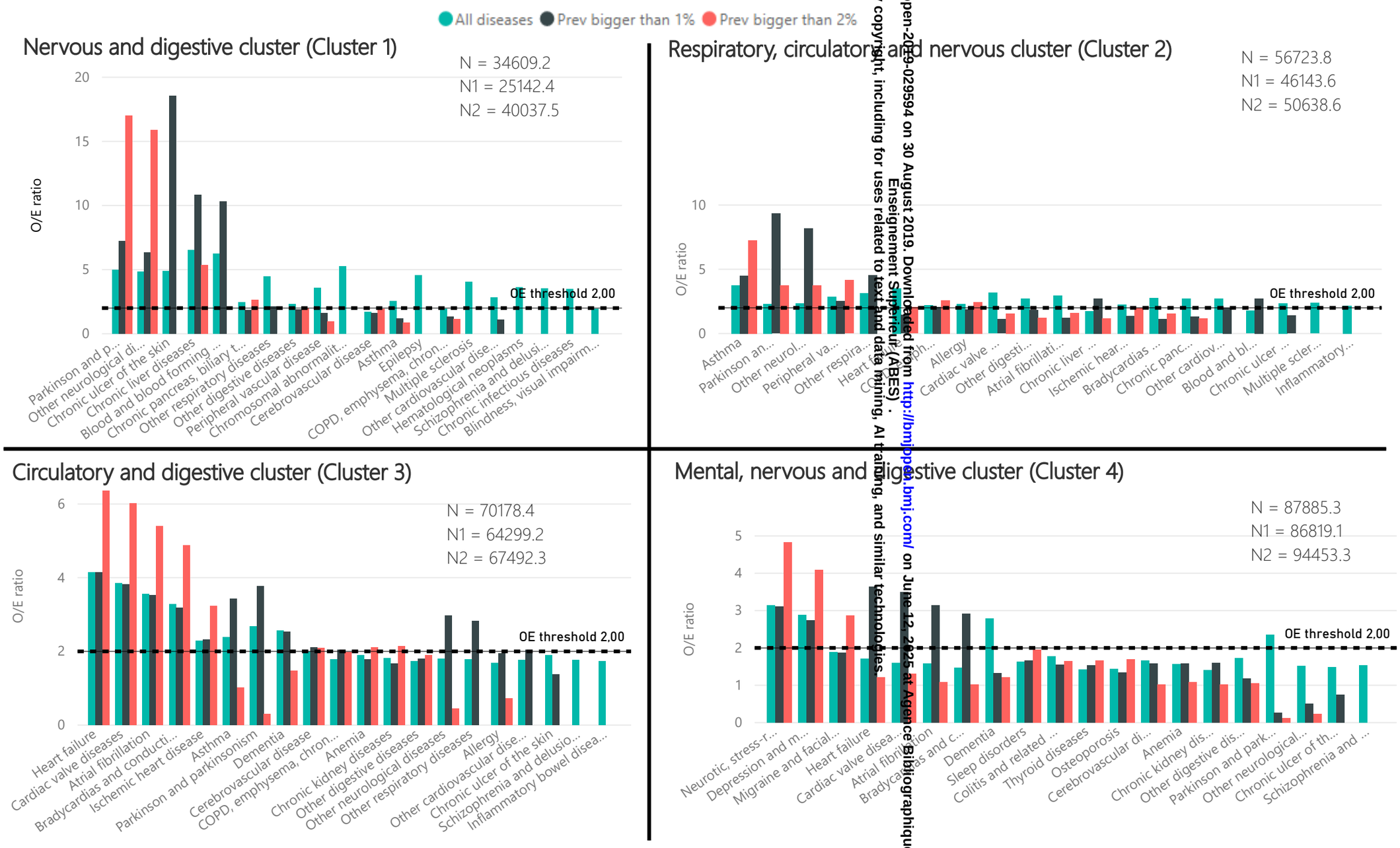


Figure 3. Cross-validation of the clustering with m = 1.1

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. In the plots above we can find the optimal values in the range from 6 to 12 clusters. Additionally, no significant variation is registered in the indices regardless of the dataset selection.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml



N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied: N for no filtering, N1 for the 1% filter and N2 for the 2% filter

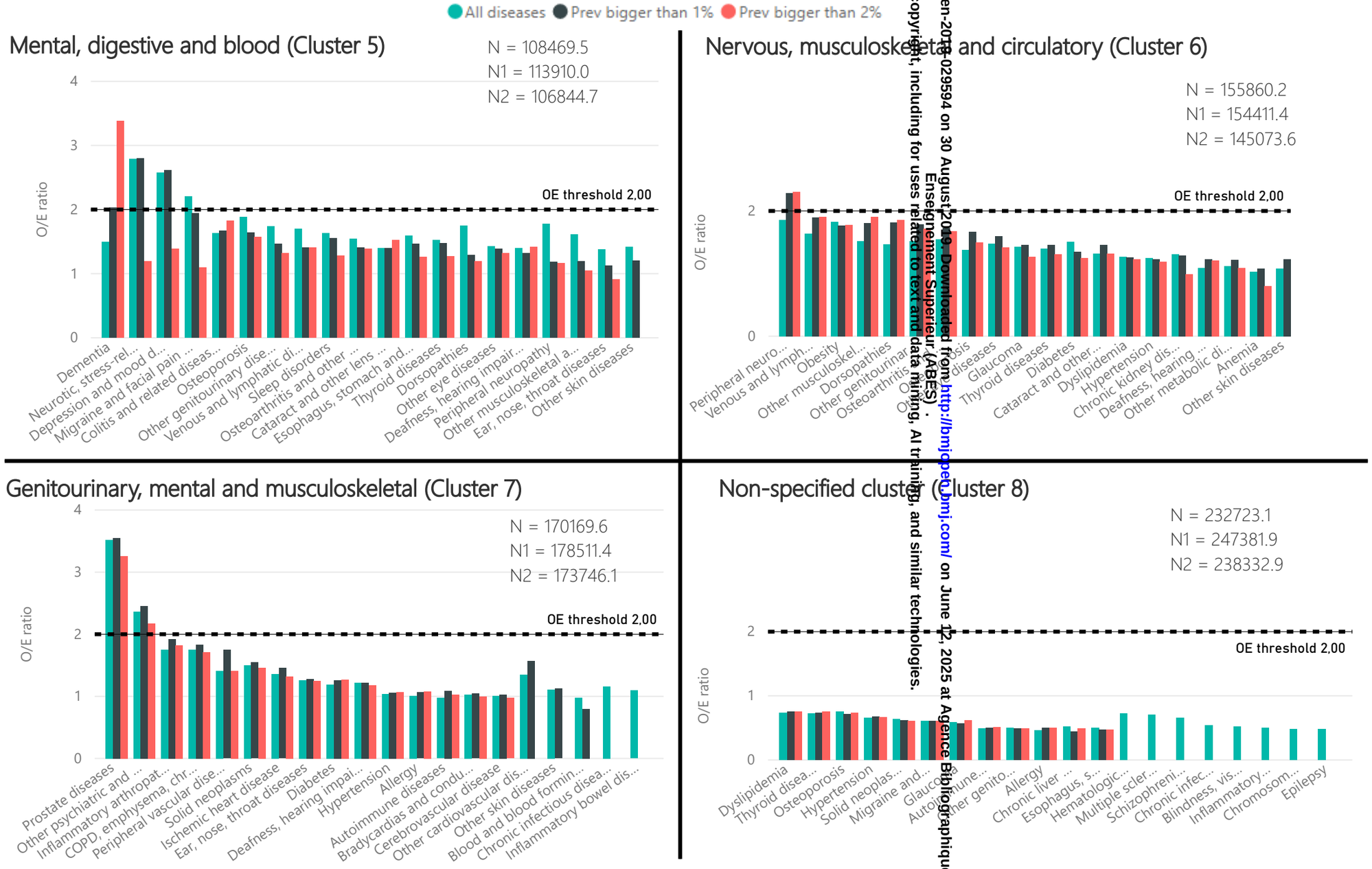
For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Supérieur (ABES) .

http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l

136/bmjopen-2019-029594 on 30 August 2019. Downloaded from



N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied: N for no filtering, N1 for the 1% filter and N2 for the 2% filter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

136/bmjopen-2018-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES) . cted by copyright, including for uses related to text and data mining, AI training, and similar technologies.

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	5
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*6*
Bias	9	Describe any efforts to address potential sources of bias	7
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	6
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	6
		(*b*) Describe any methods used to examine subgroups and interactions	6
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	
		(*e*) Describe any sensitivity analyses	7
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	Figure 1
		(c) Consider use of a flow diagram	Figure 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 1
		(b) Indicate number of participants with missing data for each variable of interest	Tables
Outcome data	15*	Report numbers of outcome events or summary measures	8-9

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	8-9 Tables
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Additional File 1
Discussion			
Key results	18	Summarise key results with reference to study objectives	10
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	12
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	12
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	14

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

BMJ Open

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-029594.R2
Article Type:	Research
Date Submitted by the Author:	24-Jul-2019
Complete List of Authors:	Violan-Fors, Concepción; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Foguet-Boreu, Quintí; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol) Universitat Autònoma de Barcelona, ; Hospital de Campdevànol, Emergency room Fernández-Bertolín, Sergio; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Guisado-Clavero, Marina; Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) Cabrera-Bean, Margarita; Universitat Politecnica de Catalunya, Signal Theory and Comunications Department Formiga, F; Hospital Universitari de Bellvitge Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol),
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, General practice / Family practice
Keywords:	Chronic conditions, Multimorbidity, Cluster analysis, EPIDEMIOLOGY

SCHOLARONE™
Manuscripts

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population

1. Concepción Violán-Fors*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

2. Quintí Foguet-Boreu*. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Department of Psychiatry, Vic University Hospital. Francesc Pla el Vigatà, 1, 08500 Vic, Barcelona, Spain. 4. Department of Basic and Methodological Sciences. Faculty of Health Sciences and Welfare. University of Vic-Central University of Catalonia (UVic-UCC)
E-mail: 42292qfb@comb.cat

3. Sergio Fernández-Bertolín. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: sfernandez@idiapjgol.org

4. Marina Guisado-Clavero. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: marina.guisado@gmail.com

5. Margarita Cabrera-Bean. Signal Theory and Comunications Department, Universitat Politècnica de Catalunya, Barcelona Tech. Campus Nord, UPC D5, Jordi Girona 1-2, 08034-Barcelona, Spain.
E-mail: marga.cabrera@upc.edu

6. Francesc Formiga. Internal Medicine Service, Hospital Universitari de Bellvitge, Hospitalet del Llobregat, Barcelona, Catalonia, Spain.
E-mail: fformiga@bellvitgehospital.cat

7. Jose M Valderas. Health Services & Policy Research Group, Academic Collaboration for Primary Care, University of Exeter Medical School, Exeter, EX1 2LU, United Kingdom.
E-mail: J.M.Valderas@exeter.ac.uk

8. Albert Roso-Llorach. 1. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

Corresponding author: Concepción Violán. IDIAPJGol
Quintí Foguet-Boreu. IDIAPJGol
Gran Via Corts Catalanes, 587 àtic.08007 Barcelona. Spain.
Telephone: 0034 93 482 41 24. FAX: 0034 93 482 41 74.
Web page: www.idiapjgol.org.E-mail: cviolan@idiapjgol.org; 42292qfb@comb.cat
Word count: 3 164

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Abstract

Objectives The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied. Fuzzy cluster analysis allows individuals to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.

Design A cross-sectional study was conducted based on data from electronic health records

Setting 284 primary health care centres in Catalonia, Spain (2012).

Participants 916 619 eligible individuals were included (women: 57.7%).

Primary and secondary outcome measures We extracted data on demographics, ICD-10 chronic diagnoses, prescribed drugs, and socioeconomic status for patients aged ≥65. Following principal component analysis of categorical and continuous variables (PCAmix) for dimensionality reduction, machine learning techniques were applied for the identification of disease clusters in a fuzzy c-means analysis. Sensitivity analyses, with different prevalence cut-off points for chronic diseases, were also conducted. Solutions were evaluated from clinical consistency and significance criteria.

Results Multimorbidity was present in 93.1%. Eight clusters were identified with a varying number of disease values: *Nervous and digestive*; *Respiratory, circulatory, and nervous*; *Circulatory, and digestive*; *Mental, nervous, and digestive, female dominant*; *Mental, digestive, and blood, female oldest-old dominant*; *Nervous, musculoskeletal, and circulatory, female dominant*; *Genitourinary, mental, and musculoskeletal, male dominant*; and *Non-specified, youngest-old dominant*. Nuclear diseases were identified for each cluster independently of the prevalence cut-off point considered.

Conclusions Multimorbidity patterns were obtained using fuzzy c-means cluster analysis. They are clinically meaningful clusters which support the development of tailored approaches to multimorbidity management and further research.

Keywords: Chronic conditions; Multimorbidity; Epidemiology; Cluster analysis.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

Strengths and limitations of this study

- Studies focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns (hard clustering forces each individual to belong to a single cluster, whereas soft clustering allows elements to be simultaneously classified into multiple cluster).
- Reliable and valid identification of disease clusters is needed for the development of evidence-based clinical practice guidelines and pathways of care for patients that correspond to the wide spectrum of diseases in patients with multimorbidity.
- Soft clustering analysis allows for diseases to be linked simultaneously to multiple clusters and is more consistent with clinical experience than other approaches frequently found in the literature.
- The different cut-off points (prevalence filters) applied to obtain multimorbidity patterns permitted the identification of common nuclear diseases which remained independent of their prevalence.
- The literature provides support for the etiopathophysiological and epidemiological associations between conditions forming part of the same cluster.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Introduction

The term multimorbidity widely refers to the existence of numerous medical conditions in a single individual (1). In many regions of the world there is evidence that a substantial, and probably growing, proportion of the adult population is affected by multiple chronic conditions. Moreover, the association of multimorbidity with increasing age leading to a two-fold prevalence in the final decades of life has been proven (2). Multimorbidity has been estimated to be at around 62% between 65 and 74 years, and around 81.5% after 85 years (3). Its true extent is, however, difficult to gauge as there is no agreed definition or classification system (4-7).

Most of the published literature focusses on diseases rather than individuals as the unit of analysis in assessing multimorbidity patterns (8). Orienting the analysis of multimorbidity patterns at an individual level, and not of disease, could have crucial implications for patients. In the current context of limited evidence on interventions for unselected patients with multimorbidity, such an approach would allow better understanding of population groups, and facilitate the development and implementation of strategies aimed at prevention, diagnosis, treatment, and prognosis. It would also elicit essential information for the development of clinical guidelines, pathways of care, and lead to better understanding of the nature and range of the required health services (9,10).

Cluster analysis involves assigning individuals so that the items (diseases) in the same cluster are as similar as possible, while individuals belonging to different clusters are as dissimilar as possible. The identification of clusters is based on similarity measures and their choice may depend on the data or the purpose of the analysis (11,12). Hard clustering forces each element to belong to a single cluster, whereas soft clustering (also referred to as fuzzy clustering) allows elements to be simultaneously classified into multiple clusters.

Empirical evidence is needed on how both established and novel techniques influence the identification of multimorbidity patterns. A recent systematic review recommended that future epidemiological studies cover a broad selection of health conditions in order to avoid missing

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

potentially key nosological associations and enhance external validity. When many conditions are considered, the clustering of individuals based on morbidity data will encounter high-dimensional issues. This is particularly important when a clustering-based approach is adopted to assess the impact of multimorbidity on individual health outcomes and health service uses (2, 8, 13-15).

The identification of multimorbidity patterns seems to be implicitly dependent on the prevalence of the included diseases (2,8,16,17). However, to the best of our knowledge no previous study has analysed the identification of multimorbidity patterns explicitly based on the prevalence of the diseases.

The aim of this study was to identify, with soft clustering methods, multimorbidity patterns in the electronic health records of a population ≥65 years, and to analyse such patterns in accordance with the different prevalence cut-off points applied.

Methods

Study population

A cross-sectional analysis was carried out in Catalonia (Spain), a Mediterranean region of 7,515,398 inhabitants (2012). The Catalan Health Institute provides universal coverage and operates 284 primary health care centres (PHC).

Data sources

Since 2006 the Information System for Research in Primary Care (SIDIAP) database includes anonymized longitudinal electronic health records from primary and secondary care which gather information on demographics, diagnoses, prescriptions, and socioeconomic status (18). In our study the inclusion criteria were individuals aged 65-99 years on 31st December 2011

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

with at least one PHC visit since 2012. Only participants that survived until 31st December 2012 (index date) were included in the analysis.

Variables

Diseases were coded in the SIDIAP using the International Classification of Diseases version 10 (ICD-10). An operational definition of multimorbidity was the simultaneous presence of more than one of the selected 60 chronic diseases previously identified by the Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (19).

Additional variables included in the study were sociodemographics (age, sex, socio-economic status (MEDEA index) (20), clinical variables (including number of chronic diseases and invoiced drugs), and use of health services (number of visits to family physicians, nurses, and emergency services).

Statistical analysis

Descriptive statistics were used to summarize overall information. Disease prevalence was computed for all the included population. Descriptive analyses were stratified by the presence of multimorbidity. Comparison was performed using t-Student or Mann-Whitney for continuous variables and Chi-Square for categorical ones.

In order to obtain the most representative clusters all patients were included irrespective of whether they presented multimorbidity or not. Sex and age variables, together with chronic diseases selected by prevalence, were included in the analysis. The number of features to be considered varied from the 62 original ones (no prevalence filtering applied) to 54 and 49, for a 1% and 2% prevalence threshold, respectively.

Due to the large number of diseases, a principal component analysis for categorical and continuous data (PCAmix) was implemented to reduce complexity. With this technique both continuous and dichotomous variables were simultaneously processed through the application of Multi Correspondence Analysis to the binary variables and PCA to the continuous ones.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Using Karlis-Saporta-Spinaki criterion to select the optimal number of dimensions to retain, the dataset of 49 features per individual per 2% prevalence cut-off was transformed to a new dimensionally reduced dataset of 13 continuous features per individual, which concentrated most of the variability of the newly transformed dataset (21).

Once the transformed dataset was obtained, clusters of chronic conditions at baseline were identified using the fuzzy c-means clustering algorithm (22). This machine learning technique forces every individual to belong to every cluster in accordance with its characteristics and by assigning a membership degree factor in (0,1) to each individual with respect to each pattern. This provides the flexibility enabling patients to belong to more than one multimorbidity pattern (23).

The main parameters in this clustering procedure were the number of clusters and a fuzziness parameter, denoted *m*, that ranged from just above 1 to infinity. High *m* values produce a fuzzy set of clusters, so that individuals are equally distributed across clusters, whereas lower ones generate non-overlapped clusters. Further details on the stability and validation techniques applied to obtain the best fuzzy c-means parameters and the set of centroids, are presented in Additional File 1.

To describe the multimorbidity patterns, frequencies and percentages of diseases (P) in each cluster were calculated. Observed/expected ratios (O/E-ratios) were calculated by dividing disease prevalence in the cluster by disease prevalence in the overall population. As the membership of each individual to any of the clusters was given by a membership degree factor, and not as a binary variable, the observed disease prevalence (O) in a cluster was computed as the sum of the disease membership degree factors corresponding to all individuals suffering the disease. Exclusivity, defined as the proportion of patients with the disease included in the cluster over the total number of patients with the disease, was also calculated. Further details on how these ratios were computed using the membership factors are given in Additional File 1. A disease was considered to be part of a multimorbidity cluster when O/E-ratio was ≥2 or

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

exclusivity value ≥25% (24). Clusters names were also defined taking into account the dominant gender or age in the cluster compared to the overall sample distribution.

We conducted a sensitivity analysis by modifying the prevalence threshold for disease inclusion in the cluster analysis. For chronic diseases we considered as alternatives no filtering, and ≥1% and ≥2% filters among the included population. In order to conform to the Karlis-Saporta-Spinaki rule, a different number of dimensions of the transformed dataset were retained to construct the clusters for every prevalence cut-off: 13 dimensions for the 2% prevalence, 14 dimensions for the 1% prevalence, and 17 dimensions with no filtering. The content of each cluster was compared across filtering approaches in terms of diseases associated with that cluster, characteristics of the included population, and cluster size. Clinical evaluation of the consistency and significance of these solutions was also conducted.

The analyses were carried out using R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). The significance level was set at 0.05.

Patient and public involvement

Patients were not involved in the study based on anonymised data.

Results

In this study 916,619 individuals were included (women: 57.7%; mean age: 75.4 (standard deviation, SD: 7.4), and 853,085 (93.1%) of them met multimorbidity criteria (Figure 1).

Participants' characteristics are summarized in Table 1. Statistically significant differences were present between the multimorbidity and non-multimorbidity groups for all the variables included in the analysis (Table 1).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Among the 60 SNAC-K chronic diseases, the most prevalent were: hypertension (71.0%), dyslipidaemia (50.9%), osteoarthritis and other degenerative joint diseases (32.8%), obesity (28.7%), diabetes (25.1%), and anaemia (18.3%) (Table 2).

Eight multimorbidity patterns were identified using fuzzy c-means algorithm with fuzziness parameter of m=1.1, after computing different validation indices to obtain the optimal number of clusters (Additional File 1). This number was the same for the three different prevalence thresholds: no filtering, and ≥1% and ≥2% filters. The cluster formed by the most prevalent diseases was designated *Non-specified, youngest-old dominant* (O/E ratio < 2 and exclusivity < 20). The remaining 7 clusters were specific: *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive; Mental, nervous, and digestive, female dominant*; *Mental, digestive, and blood, female oldest-old dominant; Nervous, musculoskeletal, and circulatory, female dominant*; and *Genitourinary, mental, and musculoskeletal, male dominant* (Table 3). Table 3 shows the results, considering a 2% prevalence filter, for each pattern based on the fifteen diseases with the higher O/E-ratios.

Women were more represented than men in almost all clusters, from 52.7% for *Respiratory, circulatory, and neurological* to 83.6% for *Mental, nervous, and digestive, female dominant.* The exception was *Genitourinary, mental, and musculoskeletal, male dominant* in which men made up 90.9% due to the presence of male reproductive system diseases (Table 4).

The highest O/E ratio and exclusivity value were observed in *Nervous and digestive* for Parkinson, parkinsonism, and other neurological diseases (17.0% and 74.3%; and 15.9% and 69.4%, respectively). The lowest values were found in *Non-specified, youngest-old dominant*. Clusters 1 to 3 presented the highest median number of visits with *Circulatory and digestive* being associated with the greatest number of visits over a one-year period (median 18 visits), and the *Non-specified, youngest-old dominant* pattern presenting the lowest median number of visits which was equal to 5 (Table 4). Additional File 2 shows tables of variables characterizing each cluster in baseline study for 1% and for no prevalence cut-off points.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Multimorbidity patterns varied according to requirements for minimal prevalence of selected conditions in the population. As an example, Figure 2 depicts the composition of Cluster 1 according to prevalence levels of disease, and the other clusters are shown in Additional file 3. Disease prevalence varied more greatly in the less populated patterns (e.g. *Non-specified, youngest-old dominant*) (Additional File 3). Nevertheless, there was a group that remained in some clusters across all prevalence levels, for instance, some in *Neurological and digestive* (Parkinson and parkinsonism, other neurological diseases, chronic liver diseases, chronic pancreas, biliary tract, and gallbladder diseases) formed part of the cluster regardless of changes in cut-off prevalence (Additional File 3). The selected level of prevalence resulted in changes in O/E ratios, with some of them doubling their values.

Discussion

The soft clustering method we employed identified eight multimorbidity patterns, regardless of the prevalence selected. The *Non-specified, youngest-old dominant* cluster included not only the largest number of individuals, but also those who presented the smallest multimorbidity prevalence. In this pattern diseases did not exhibit an association higher than chance because values of the O/E ratio and exclusivity were less than 2% and 20%, respectively. This suggests that such patients during their lives could change group. Two clusters presenting gender dominance were observed: *Nervous, musculoskeletal and circulatory, female dominant* was predominately made up of women >70 years, while *Genitourinary, mental and musculoskeletal, male dominant* was mostly formed of men of the same age. Such patterns represent 61% of the elderly participants included in the study. The rest had fewer individuals and some diseases were over-represented such as Parkinson and parkinsonism in *Nervous and digestive,* and asthma in *Respiratory, circulatory, and nervous*.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

We observed that some diseases with O/E ratios ≥ 2 were consistently associated with each other as part of the same clusters (for instance, *Nervous and digestive*; *Respiratory, circulatory, and nervous; Circulatory and digestive*; and *Mental, nervous, and digestive, female dominant*) regardless of the prevalence threshold that had been set. They can be considered core components of those clusters. Further research is needed to establish the role of these conditions from a longitudinal perspective.

Comparison with the literature

Comparison with other studies is hindered by variations in methods, data sources and structures, populations, and diseases studied. Nevertheless, there are similarities with other authors. The non-specified pattern is the one most replicated in the literature, for example Prados et al who employed an exploratory factor analysis (25) and our group with k-means (24). Specifically, although the age range and the exclusivity threshold in our previous study were different, the hard clustering method provided clusters that overlap with some of the patterns obtained in this study, since both clustering results were predominantly defined by the O/E ratio (≥2) criteria. However, the soft approach allows a more flexible distribution of the individual and diseases.

Recent research has provided support for physio-pathological and genetic associations that explain the observed multimorbidity patterns. For instance, *Neurological and digestive* included chronic liver disease which has been linked to Parkinson through the accumulation of toxic substances in the brain (ammonia and manganese) and neuroinflammation (26). A higher risk of Parkinson among patients with chronic hepatitis C virus has also been reported (OR: 1.35) (27), in addition to associations between digestive diseases and neurodegenerative ones (e.g. Parkinson and Alzheimer) through the microbiome-gut-brain axis (27). A possible link between microbiota and digestive diseases such as chronic pancreatitis and pancreatic cancer has also been suggested (28,29). For the *Respiratory, circulatory, and neurological* cluster there is evidence of an association between chronic bronchial pathology, particularly asthma and obstructive pulmonary disease (COPD), and the risk of cardiovascular events (30). Longitudinal

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

studies have observed an increased risk of developing Parkinson among individuals suffering from asthma and/or COPD (31,32). The association between asthma and allergy is known, and its coexistence defines a specific phenotype. For the *Circulatory and digestive c*luster, non-alcoholic fatty liver disease has been associated with the development of atrial fibrillation (33), and hepatitis C infection with an increase in the risk of developing cardio- and cerebrovascular events (34). In addition, anaemia has been associated with advanced stages of chronic renal diseases and erythropoietin deficiency (35). Iron-deficiency anaemia has been associated with an increased risk of stroke (36) through thromboembolic phenomena secondary to reactive thrombocytosis. Chronic kidney disease produces auricle injuries (dilatation, fibrosis) and systemic inflammation, both of which can favour the onset and maintenance of atrial fibrillation (37).

Strengths and limitations

A major strength of this study is that it has employed a large, high-quality database made up of primary care records representative of the Catalan population aged ≥ 65 years (18). Patterns of multimorbidity have been studied based on the whole eligible sample. This approach is epidemiologically robust as the prevalence of diseases has been estimated on the whole sample rather than limited to patients with multimorbidity (2). Another strength is that individuals rather than diseases have been considered as the unit of analysis (8, 24). Such an approach permits a more realistic and rational monitoring of participants than cohort studies in order to analyse multimorbidity patterns along time. Moreover, the use of different prevalence cut-offs to obtain multimorbidity patterns has allowed the identification of nuclear diseases. We selected the higher prevalence (2%) because the patterns obtained had more clinical representativeness. The inclusion of all the potential diagnoses may have signified a greater complexity that would have hindered both the interpretation of findings and comparison with other studies.

Compared to hierarchical clustering, fuzzy c-means cluster analysis is less susceptible to: outliers in the data, choice of distance measure, and the inclusion of inappropriate or irrelevant variables (38). Nevertheless, some disadvantages of the method are that different solutions for

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

each set of seed points can occur and there is no guarantee of optimal clustering (11). To minimize this shortcoming, we carried out 100 cluster realizations with different seeds to finally use the average result of all of them. In addition, the method is not efficient when a large number of potential cluster solutions are to be considered (38). To address this limitation, we computed the optimal number of clusters using analytical indexes (Additional File 1).

Other limitations need to be taken into account. The dimensional reduction method performed in this work to reduce data complexity was PCAmix. Such methods can produce low percentages of variation on principal axes and make it difficult to choose the number of dimensions to retain. In order to decide on the most suitable number of dimensions we applied the Karlis-Saporta-Spinaki rule (27) which resulted in a 13-dimensional space for the 2% prevalence cut-off. Furthermore, the feasibility of developing clinical practice guidelines in accordance with these patterns might prove difficult due to the dimension of the diseases included in each pattern. Nonetheless, new clinical practice guidelines should consider the diseases that are overrepresented (O/E ratio≥2).

Implications for practice, policy, and research

Soft clustering methods offer a new methodological approach to understanding the relationships between specific diseases in individuals. This is an essential step in improving the care of patients and health systems. Analysing multimorbidity patterns permits the identification of patient subgroups with different associated diseases. Our analysis focuses on groups of patients as opposed to diseases. In this case, a disease is present in all patterns (clusters), but in different degrees. In this context, the observed/expected ratios (O/E-ratios) are used to measure which diseases are overrepresented in each cluster and to lead the clinical practice guidelines. The inclusion of varying cut-off points (prevalence filters) of the diseases that form the multimorbidity patterns allowed us to identify common nuclear diseases that remained independent from the prevalence that build such patterns.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

It is noteworthy that 60% of the population ≥65 years was included in multimorbidity patterns made up of the most prevalent diseases. The rest of the population was grouped into five more specific patterns which permitted their better management.

Whilst clinical guidelines are currently aimed at covering the management of the diseases found in the *Non-specified, youngest-old dominant* cluster, there is a lack of information regarding the associated diseases in the other patterns. The challenge will be to refocus healthcare policy from that based on individual diseases, with the accompanying consequences (increased risk of functional decline, poorer quality of life, greater use of services, polypharmacy, and increased mortality), to a multimorbidity orientation (39).

Further investigation on this topic is called for with particular focus on five major issues. First, the genetic study of these patterns will help the identification of risk subgroups. Second, research is needed on the life style and environmental factors (diet, physical exercise, toxics) associated with such patterns. Third, longitudinal studies should be performed to establish the onset order of the core diseases. Fourth, alternative approaches to handle covariates in cluster analysis should be addressed in future analysis plan. Recently, a new method that allows the covariates to be incorporated into the membership factor to model individual probabilities of cluster membership has been proposed (40). And fifth, the characteristics of the diseases in the same cluster and their potential implication on the quality of primary care should be ascertained in greater detail.

Our findings suggest non-hierarchical cluster analysis identified multimorbidity patterns and phenotypes of certain sub-groups of patients that were more consistent with clinical practice.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Supplementary Data

Additional File 1. Extracting and validating multimorbidity patterns by applying the fuzzy c-means clustering algorithm and Computation of the observed/expected ratio and the exclusivity ratio.

Additional File 2. Variables characterizing each cluster in baseline study for 1% and for no prevalence cut-off points.

Additional File 3. Composition of multimorbidity patterns according to disease levels of prevalence.

Footnotes

CVF and QFB contributed equally.

Contributors: All authors contributed to the design of the study, revised the article and approved the final version. CV, ARL and SFB obtained the funding. CV, QFB and SFB drafted the article. CV, QFB, SFB, MGC, MCB, FF, JMV and ARL contributed to the analysis and interpretation of data. CV, QFB and SFB wrote the first draft, and all authors contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript.

Funding: This work was supported by a research grant from the Carlos III Institute of Health, Ministry of Economy and Competitiveness (Spain), awarded on the 2016 call under the Health Strategy Action 2013-2016, within the National Research Program oriented to Societal Challenges, within the Technical, Scientific and Innovation Research National Plan 2013-2016 '[grant number PI16/00639]', co-funded with European Union ERDF funds (European Regional Development Fund) and Department of Health of the Catalan Government, in the call corresponding to 2017 for the granting of subsidies from the Strategic Plan for Research in Health (*Pla Estratègic de Recerca i Innovació en Salut*, PERIS) 2016-2020, modality research oriented to Primary care '[grant number SLT002/16/00058]' and from the Catalan Government '[grant number AGAUR 2017 SGR 578].

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Disclaimer: The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service, the National Institute for Health Research or the National Department of Health.

Competing interests None declared.

Ethics approval: The protocol of the study was approved by the Committee on the Ethics of Clinical Research, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol) (P16/151). All data were anonymized and the confidentiality of EHR was respected at all times in accordance with national and international law.

Data sharing statement: The datasets are not available because researchers have signed an agreement with the Information System for the Development of Research in Primary Care (SIDIAP) concerning confidentiality and security of the dataset that forbids providing data to third parties. This organisation is subject to periodic audits to ensure the validity and quality of the data.

Patient consent: Not required.

For Peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

References

1. Valderas Starfield B, Sibbald B, Salisbuty C, Roland M JM. Defining Comorbidity: Implications for Understanding Health and Health Services. Ann Fam Med 2009; 7:357–63.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, Salisbury C, Blom J, Freitag M, et al. Prevalence, determinants and patterns of multimorbidity in Primary Care: a systematic review of observational studies. PLOS One 2014; 21;9(7): e102149.

3. Salive ME. Multimorbidity in Older Adults. Epidemiol Rev 2013; 35:75-83.

4. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. Lancet. 2012; 380(9836):37-43.

5. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015; 386 (9995):743-800.

6. Gruneir A, Bronskill SE, Maxwell CJ, Bai YQ, Kone AJ, Thavorn K, et al. The association between multimorbidity and hospitalization is modified by individual demographics and physician continuity of care: a retrospective cohort study. BMC Health Serv Res 2016; 16:154.

7. Rocca WA, Boyd CM, Grossardt BR, Bobo WV, Finney Rutten LJ, Roger VL, et al. Prevalence of multimorbidity in a geographically defined American population: patterns by age, sex, and race/ethnicity. Mayo Clin Proc 2014; 89(10):1336-49.

8. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. J Clin Epidemiol 2014; 67(3):254-66.

9. Muth C, Blom JW, Smith SM, Johnell K, Gonzalez-Gonzalez AI, Nguyen TS, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. J Intern Med 2018; [Epub ahead of print]

10. Palmer K, Marengoni A, Forjaz MJ, Jureviciene E, Laatikainen T, Mammarella F, et al. Multimorbidity care model: Recommendations from the consensus meeting of the Joint Action on Chronic Diseases and Promoting Healthy Ageing across the Life Cycle (JA-CHRODIS). Health Policy 2018;122(1):4-11.

11.Wolfram. Fuzzy Clustering [Internet]. Available from: https://reference.wolfram.com/legacy/applications/fuzzylogic/Manual/12.html

12. MathWorks. Fuzzy Clustering [Internet]. Available from: https://www.mathworks.com/help/fuzzy/fuzzy-clustering.html

13. France EF, Wyke S, Gunn JM, Mair FS, McLean G, Mercer SW. Multimorbidity in primary care: a systematic review of prospective cohort studies. Br J Gen Pract 2012; 62 (597): e297-307.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

14. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018; 47(5):1687-1704.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, Rodriguez-Blanco T, Pons-Vigués M, Pujol-Ribera E, et al. Burden of multimorbidity, socioeconomic status and use of health services across stages of life in urban areas: a cross-sectional study. BMC Public Health 2014;14(1):530.

16. Willadsen TG, Bebe A, Køster-Rasmussen R, Jarbøl DE, Guassora AD, Waldorff FB, et al. The role of diseases, risk factors and symptoms in the definition of multimorbidity – a systematic review. Scand J Prim Health Care 2016;34(2):112–21.

17. Xu X, Mishra GD, Jones M. Evidence on multimorbidity from definition to intervention: An overview of systematic reviews. Ageing Res Rev 2017; 7:53-68.

18. Del Mar García-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care 2012;19(3):135–45.

19. Calderón-Larrañaga A, Vetrano DL, Onder G, Gimeno-Feliu LA, Coscollar-Santaliestra C, Carfí A, et al. Assessing and Measuring Chronic Multimorbidity in the Older Population: A Proposal for Its Operationalization. J Gerontol A Biol Sci Med Sci 2017; 72 (10):1417-1423.

20. Domínguez-Berjón MF, Borrell C, Cano-Serral G, Esnaola S, Nolasco A, Pasarín MI, et al. Constructing a deprivation index based on census data in large Spanish cities (the MEDEA project)]. Gac Sanit 2008; 22(3):179-87.

21. Karlis D, Saporta G, Spinakis A. A simple rule for the selection of principal components. Commun Stat- Theory Methods 2003;32(3):643–66.

22. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Comput Geosci 1984;10(2):191–203.

23. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

24. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018;19(1): 108.

25. Prados-Torres A, Poblador-Plou B, Calderón-Larrañaga A, Gimeno-Feliu LA, González-Rubio F, Poncel-Falcó A, et al. Multimorbidity Patterns in Primary Care: Interactions among Chronic Diseases Using Factor Analysis. PLoS One 2012; 7 (2): e32190.

26. Shin HW, Park HK. Recent Updates on Acquired Hepatocerebral Degeneration. Tremor Other Hyperkinet Mov (N Y) 2017;7:463.

27. Wijarnpreecha K, Chesdachai S, Jaruvongvanich V, Ungprasert P. Hepatitis C virus infection and risk of Parkinson's disease: A systematic review and meta-analysis. Eur J Gastroenterol Hepatol 2018;30(1):9–13.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Enseignement Superieur (ABES) .

28. Westfall S, Lomis N, Kahouli I, Dia SY, Singh SP, Prakash S. Microbiome, probiotics and neurodegenerative diseases: deciphering the gut brain axis. Cell Mol Life Sci 2017; 74(20):3769–87.

29. Memba R, Duggan SN, Ni Chonchubhair HM, Griffin OM, Bashir Y, O'Connor DB, et al. The potential role of gut microbiota in pancreatic disease: A systematic review. Pancreatology 2017;17(6):867–74.

30. Xu M, Xu J, Yang X. Asthma and risk of cardiovascular disease or all-cause mortality: A meta-analysis. Ann Saudi Med 2017;37(2):99–105.

31. Cheng CM, Wu YH, Tsai SJ, Bai YM, Hsu JW, Huang KL, et al. Risk of developing Parkinson's disease among patients with asthma: A nationwide longitudinal study. Allergy 2015;70(12):1605–12.

32. Li CH, Chen WC, Liao WC, Tu CY, Lin CL, Sung FC, et al. The association between chronic obstructive pulmonary disease and Parkinson's disease: A nationwide population-based retrospective cohort study. Qjm. 2015;108(1):39–45.

33. Wijarnpreecha K, Boonpheng B, Thongprayoon C, Jaruvongvanich V, Ungprasert P. The association between non-alcoholic fatty liver disease and atrial fibrillation: A meta-analysis. Clin Res Hepatol Gastroenterol 2017 Oct;41(5):525-532.

34. Ambrosino P, Lupoli R, Di Minno A, Tarantino L, Spadarella G, Tarantino P, et al. The risk of coronary artery disease and cerebrovascular disease in patients with hepatitis C: A systematic review and meta-analysis. Int J Cardiol 2016;221:746-54.

35. Kepez A, Mutlu B, Degertekin M, Erol C. Association between left ventricular dysfunction, anemia, and chronic renal failure. Analysis of the Heart Failure Prevalence and Predictors in Turkey (HAPPY) cohort. Herz 2015;40(4):616–23.

36. Chang YL, Hung SH, Ling W, Lin HC, Li HC, Chung SD. Association between ischemic stroke and iron-deficiency anemia: a population-based study. PLoS One 2013;8(12):e82952.

37. Turakhia MP, Blankestijn PJ, Carrero JJ, Clase CM, Deo R, Herzog CA, et al. Chronic kidney disease and arrhythmias: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Eur Heart J 2018;39(24):2314–2325e.

38. Badsha MB, Mollah MN, Jahan N, Kurata H. Robust complementary hierarchical clustering for gene expression data analysis by β-divergence. J Biosci Bioeng 2013;116(3):397-407.

39. Yarnall AJ, Sayer AA, Clegg A, Rockwood K, Parker S, Hindle J V. New horizons in multimorbidity in older adults. Age Ageing 2017;46(6):882–8.

40. Ng SK. Tawiah R, McLachlan GJ. Unsupervised pattern recognition of mixed data structures with numerical and categorical features using a mixture regression modelling framework. Patter Recognition. 2019; 88: 261-271.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

Table 1. Characteristics of study participants aged 65-94 years stratified by multimorbidity and non-multimorbidity (N= 916 619, Catalonia, 2012)

Variables*	Multimorbidity (n= 853 085)	Non-multimorbidity (n= 63 534)	All (N=916 619)
Sex, women, n (%)	496 294 (58.2)	32 837 (51.7)	529 131 (57.7)
Age, mean (SD)	75.6 (7.4)	73.2 (7.3)	75.4 (7.4)
Age (categories), n (%)			
[65,70)	225 514 (26.4)	26 664 (42.0)	252 178 (27.5)
[70,80)	370 356 (43.4)	24 230 (38.1)	394 586 (43.0)
[80,90)	224 143 (26.3)	10 601 (16.7)	234 744 (25.6)
≥90	33 072 (3.9)	2039 (3.2)	35 111 (3.8)
MEDEA index†			
Q1	130 894 (16.5)	13 897 (23.4)	144 791 (17.0)
Q2	126 537 (16.0)	9894 (16.6)	136 431 (16.0)
Q3	129 246 (16.3)	8976 (15.1)	138 222 (16.2)
Q4	125 322 (15.8)	7666 (12.9)	132 988 (15.6)
Q5	110 916 (14.0)	5967 (10.0)	116 883 (13.7)
Rural	169 190 (21.4)	13 059 (22.0)	182 249 (21.4)
Number of chronic diseases, median [IQR]	6.0 [4.0;8.0]	1.0 [0.0;1.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)			
0	0 (0.0)	25 380 (39.9)	25 380 (2.8)
1	0 (0.0)	38 154 (60.1)	38 154 (4.2)
[2, 5)	268 836 (31.5)	0 (0.0)	268 836 (29.3)
[5,10)	463 709 (54.4)	0 (0.0)	463 709 (50.6)
≥10	120 540 (14.1)	0 (0.0)	120 540 (13.2)
Number of drugs, median [IQR]	5.0 [3.0;8.0]	0.0 [0.0;1.0]	5.0 [2.0;8.0]
Number of drugs (categories):			
0	72 557 (8.5)	40 811 (64.2)	113 368 (12.4)
1	48 704 (5.7)	8378 (13.2)	57 082 (6.2)
[2, 5)	247 095 (29.0)	11 572 (18.2)	258 667 (28.2)
[5,10)	360 030 (42.2)	2651 (4.2)	362 681 (39.6)
≥10	124 699 (14.6)	122 (0.2)	124 821 (13.6)
Number of visits, median [IQR]	10.0 [6.0;17.0]	1.0 [0.0;4.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)			
0	24 543 (2.9)	23,402 (36.8)	47 945 (5.2)
1	24 281 (2.8%)	9603 (15.1%)	33 884 (3.7)
[2, 5)	114 198 (13.4%)	16 241 (25.6%)	130 439 (14.2%)
[5, 10)	239 181 (28.0%)	10 168 (16.0%)	249 349 (27.2%)
≥10	450 882 (52.9%)	4120 (6.5%)	455 002 (49.6%)

All comparisons between variables in multimorbidity and non-multimorbidity showed P<0.001

†MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 2. Prevalence of the 60 chronic diseases included in the study in individuals aged 65-94 years (N= 916 619, Catalonia, 2012). In three last columns, list of diseases included by prevalence cut off (1%, 2%, All)

Rank	*Chronic conditions*	*Frequency*	*Percentage (%)*	*All diseases included*	*1%*	*2%*
1	Hypertension	650 899	71.0			
2	Dyslipidaemia	466 585	50.9			
3	Osteoarthritis and other degenerative joint diseases	300 803	32.8			
4	Obesity	262 888	28.7			
5	Diabetes	230 460	25.1			
6	Anaemia	167 577	18.3			
7	Cataract and other lens diseases	156 622	17.1			
8	Chronic kidney diseases	153 756	16.8			
9	Prostate diseases	153 635	16.8			
10	Osteoporosis	151 847	16.6			
11	Depression and mood diseases	148 751	16.2			
12	Solid neoplasms	137 045	15.0			
13	Colitis and related diseases	131 512	14.4			
14	Venous and lymphatic diseases	126 997	13.9			
15	Other musculoskeletal and joint diseases	124 765	13.6			
16	Dorsopathies	124 603	13.6			
17	Neurotic, stress-related and somatoform diseases	123 395	13.5			
18	COPD, emphysema, chronic bronchitis	109 603	12.0			
19	Ischemic heart disease	95 434	10.4			
20	Deafness, hearing impairment	90 261	9.9			
21	Sleep disorders	88 739	9.7			
22	Thyroid diseases	88 445	9.7			
23	Other genitourinary diseases	85 468	9.3			
24	Cerebrovascular disease	80 264	8.8			
25	Atrial fibrillation	80 247	8.8			
26	Esophagus, stomach and duodenum diseases	80 043	8.7			
27	Heart failure	74 077	8.1			
28	Other eye diseases	68 939	7.5			
29	Glaucoma	66 162	7.2			
30	Inflammatory arthropathies	62 450	6.8			
31	Dementia	59 213	6.5			
32	Cardiac valve diseases	52 100	5.7			
33	Peripheral neuropathy	49 127	5.4			
34	Other psychiatric and behavioural diseases	46 841	5.1			
35	Asthma	43 663	4.8			
36	Allergy	40 394	4.4			
37	Autoimmune diseases	39 350	4.3			
38	Ear, nose, throat diseases	38 752	4.2			
39	Peripheral vascular disease	30 674	3.4			
40	Other neurological diseases	28 541	3.1			
41	Chronic pancreas, biliary tract and gallbladder diseases	27 321	3.0			
42	Migraine and facial pain syndromes	25 999	2.8			
43	Bradycardias and conduction diseases	25 476	2.8			
44	Chronic liver diseases	22 633	2.5			
45	Other digestive diseases	22 022	2.4			
46	Parkinson and parkinsonism	20 833	2.3			
47	Other metabolic diseases	18 997	2.1			
48	Other cardiovascular diseases	16 833	1.8			
49	Other skin diseases	15 363	1.7			
50	Chronic ulcer of the skin	13 869	1.5			
51	Blood and blood forming organ diseases	13 575	1.5			
52	Other respiratory diseases	9974	1.1			
53	Epilepsy	8981	1.0			
54	Haematological neoplasms	8174	0.9			
55	Chronic infectious diseases	6647	0.7			
56	Inflammatory bowel diseases	5549	0.6			
57	Schizophrenia and delusional diseases	4792	0.5			
58	Blindness, visual impairment	4772	0.5			
59	Multiple sclerosis	576	0.1			
60	Chromosomal abnormalities	77	0.0			

Abbreviations: COPD: Chronic obstructive Pulmonary Disease.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

Table 3. Most frequent 15 diseases found in multimorbidity patterns in individuals aged 65-94 years (N= 916 619, Catalonia, 2012)

Pattern	Disease	O	O/E ratio	EX	Pattern	Disease	O	O/E ratio	EX
1 Nervous and digestive (n= 40 037)	Parkinson and parkinsonism	38.7	17.0	74.3	2 Respiratory, circulatory and nervous (n= 50 639)	Asthma	34.5	7.2	40.0
	Other neurological diseases	49.5	15.9	69.4		Peripheral vascular disease	13.9	4.2	22.9
	Chronic liver diseases	13.2	5.4	23.4		Parkinson and parkinsonism	8.5	3.8	20.8
	Chronic pancreas, biliary tract and gallbladder diseases	7.9	2.7	11.6		Other neurological diseases	11.7	3.7	20.7
	Dementia	14.7	2.3	9.9		COPD, emphysema, chronic bronchitis	31.0	2.6	14.3
	Other digestive diseases	4.8	2.0	8.7		Allergy	10.8	2.4	13.5
	Cerebrovascular disease	16.9	1.9	8.4		Heart failure	16.6	2.0	11.3
	Colitis and related diseases	24.1	1.7	7.3		Ischemic heart disease	21.1	2.0	11.2
	Other metabolic diseases	3.4	1.7	7.2		Other eye diseases	14.0	1.9	10.3
	Depression and mood diseases	25.0	1.5	6.7		Autoimmune diseases	7.2	1.7	9.3
	Anaemia	26.1	1.4	6.2		Other psychiatric and behavioural diseases	8.5	1.7	9.2
	Esophagus, stomach and duodenum diseases	11.3	1.3	5.6		Ear. nose. throat diseases	7.1	1.7	9.2
	Sleep disorders	12.4	1.3	5.6		Anaemia	30.4	1.7	9.2
	Other eye diseases	9.6	1.3	5.6		Peripheral neuropathy	8.8	1.6	9.1
	Dorsopathies	17.0	1.2	5.4		Cerebrovascular disease	14.3	1.6	9.0
3 Circulatory and digestive (n= 67 492)	Heart failure	51.4	6.4	46.9	4 Mental, nervous and digestive, female dominant (n= 94 453)	Neurotic, stress-related and somatoform diseases	64.9	4.8	49.7
	Cardiac valve diseases	34.2	6.0	44.3		Depression and mood diseases	66.4	4.1	42.1
	Atrial fibrillation	47.3	5.4	39.8		Migraine and facial pain syndromes	8.2	2.9	29.6
	Bradycardias and conduction diseases	13.5	4.9	35.9		Sleep disorders	19.0	2.0	20.2
	Ischemic heart disease	33.7	3.2	23.8		Esophagus. stomach and duodenum diseases	14.9	1.7	17.6
	Chronic pancreas, biliary tract and gallbladder diseases	8.0	2.7	19.7		Osteoporosis	28.0	1.7	17.4
	Chronic liver diseases	6.1	2.5	18.2		Thyroid diseases	16.0	1.7	17.1
	Chronic kidney diseases	35.9	2.1	15.8		Colitis and related diseases	23.7	1.7	17.0
	Anemia	38.6	2.1	15.5		Other genitourinary diseases	14.4	1.5	15.9
	Cerebrovascular disease	18.3	2.1	15.4		Ear, nose, throat diseases	6.2	1.5	15.2
	COPD, emphysema, chronic bronchitis	23.6	2.0	14.5		Venous and lymphatic diseases	19.9	1.4	14.8
	Other digestive diseases	4.6	1.9	14.0		Allergy	6.1	1.4	14.3
	Peripheral vascular disease	6.1	1.8	13.3		Osteoarthritis and other degenerative joint diseases	45.0	1.4	14.1
	Other metabolic diseases	3.2	1.5	11.3		Dorsopathies	18.0	1.3	13.7
	Dementia	9.5	1.5	10.9		Cardiac valve diseases	7.4	1.3	13.5
5 Mental, digestive and blood, female oldest-old dominant (n= 106 845)	Dementia	21.8	3.4	39.4	6 Nervous, musculoskeletal and circulatory, female dominant (n= 145 074)	Peripheral neuropathy	12.4	2.3	36.6
	Other digestive diseases	5.8	2.4	28.1		Other musculoskeletal and joint diseases	26.0	1.9	30.2
	Anemia	38.5	2.1	24.6		Venous and lymphatic diseases	26.4	1.9	30.2
	Chronic kidney diseases	33.3	2.0	23.1		Dorsopathies	25.3	1.9	29.4
	Colitis and related diseases	26.2	1.8	21.3		Obesity	51.0	1.8	28.2
	Cerebrovascular disease	14.8	1.7	19.7		Other genitourinary diseases	16.0	1.7	27.2
	Osteoporosis	26.0	1.6	18.3		Osteoarthritis and other degenerative joint diseases	55.0	1.7	26.5
	Cataract and other lens diseases	25.9	1.5	17.7		Osteoporosis	24.8	1.5	23.7
	Deafness. hearing impairment	14.0	1.4	16.5		Other eye diseases	10.7	1.4	22.4
	Venous and lymphatic diseases	19.5	1.4	16.4		Cataract and other lens diseases	22.5	1.3	20.8
	Osteoarthritis and other degenerative joint diseases	45.5	1.4	16.2		Thyroid diseases	12.6	1.3	20.7
	Depression and mood diseases	22.5	1.4	16.1		Glaucoma	9.2	1.3	20.1
	Other genitourinary diseases	12.3	1.3	15.4		Diabetes	31.3	1.2	19.7
	Other eye diseases	9.9	1.3	15.4		Ear, nose, throat diseases	5.2	1.2	19.5
	Sleep disorders	12.4	1.3	14.9		Dyslipidemia	62.7	1.2	19.5
7 Genitourinary, mental and musculoskeletal, male dominant (n=173 746)	Prostate diseases	54.7	3.3	61.8	8 Non-specified, youngest-old dominant(n=238 333)	Dyslipidemia	38.4	0.8	19.6
	Other psychiatric and behavioural diseases	11.1	2.2	41.2		Thyroid diseases	7.3	0.8	19.6
	Inflammatory arthropathies	12.4	1.8	34.5		Osteoporosis	12.2	0.7	19.2
	COPD, emphysema, chronic bronchitis	20.5	1.7	32.5		Hypertension	47.6	0.7	17.4
	Solid neoplasms	21.8	1.5	27.7		Glaucoma	4.4	0.6	16.0
	Peripheral vascular disease	4.7	1.4	26.7		Solid neoplasms	9.1	0.6	15.7
	Ischemic heart disease	13.7	1.3	25.0		Migraine and facial pain syndromes	1.7	0.6	15.7
	Diabetes	31.8	1.3	24.0		Autoimmune diseases	2.2	0.5	13.4
	Ear, nose, throat diseases	5.3	1.3	23.7		Other metabolic diseases	1.1	0.5	13.3
	Deafness, hearing impairment	11.6	1.2	22.3		Allergy	2.2	0.5	13.0
	Allergy	4.8	1.1	20.5		Chronic liver diseases	1.2	0.5	12.8
	Hypertension	75.8	1.1	20.2		Other genitourinary diseases	4.5	0.5	12.7
	Glaucoma	7.5	1.0	19.6		Esophagus, stomach and duodenum diseases	4.1	0.5	12.2
	Autoimmune diseases	4.4	1.0	19.4		Other psychiatric and behavioral diseases	2.4	0.5	12.0
	Obesity	29.0	1.0	19.2		Diabetes	10.8	0.4	11.2

Abbreviations: O: Disease prevalence in the cluster; O/E ratio: observed/expected ratio; Ex: exclusivity; COPD: Chronic obstructive Pulmonary

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES)

Table 4. Variables characterizing each cluster in baseline study for 2% prevalence cut-off point (N= 916 619).

	1.Nervous and digestive	2. Respiratory, circulator and nervous	3. Circulatory and digestive	4. Mental, nervous and digestive, female dominant	5. Mental, digestive and blood, female oldest-old dominant	6. Nervous, musculoskeletal and circulatory, female dominant	7. Genitourinary, mental and musculoskeletal, male dominant	8. Non-specified, youngest-old dominant	All
Number of people, n	40 037	50 639	67 492	94 453	106 845	145 0[illegible]	173 746	238 333	916 619
Multimorbidity, n (%)	39 776 (99.3)	50 513 (99.8)	67 443 (99.9)	94 442 (100.0)	106 696 (99.9)	144 86[illegible] (99.[illegible])	171 983 (99.0)	177 363 (74.4)	853 085 (93.1)
Polypharmacy, n (%)	28 484 (71.1)	38 869 (76.8)	54 658 (81.0)	64 154 (67.9)	71 830 (67.2)	86 317 (59.[illegible])	90 603 (52.1)	52 588 (22.1)	487 502 (53.1)
Women, n (%)	22 628 (56.5)	26 690 (52.7)	38 023 (56.3)	78 922 (83.6)	85 735 (80.2)	113 62[illegible] (78.[illegible])	15 730 (9.1)	147 773 (62.0)	529 131 (57.7)
Men, n (%)	17 409 (43.5)	23 949 (47.3)	29 469 (43.7)	15 531 (16.4)	21 110 (19.8)	31 44[illegible] (21.[illegible])	158 016 (90.9)	90 560 (38.0)	387 488 (42.3)
Age (categories), n (%)									
[65,70)	7188 (18.0)	10 400 (20.5)	7233 (10.7)	28 305 (30.0)	12 036 (11.3)	38 829 (26.[illegible])	52 003 (29.9)	96 184 (40.4)	252 178 (27.5)
[70,80)	17 804 (44.5)	22 743 (44.9)	24 724 (36.6)	40 577 (43.0)	33 624 (31.5)	70 643 (48.[illegible])	84 037 (48.4)	100 435 (42.1)	394 586 (43.0)
[80,90)	13 460 (33.6)	15 568 (30.7)	29 908 (44.3)	22 638 (24.0)	48 453 (45.3)	32 714 (22.[illegible])	34 785 (20.0)	37 217 (15.6)	234 744 (25.6)
[90,99]	1587 (4.0)	1927 (3.8)	5628 (8.3)	2934 (3.1)	12 732 (11.9)	2888 (2.[illegible])	2920 (1.7)	4497 (1.9)	35 111 (3.8)
MEDEA* index									
R	7831 (21.8)	9300 (20.2)	13 718 (23.2)	17 266 (19.7)	22 183 (23.0)	27 40[illegible] (21.[illegible])	35 145 (21.5)	49 405 (21.9)	182249 (21.4)
U1	6010 (16.7)	6890 (15.0)	9537 (16.1)	15 027 (17.2)	16 556 (17.2)	19 599 (15.[illegible])	25 656 (15.7)	45 516 (20.2)	144791 (17.0)
U2	5690 (15.8)	7134 (15.5)	9140 (15.4)	14 335 (16.4)	15 272 (15.8)	21 379 (16.[illegible])	25 951 (15.9)	37 530 (16.6)	136431 (16.0)
U3	5941 (16.5)	7520 (16.4)	9187 (15.5)	14 223 (16.3)	15 421 (16.0)	23 26[illegible] (17.[illegible])	26 908 (16.5)	35 761 (15.8)	138222 (16.2)
U4	5540 (15.4)	7686 (16.7)	9016 (15.2)	14 012 (16.0)	14 272 (14.8)	23 78[illegible] (17.[illegible])	26 526 (16.2)	32 157 (14.2)	132988 (15.6)
U5	4982 (13.8)	7421 (16.2)	8638 (14.6)	12 652 (14.5)	12 699 (13.2)	21 923 (16.[illegible])	23 064 (14.1)	25 506 (11.3)	116883 (13.7)
Number of chronic diseases, median [IQR]	8.0 [6.0;10.0]	8.0 [6.0;10.0]	8.0 [7.0;11.0]	7.0 [6.0;9.0]	7.0 [5.0;9.0]	6.0 [5.0;8.0]	5.0 [4.0;7.0]	3.0 [3.0;4.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)									
0	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.0%)	0 (0.0%)	235 (0.1)	25 144 (10.5)	25 380 (2.8)
1	262 (0.7)	125 (0.2)	49 (0.1)	11.0 (0.0)	149 (0.1)	204 (0.[illegible])	1528 (0.9)	35 826 (15.0)	38 154 (4.2)
[2, 5)	5409 (13.5	4507 (8.9)	4275 (6.3)	8781 (9.3)	14 601 (13.7)	22 409 (15.4)	57 561 (33.1)	151 302 (63.5)	268 836 (29.3)
[5,10)	23 502 (58.7)	30 257 (59.8)	37 910 (56.2)	62 490 (66.2)	73 427 (68.7)	105 62[illegible] (72.[illegible])	104 915 (60.4)	25 588 (10.7)	463 709 (50.6)
≥10	10 864 (27.1)	15 749 (31.1)	25 259 (37.4)	231 715 (24.5)	18 668 (17.5)	16 856 (11.6)	9506 (5.5)	473 (0.2)	120 540 (13.2)
Number of drugs, median [IQR]	7.0 [4.0;9.0]	7.0 [5.0;10.0]	8.0 [5.0;11.0]	6.0 [4.0;9.0]	6.0 [4.0;9.0]	5.0 [3.0;8.0]	5.0 [3.0;7.0]	2.0 [0.0;4.0]	5.0 [2.0;8.0]
Number of drugs (categories)									
0	2576 (6.4)	2491 (4.9)	3349 (5.0)	5636 (6.0)	7037 (6.6)	8330 (5.7)	13 389 (7.7)	70 561 (29.6)	113 368 (12.4)
1	1212 (3.0)	1072 (2.1)	1015 (1.5)	2939 (3.1)	3390 (3.2)	6772 (4.7)	11 440 (6.6)	29 242 (12.3)	57 082 (6.2)
[2, 5)	7766 (19.4)	8207 (16.2)	8471 (12.6)	21 725 (23.0)	24 587 (23.0)	43 656 (30.[illegible])	58 314 (33.6)	85 942 (36.1)	258 667 (28.2)
[5,10)	18 510 (46.2)	23 597 (46.6)	31 850 (47.2)	46 022 (48.7)	52 653 (49.3)	68 193 (47.[illegible])	73 694 (42.4)	48 161 (20.2)	362 681 (39.6)
≥10	9973 (24.9)	15 272 (30.2)	22 808 (33.8)	18 132 (19.2)	19 177 (17.9)	18 123 (12.[illegible])	16 909 (9.7)	4427 (1.9)	124 821 (13.6)
Number of visits 2012, median [IQR]	12.0 [7.0;20.0]	14.0 [8.0;22.0]	18.0 [9.0;30.0]	11.0 [6.0;19.0]	12.0 [7.0;19.0]	11.0 [6.0;17.0]	9.0 [5.0;15.0]	5.0 [2.0;9.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)									
0	976 (2.4)	871 (1.7)	1143 (1.7)	2219 (2.3)	2515 (2.4)	2410.3 (1.7)	4137 (2.4)	33 673 (14.1)	47 945 (5.2)
1	874 (2.2)	754 (1.5)	929 (1.4)	2055 (2.2)	2238 (2.1)	2412.4 (1.7)	4685 (2.7)	19 938 (8.4)	33 884 (3.7)
[2, 5)	4000 (10.0)	3918 (7.7)	4329 (6.4)	10 589 (11.2)	11 018 (10.3)	14943.7 (10.[illegible])	24 319 (14.0)	57 322 (24.1)	130 439 (14.2)
[5, 10)	9158 (22.9)	10 774 (21.3)	10 883 (16.1)	24 504 (25.9)	27 003 (25.3)	42180.7 (29.[illegible])	54 212 (31.2)	70 634 (29.6)	249 349 (27.2)
≥10	25 030 (62.5)	34 322 (67.8)	50 209 (74.4)	55 085 (58.3)	64 071 (60.0)	83126.5 (57.[illegible])	86 393 (49.7)	56 766 (23.8)	455 002 (49.6)

For the sake of simplicity, all numbers in the table were rounded to its closest natural number. *MEDEA index goes from 1 (least deprived) to 5 (most deprived), in this variable n=851 564.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES). Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Figure 1. Study population flow chart

*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

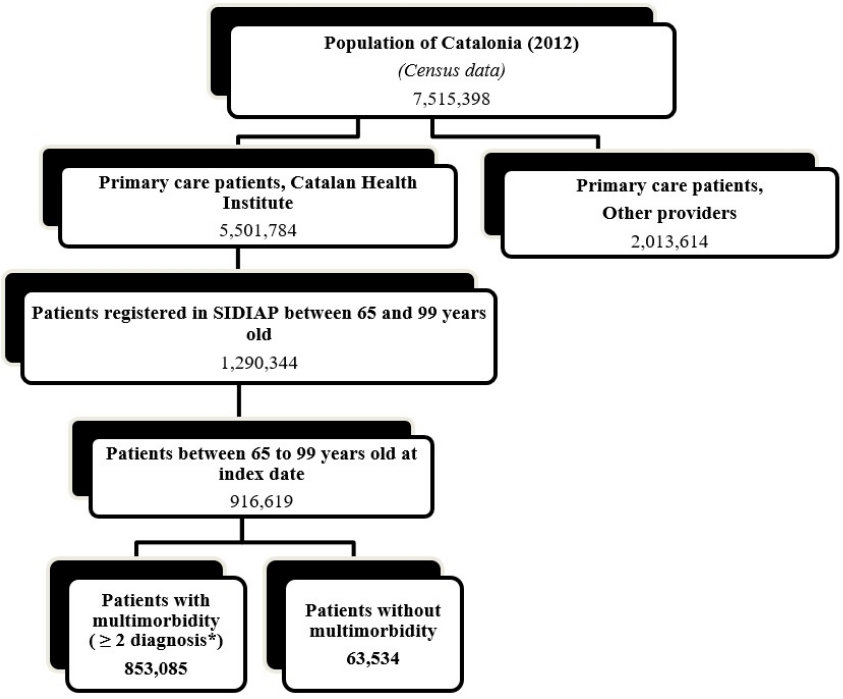


Figure 1. Study population flow chart
*See 60 chronic diseases group defined in Swedish National study of Aging and Care in Kungsholmen (SNAC-K) (25).

217x161mm (115 x 115 DPI)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

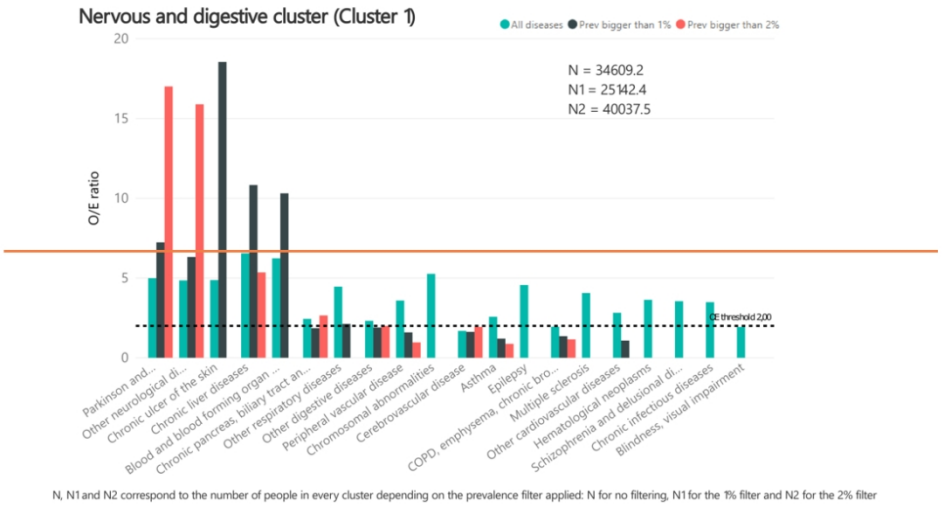


Figure 2. Composition of cluster 1 (Nervous and digestive) in individuals aged 65-94 years according to disease levels of prevalence (N= 916 619, Catalonia, 2012)

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Additional File 1

A) Extracting and Validating Multimorbidity Patterns by applying the Fuzzy C Means Clustering algorithm.

In this annex we present a description of the procedure followed to obtain a set of multimorbidity patterns characterizing a patient population aged 65 or more in Catalonia (Spain).

Dataset dimension reduction.

The initial dataset was composed on 31st December, 2012, of a registered active diagnosis with a certain prevalence value, out of 60 possible diseases for the N=916,619 patients included in the study. Additionally, considering age and the gender, each patient was initially characterized by a vector of 62 features, most of which were binary variables indicating the presence/absence of a disease at the end of 2012. For most of the study, diseases with prevalence ≥2% were filtered, resulting in 47 diseases and the corresponding 49 features (adding age and gender). Since most of the selected features were categorical instead of quantitative, the dataset was a mixture of numerical and categorical variables. We processed this dataset by applying a mixture of the well-known Principal Component Analysis (PCA) to the numeric original features and a Multiple Correspondence Analysis (MCA) to the binary ones, in order to obtain a new dataset of reduced dimension. We selected the PCAmix algorithm, as described by Chavent et al, to perform the dimensionality reduction. It follows the criterion based on concentrating most of the variability of the new transformed features, that is to say, variance of the data in the low-dimensional representation were maximized. The Karlis-Saporta-Spinaki rule was followed to select the first 13 dimensions out of the 49 for the 2% prevalence filtering, according to the eigenvalues of the PCAmix and the number of features and individuals in the dataset. As a result, after the PCAmix transformation and the extraction of the optimal number of dimensions, the new dataset was composed of N=916,619 vectors of $d = 13$ features each one. In the following we denote this new dataset as $\boldsymbol{\mathcal{Y}} \coloneqq \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$, denoting by $\boldsymbol{y}_2 \in \mathbb{R}^{13}$ for $n = 1, \dots, N$ the new vector representing patient n.

Soft clustering algorithm

Once the transformed dataset $\boldsymbol{\mathcal{Y}}$ was computed, a soft clustering algorithm was applied to fuzzily distribute the population into a set of clusters, corresponding to the different multimorbidity patterns. In a traditional clustering procedure patients are grouped in an exclusive way, so that if a certain patient belongs to a definite cluster then s/he cannot be included in another one. In contrast, an overlapping clustering, such as the Fuzzy C Means (FCM) algorithm, uses fuzzy sets to cluster patients, so that each patient belongs to all clusters with different degrees of

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

membership. The choice between a hard or a soft clustering algorithm is traditionally made based on the application and the performance obtained. In our case, the use of the FCM algorithm presented performance results similar to those of the hard clustering algorithm Kmeans, but clinically more solid. It was, therefore, chosen as the most appropriate method for the description of the multimorbidity patterns.

FCM was originally introduced by Bezdek and yields an unsupervised form of grouping in which individuals can belong to more than one cluster. To do so, they are associated with an appropriate set of K membership values, where K denotes the number of clusters. The parameters that determine the clustering process are a set of K centroids $\mathbf{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_K\}$ where $v_k \in \mathbb{R}^{13}$ for $k = 1, \dots, K$ and a set of membership factors $\mathbf{U} = \{u_{jn}; j = 1, \dots, K; n = 1, \dots, N\}$ with $0 \leq u_{jn} \leq 1$. Factor u_{jn} indicates the degree to which individual n^{th} belongs to cluster j^{th}. Both centroids $\mathbf{V}$ and membership factors $\mathbf{U}$ are obtained by iteratively minimizing the objective function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$, which is the weighted sum of squared errors within clusters

$$J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) = \sum_{n=1}^{N} \sum_{j=1}^{K} (u_{jn})^m \left\| \boldsymbol{y}_n - \boldsymbol{v}_j \right\|^2; \quad 1 < m < \infty \quad (1)$$

Thus, the similarity between an individual and a cluster centroid is measured through the squared error between the vector associated with the patient and the centroid prototyping the cluster. The fuzziness weighting parameter m, is selected to adjust the blending of the different clusters and it is any real number greater than 1. High m values would produce a fuzzy set of clusters so that individuals would tend to be equally distributed across clusters, whereas lower ones would generate a non-overlapped set of clusters. The FCM method iteratively alternates between computing the centroids in $\mathbf{V}$ as the average of the individual's features in $\boldsymbol{\mathcal{Y}}$ previously weighted by the correspondent membership factors and estimating the membership factors in $\mathbf{U}$ in order to maximize the cost function $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$ given the updated centroids in $\mathbf{V}$. In our work, we randomly initialized the set of centroids $\mathbf{V}$ and halted the iterative process when $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}}) < \epsilon$, where $0 < \epsilon \ll 1$. This procedure converges to a local minimum or saddle point of $J_m(\mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{Y}})$.

Cluster stability validation.

Stable clusters are required in order to characterize multimorbidity patterns, consequently we applied 100 FCM independent runs to the transformed dataset $\boldsymbol{\mathcal{Y}}$ and averaged both the membership factors and the centroid vectors, after ordering the clusters in descending order in terms of the summation of memberships to clusters, measured as $\sum_{n=1}^{N} (u_{jn})^m$. This is equivalent to selecting the centroid and membership factors associated with the cluster with more population in each run and averaging them. Then after removing the selected cluster from each set, the procedure is repeated until a final set of clusters, composed of the K averaged

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

centroids and the corresponding averaged membership factors, is obtained. In this averaging process we previously verified the similarity between the averaged parameters by a heuristic inspection of some randomly selected run results

Number of clusters and fuzziness parameter validation.

Since clustering algorithms are unsupervised, machine-learning techniques, the model fitting the dataset is traditionally computed through cost functions that depend on both the dataset and the clustering parameters and are denoted as validation indices. We computed three different well-known validation indices to obtain the optimal number of clusters K and the optimal value of the fuzziness parameter m: the partition coefficient validation index whose cost function is maximum for the optimal model, the Xie-Beni, and the partition entropy validation indices whose cost functions are minimum for the optimal models. A cross-validation technique was applied using a split sample approach, by randomly dividing the individuals into two different datasets, a first (50%) training dataset used for obtaining the averaged FCM clusters, and a second (50%) test dataset used to verify the model fitting the data.

This validation procedure was applied to the set of clusters obtained after the previously explained averaging process, with the 2% prevalence filtering and considering 49 features before PCAmix reduction. We checked $m = 1.1, 1.2,$ and 1.5 and $K = 5,..,20$. In Figure1 the performance obtained through the three validation indices is depicted. The best behaviour is obtained for m=1.1 and as is shown in Figure 2 and Figure 3 we can conclude that the optimal number of clusters for m=1.1 ranges from 6 to 12, validated with both the training dataset and the test dataset (more details are given in figures).

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

B) Computation of the observed/expected ratio and the exclusivity ratio.

The observed/expected $(O/E)_{dj}$ ratio and the exclusivity ratio EX_{dj} have been used in this work in order to decide whether a disease d is overrepresented or not in any given cluster j.

The $(O/E)_{dj}$ ratio was calculated by dividing disease prevalence in the cluster O_{dj} by disease prevalence in the overall population E_d. As membership of an individual n in a cluster j was denoted by a membership degree factor u_{nj}, and not as a binary variable, the observed disease prevalence O_{dj} in a cluster j was computed as the ratio between the summation of the membership degree factors corresponding to all individuals suffering the disease d and the summation of all the membership degree factors corresponding to the cluster j. Let us assume that there are n_d individuals suffering the disease d and that they are grouped in the set I_d, then the observed prevalence was computed as

$$O_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{\sum_{n=1}^{N} u_{nj}}$$

while the expected prevalence was computed as

$$E_d = \frac{n_d}{N}$$

Exclusivity ratio EX_{dj}, defined as the proportion of individuals with the disease d included in the cluster j over the total number of individuals with the disease n_d, was computed as

$$EX_{dj} = \frac{\sum_{n \in I_d} u_{nj}}{n_d}$$

References

1. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate analysis of mixed data: The PCAmixdata R package. 2014; eprint arXiv:1411.4911.

2. Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press; 1981.

3. Bora D, Kumar Gupta A. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int J Comput Trends Technol 2014;10(2):108–13.

4. Pal NR, Bezdek JC. On Cluster Validity for the Fuzzy c-Means Model. IEEE Trans Fuzzy Syst 1995;3(3):370–9.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

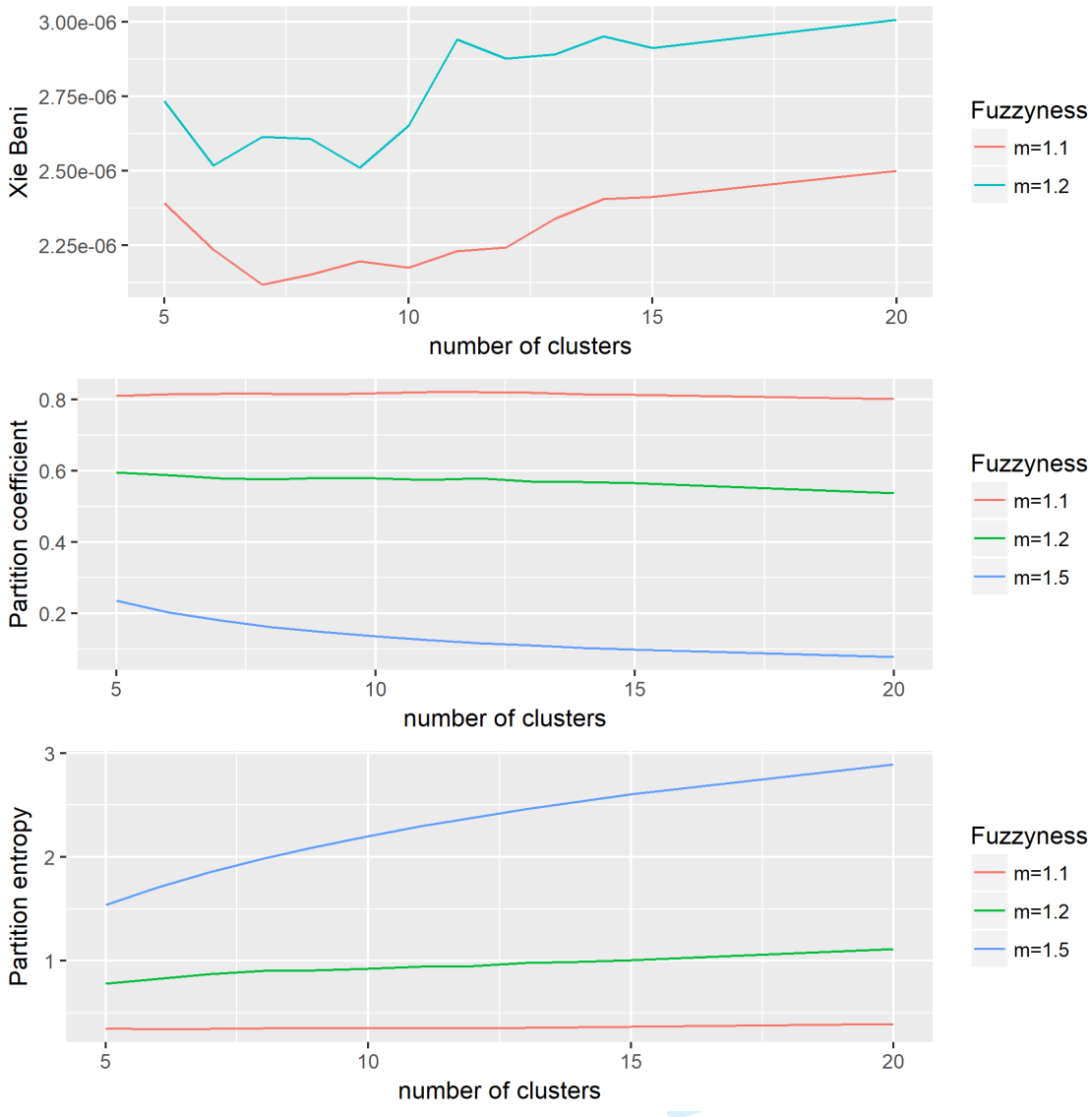


Figure 1. Selection of the optimal m parameter

Index m = 1.5 was also computed for Xie-Beni indices, but not included in the graph because the curve is significantly higher than the other two in the plot.

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. For this reason, all plots are showing that m = 1.1 is the best parameter to optimize all the computed indices.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

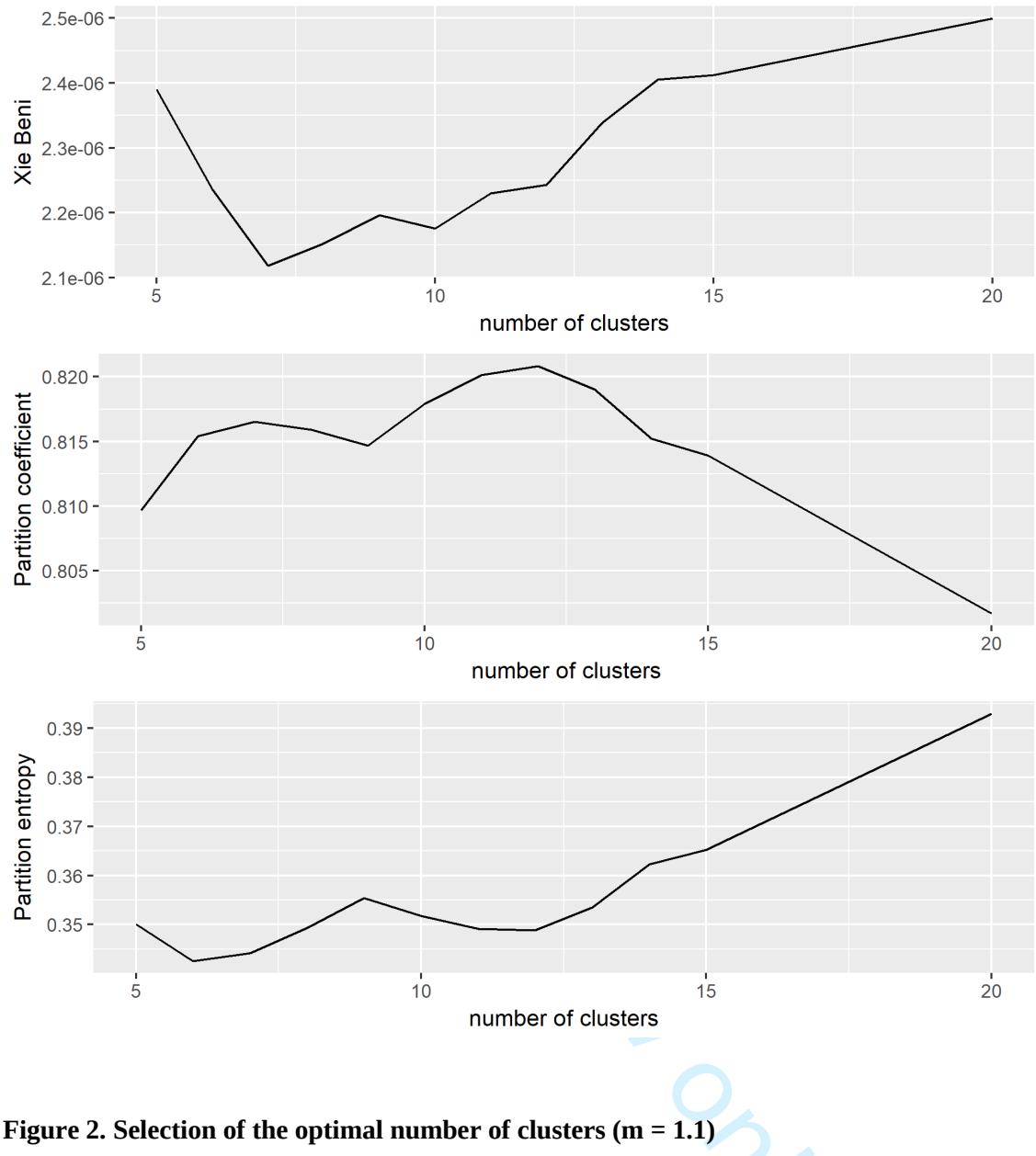


Figure 2. Selection of the optimal number of clusters (m = 1.1)

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. Within the plots above, optimal values are located in the range from 6 to 12 clusters.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

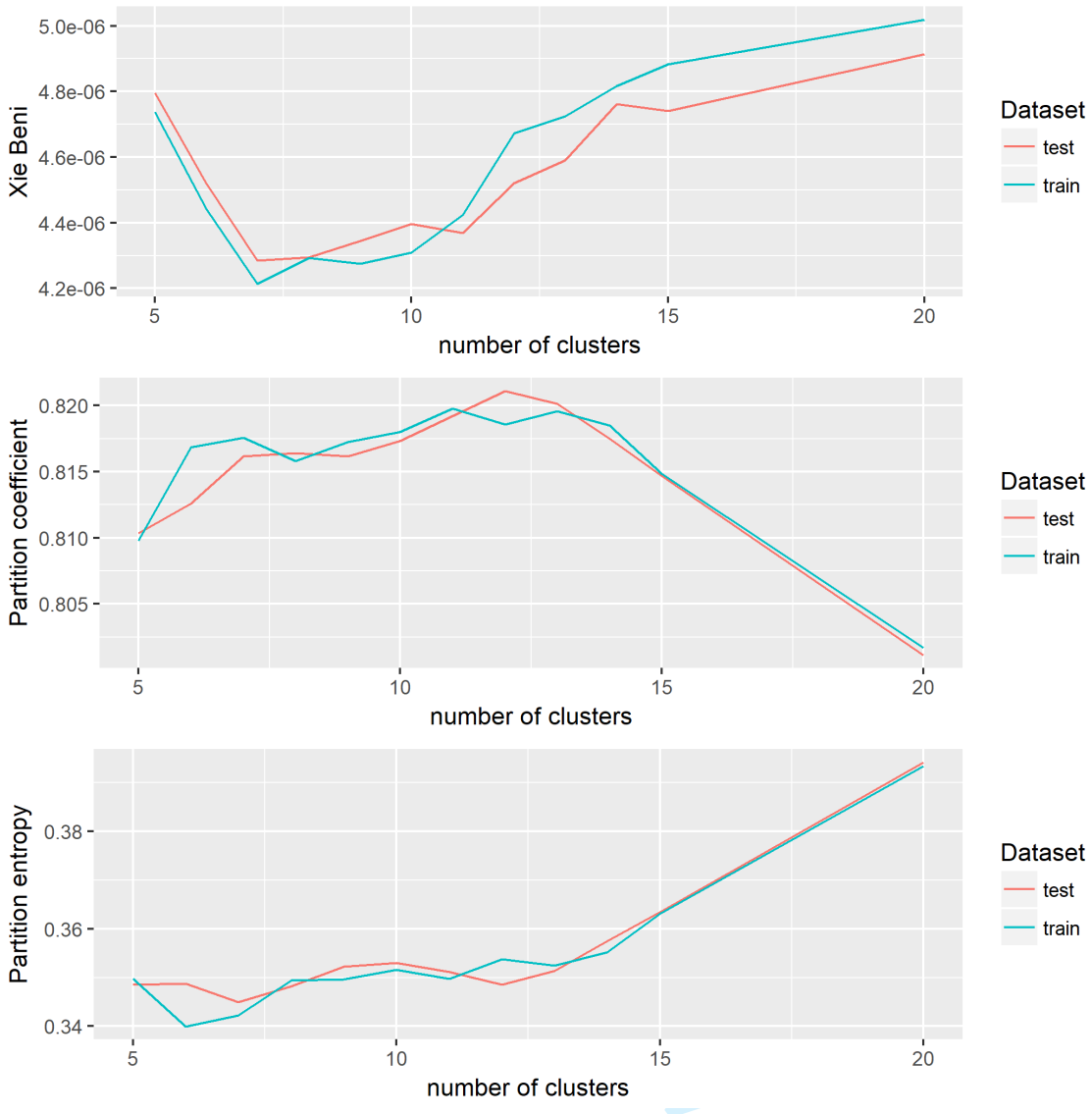


Figure 3. Cross-validation of the clustering with m = 1.1

Optimum Xie-Beni and partition entropy indices are at the minimum, whereas optimal choice for partition coefficient is at the maximum. In the plots above we can find the optimal values in the range from 6 to 12 clusters. Additionally, no significant variation is registered in the indices regardless of the dataset selection.

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Additional File 2.

Table 1. Variables characterizing each cluster in baseline study for 1% prevalence cut-off point (N= 916 619)

	1.Nervous and digestive	2. Respiratory, circulator and nervous	3. Circulatory and digestive	4. Mental, nervous and digestive, female dominant	5. Mental, digestive and blood, female oldest-old dominant	6. Nervous, musculoskeletal and circulatory, female dominant	7. Genitourinary, mental and musculoskeletal, male dominant	8. Non-specified, youngest-old dominant	All
Number of people, n	25 142	46 144	64 299	86 819	113 910	154 411	178 511	247 382	916 619
Multimorbidity, n (%)	25 011 (99.5)	45 969 (99.6)	64 210 (99.9)	86 815 (100.0)	113 869 (100.0)	154 406 (100.0)	177 392 (99.4)	185 414 (75.0)	853 085 (93.1)
Polypharmacy, n (%)	16 859 (67.1)	33 629 (72.9)	49 776 (77.4)	64 969 (74.8)	76 376 (67.0)	96 657 (62.6)	94 463 (52.9)	54 773 (22.1)	487 502 (53.1)
Women, n (%)	14 637 (58.2)	26 113 (56.6)	38 930 (60.5)	61 441 (70.8)	95 491 (83.8)	135 476 (87.7)	4 675 (2.6)	152 368 (61.6)	529 131 (57.7)
Men, n (%)	10 506 (41.8)	20 031 (43.4)	25 369 (39.5)	25 378 (29.2)	18 419 (16.2)	18 935 (12.3)	173 836 (97.4)	95 014 (38.4)	387 488 (42.3)
Age (categories), n (%)									
[65,70)	4 766 (19.0)	8 485 (18.4)	8 980 (14.0)	18 070 (20.8)	23 078 (20.3)	35 167 (22.8)	53 918 (30.2)	99 715 (40.3)	252 178 (27.5)
[70,80)	10 562 (42.0)	19 970 (43.3)	24 698 (38.4)	34 460 (39.7)	43 362 (38.1)	72 030 (46.6)	86 357 (48.4)	103 146 (41.7)	394 586 (43.0)
[80,90)	8 367 (33.3)	15 458 (33.5)	25 810 (40.1)	29 261 (33.7)	39 382 (34.6)	41 966 (27.2)	35 304 (19.8)	39 197 (15.8)	234 744 (25.6)
[90,99]	1 448 (5.8)	2 230 (4.8)	4 811 (7.5)	5 028 (5.8)	8 089 (7.1)	5 248 (3.4)	2 933 (1.6)	5 324 (2.2)	35 111 (3.8)
MEDEA* index									
R	4 921 (19.6)	8 815 (19.1)	12 845 (20.0)	16 718 (19.3)	22 224 (19.5)	29 369 (19.0)	35 849 (20.1)	51 507 (20.8)	182249 (21.4)
U1	3 669 (14.6)	6 651 (14.4)	9 244 (14.4)	13 108 (15.1)	17 669 (15.5)	21 028 (13.6)	26 416 (14.8)	47 006 (19.0)	144791 (17.0)
U2	3 513 (14.0)	6 502 (14.1)	8 859 (13.8)	12 527 (14.4)	16 843 (14.8)	22 642 (14.7)	26 697 (15.0)	38 847 (15.7)	136431 (16.0)
U3	3 624 (14.4)	6 806 (14.7)	9 057 (14.1)	12 495 (14.4)	16 973 (14.9)	24 536 (15.9)	27 619 (15.5)	37 112 (15.0)	138222 (16.2)
U4	3 452 (13.7)	6 586 (14.3)	8 808 (13.7)	12 279 (14.1)	16 327 (14.3)	24 859 (16.1)	27 294 (15.3)	33 383 (13.5)	132988 (15.6)
U5	3 206 (12.8)	6 188 (13.4)	8 305 (12.9)	11 362 (13.1)	14 676 (12.9)	23 003 (14.9)	23 650 (13.2)	26 493 (10.7)	116883 (13.7)
Number of chronic diseases, median [IQR]	8.0 [6.0; 10.0]	8.0 [6.0; 10.0]	8.0 [6.0; 10.0]	8.0 [6.0; 10.0]	7.0 [5.0; 9.0]	6.0 [5.0; 8.0]	5.0 [4.0; 7.0]	3.0 [1.0; 4.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)									
0	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	25 380 (10.3)	25 380 (2.8)
1	131 (0.5)	175 (0.4)	90 (0.1)	5 (0.0)	41 (0.0)	5 (0.0)	1 120 (0.6)	36 588 (14.8)	38 154 (4.2)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES). Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

[2, 5)	3 207 (12.8)	5 466 (11.8)	5 560 (8.6)	6 424 (7.4)	13 367 (11.7)	18 862 (12.2)	57 441 (32.2)	158 509 (64.1)	268 836 (29.3)
[5,10)	14 285 (56.8)	27 482 (59.6)	37 649 (58.6)	54 013 (62.2)	78 670 (69.1)	116 135 (75.2)	109 238 (61.2)	26 237 (10.6)	463 709 (50.6)
≥10	7 520 (29.9)	13 021 (28.2)	21 000 (32.7)	26 377 (30.4)	21 832 (19.2)	19 409 (12.6)	10 713 (6.0)	668 (0.3)	120 540 (13.2)
Number of drugs, median [IQR]	6.0 [4.0; 9.0]	7.0 [4.0; 10.0]	7.0 [5.0; 10.0]	7.0 [4.0; 10.0]	6.0 [4.0; 9.0]	6.0 [3.0; 8.0]	5.0 [3.0; 7.0]	2.0 [0.0; 4.0]	5.0 [2.0;8.0]
Number of drugs (categories)									
0	1 988 (7.9)	2 733 (5.9)	3 420 (5.3)	4 605 (5.3)	6 936 (6.1)	8 160 (5.3)	13 098 (7.3)	72 427 (29.3)	113 368 (12.4)
1	965 (3.8)	1 256 (2.7)	1 268 (2.0)	1 913 (2.2)	3 633 (3.2)	6 072 (3.9)	11 575 (6.5)	30 400 (12.3)	57 082 (6.2)
[2, 5)	5 330 (21.2)	8 526 (18.5)	9 835 (15.3)	15 332 (17.7)	26 965 (23.7)	43 522 (28.2)	59 374 (33.2)	89 782 (36.3)	258 667 (28.2)
[5,10)	11 033 (43.9)	21 308 (46.2)	30 250 (47.0)	42 078 (48.5)	56 341 (49.5)	75 147 (48.7)	76 377 (42.7)	50 148 (20.3)	362 681 (39.6)
≥10	5 826 (23.2)	12 321 (26.7)	19 525 (30.4)	22 891 (26.4)	20 036 (17.6)	21 510 (13.9)	18 086 (10.1)	4 625 (1.9)	124 821 (13.6)
Number of visits 2012, median [IQR]	13.0 [7.0; 23.0]	13.0 [7.0; 21.0]	15.0 [8.0; 26.0]	14.0 [7.0; 24.0]	11.0 [6.0; 18.0]	11.0 [7.0; 17.0]	10.0 [6.0; 16.0]	5.0 [2.0; 9.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)									
0	667 (2.7)	983 (2.1)	1 212 (1.9)	1 727 (2.0)	2 563 (2.3)	2 459 (1.6)	3 916 (2.2)	34 418 (13.9)	47 945 (5.2)
1	550 (2.2)	887 (1.9)	1 070 (1.7)	1 536 (1.8)	2 342 (2.1)	2 282 (1.5)	4 671 (2.6)	20 546 (8.3)	33 884 (3.7)
[2, 5)	2 389 (9.5)	4 242 (9.2)	5 030 (7.8)	7 700 (8.9)	12 166 (10.7)	14 734 (9.5)	24 789 (13.9)	59 389 (24.0)	130 439 (14.2)
[5, 10)	5 390 (21.4)	10 384 (22.5)	12 356 (19.2)	18 483 (21.3)	29 941 (26.3)	43 668 (28.3)	55 517 (31.1)	73 610 (29.8)	249 349 (27.2)
≥10	16 146 (64.2)	29 647 (64.3)	44 631 (69.4)	57 373 (66.1)	66 898 (58.7)	91 267 (59.1)	89 618 (50.2)	59 420 (24.0)	455 002 (49.6)

For the sake of simplicity, all numbers in the table were rounded to its closest natural number

For peer review only

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . d by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 2. Variables characterizing each cluster in baseline study for no prevalence cut-off point (N= 916 619)

	1.Nervous and digestive	2. Respiratory, circulator and nervous	3. Circulatory and digestive	4. Mental, nervous and digestive, female dominant	5. Mental, digestive and blood, female oldest-old dominant	6. Nervous, musculoskeletal and circulatory, female dominant	7. Genitourinary, mental and musculoskeletal, male dominant	8. Non-specified, youngest-old dominant	All
Number of people, n	34 609	56 724	70 178	87 885	108 469	155 860	170 170	232 723	916 619
Multimorbidity, n (%)	34 446 (99.5)	56 618 (99.8)	70 069 (99.8)	87 773 (99.9)	108 415 (100.0)	155 823 (100.0)	168 285 (98.9)	171 654 (73.8)	853 085 (93.1)
Polypharmacy, n (%)	24 747 (71.5)	42 025 (74.1)	52 458 (74.8)	62 327 (70.9)	72 520 (66.9)	95 673 (61.4)	87 676 (51.5)	52 074 (22.4)	487 502 (53.1)
Women, n (%)	17 458 (50.4)	31 444 (55.4)	42 390 (60.4)	66 619 (75.8)	91 266 (84.1)	129 678 (83.2)	6 227 (3.7)	144 047 (61.9)	529 131 (57.7)
Men, n (%)	17 151 (49.6)	25 280 (44.6)	27 788 (39.6)	21 266 (24.2)	17 203 (15.9)	26 182 (16.8)	163 943 (96.3)	88 676 (38.1)	387 488 (42.3)
Age (categories), n (%)									
[65,70)	6 968 (20.1)	9 731 (17.2)	10 239 (14.6)	17 869 (20.3)	25 715 (23.7)	36 946 (23.7)	51 412 (30.2)	92 307 (39.7)	252 178 (27.5)
[70,80)	15 290 (44.2)	23 241 (41.0)	26 372 (37.6)	33 246 (37.8)	44 982 (41.5)	72 562 (46.6)	81 920 (48.2)	96 693 (41.5)	394 586 (43.0)
[80,90)	10 875 (31.4)	20 373 (35.9)	27 952 (39.8)	30 488 (34.7)	32 319 (29.8)	41 430 (26.6)	33 959 (20.0)	38 357 (16.5)	234 744 (25.6)
[90,99]	1 476 (4.3)	3 379 (6.0)	5 615 (8.0)	6 282 (7.1)	5 454 (5.0)	4 922 (3.2)	2 878 (1.7)	5 367 (2.3)	35 111 (3.8)
MEDEA* index									
R	7 199 (20.8)	12 283 (21.7)	16 063 (22.9)	19 200 (21.8)	21 807 (20.1)	32 218 (20.7)	36 483 (21.4)	50 996 (21.9)	182249 (21.4)
U1	5 502 (15.9)	9 073 (16.0)	11 462 (16.3)	15 001 (17.1)	17 925 (16.5)	22 513 (14.4)	27 114 (15.9)	47 040 (20.2)	144791 (17.0)
U2	5 445 (15.7)	8 862 (15.6)	10 921 (15.6)	14 028 (16.0)	17 500 (16.1)	24 185 (15.5)	27 171 (16.0)	38 667 (16.6)	136431 (16.0)
U3	5 642 (16.3)	9 051 (16.0)	11 105 (15.8)	14 065 (16.0)	17 848 (16.5)	26 174 (16.8)	28 023 (16.5)	36 842 (15.8)	138222 (16.2)
U4	5 550 (16.0)	8 930 (15.7)	10 702 (15.2)	13 452 (15.3)	17 525 (16.2)	26 424 (17.0)	27 581 (16.2)	33 017 (14.2)	132988 (15.6)
U5	5 272 (15.2)	8 525 (15.0)	9 926 (14.1)	12 139 (13.8)	15 864 (14.6)	24 346 (15.6)	23 798 (14.0)	26 161 (11.2)	116883 (13.7)
Number of chronic diseases, median [IQR]	8.0 [6.0; 10.0]	8.0 [6.0; 10.0]	8.0 [6.0; 10.0]	7.0 [6.0; 10.0]	7.0 [6.0; 9.0]	6.0 [5.0; 8.0]	5.0 [4.0; 7.0]	3.0 [1.0; 4.0]	6.0 [4.0;8.0]
Number of chronic diseases (categories), n (%)									
0	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	340 (0.2)	24 718 (10.6)	25 380 (2.8)
1	150 (0.4)	140 (0.2)	144 (0.2)	104 (0.1)	46 (0.0)	61 (0.0)	1 747 (1.0)	35 302 (15.2)	38 154 (4.2)
[2, 5)	4 022 (11.6)	5 351 (9.4)	7 343 (10.5)	9 477 (10.8)	10 628 (9.8)	27 127 (17.4)	58 129 (34.2)	144 766 (62.2)	268 836 (29.3)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

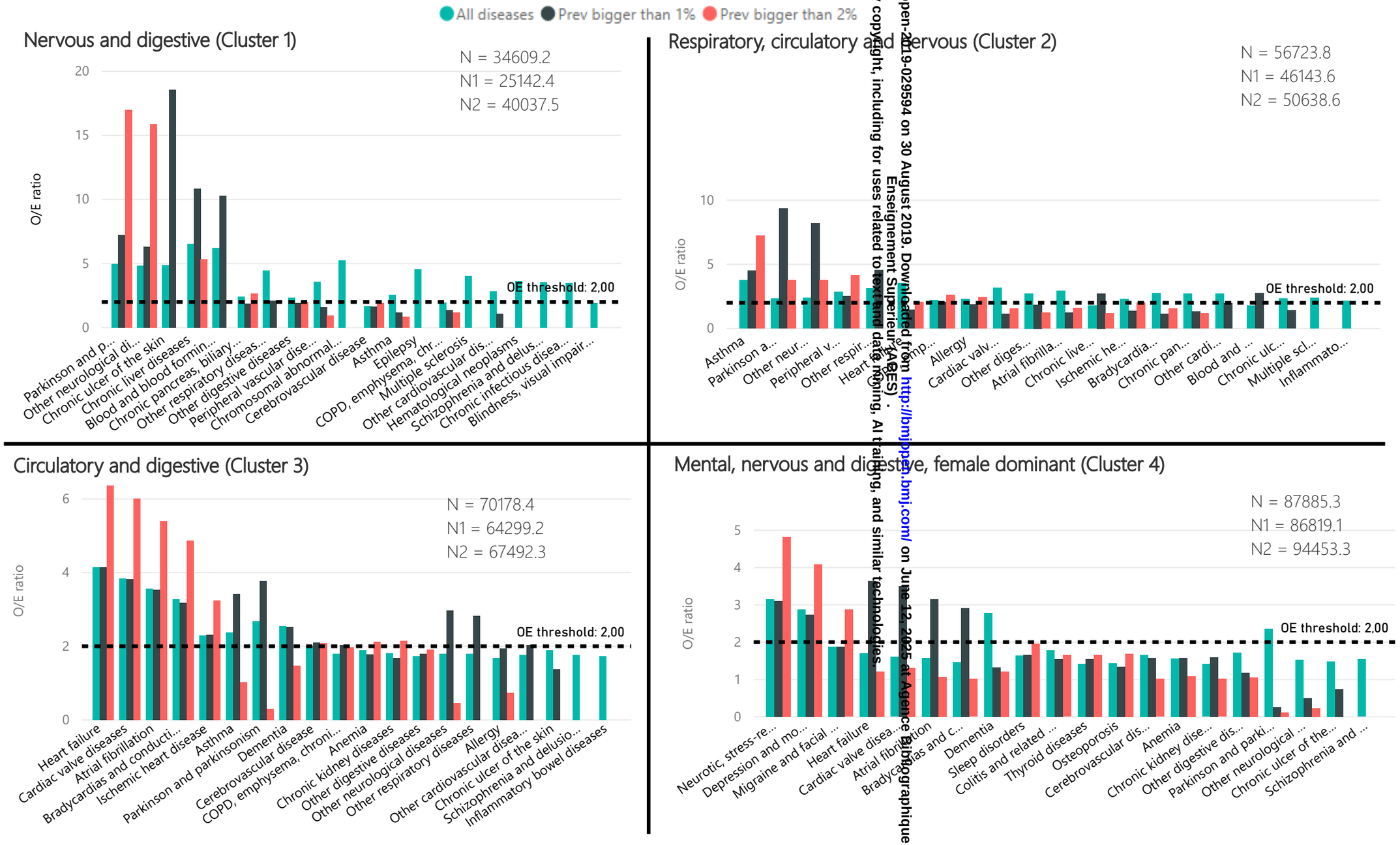
bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES). Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

[5,10)	20 440 (59.1)	32 996 (58.2)	41 917 (59.7)	56 331 (64.1)	74 196 (68.4)	112 073 (71.9)	100 295 (5[illegible].9)	26 838 (11.5)	463 709 (50.6)
≥10	9 997 (28.9)	18 237 (32.2)	20 774 (29.6)	21 973 (25.0)	23 599 (21.8)	16 600 (10.7)	9 659 (5.7[illegible])	1 099 (0.5)	120 540 (13.2)
Number of drugs, median [IQR]	7.0 [4.0; 10.0]	7.0 [4.0; 10.0]	7.0 [4.0; 10.0]	7.0 [4.0; 9.0]	6.0 [4.0; 9.0]	5.0 [3.0; 8.0]	5.0 [3.0; 7[illegible]]	2.0 [0.0; 4.0]	5.0 [2.0;8.0]
Number of drugs (categories)									
0	2 174 (6.3)	3 310 (5.8)	4 049 (5.8)	5 328 (6.1)	6 377 (5.9)	8 768 (5.6)	13 693 (8.[illegible])	68 920 (29.6)	113 368 (12.4)
1	1 052 (3.0)	1 508 (2.7)	1 665 (2.4)	2 406 (2.7)	3 600 (3.3)	6 433 (4.1)	11 557 (6.[illegible])	28 489 (12.2)	57 082 (6.2)
[2, 5)	6 636 (19.2)	9 880 (17.4)	12 006 (17.1)	17 824 (20.3)	25 972 (23.9)	44 986 (28.9)	57 244 (33.[illegible])	83 239 (35.8)	258 667 (28.2)
[5,10)	15 840 (45.8)	26 051 (45.9)	32 957 (47.0)	42 480 (48.3)	52 995 (48.9)	74 918 (48.1)	71 115 (41.[illegible])	47 190 (20.3)	362 681 (39.6)
≥10	8 908 (25.7)	15 974 (28.2)	19 502 (27.8)	19 847 (22.6)	19 525 (18.0)	20 755 (13.3)	16 561 (9.[illegible])	4 885 (2.1)	124 821 (13.6)
Number of visits 2012, median [IQR]	13.0 [7.0; 22.0]	14.0 [8.0; 25.0]	14.0 [8.0; 25.0]	12.0 [7.0; 21.0]	11.0 [7.0; 18.0]	11.0 [7.0; 17.0]	9.0 [5.0; 1[illegible].0]	5.0 [2.0; 9.0]	9.0 [5.0;16.0]
Number of visits 2012 (categories), n (%)									
0	766 (2.2)	1 122 (2.0)	1 435 (2.0)	2 027 (2.3)	2 274 (2.1)	2 771 (1.8)	4 278 (2.5)	32 903 (14.1)	47 945 (5.2)
1	675 (1.9)	959 (1.7)	1 302 (1.9)	1 871 (2.1)	2 089 (1.9)	2 572 (1.7)	4 798 (2.8[illegible])	19 408 (8.3)	33 884 (3.7)
[2, 5)	3 171 (9.2)	4 578 (8.1)	6 024 (8.6)	8 987 (10.2)	11 289 (10.4)	15 678 (10.1)	24 339 (14.[illegible])	55 804 (24.0)	130 439 (14.2)
[5, 10)	7 708 (22.3)	11 373 (20.1)	14 299 (20.4)	20 681 (23.5)	28 386 (26.2)	44 934 (28.8)	52 979 (31.[illegible])	68 497 (29.4)	249 349 (27.2)
≥10	22 289 (64.4)	38 692 (68.2)	47 118 (67.1)	54 320 (61.8)	64 431 (59.4)	89 904 (57.7)	83 776 (49.[illegible])	56 110 (24.1)	455 002 (49.6)

For the sake of simplicity, all numbers in the table were rounded to its closest natural number

For peer review only

bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

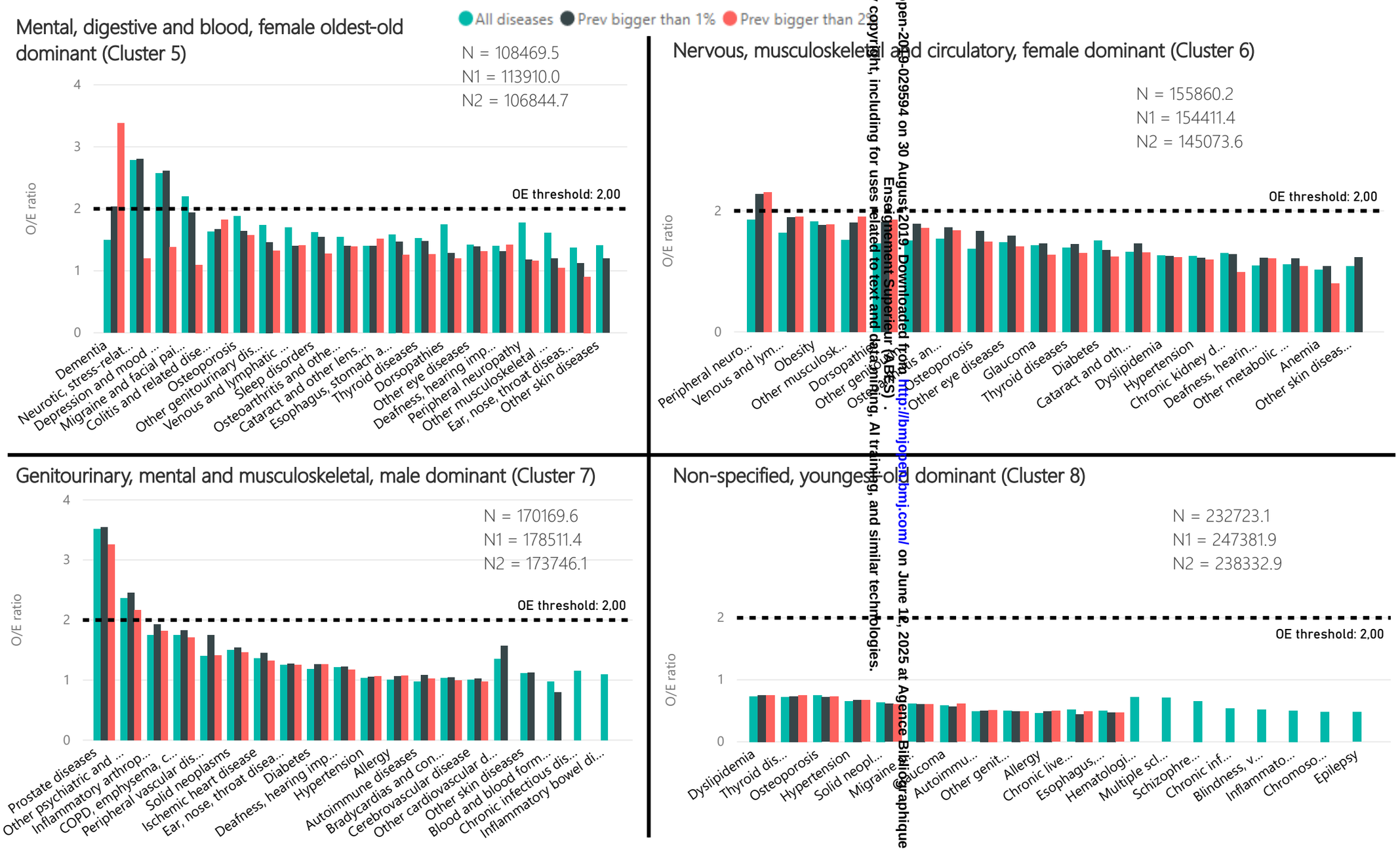


N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied. N for no filtering, N1 for the 1% filter and N2 for the 2% filter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open: first published as 10.1136/bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l'Enseignement Superieur (ABES).



N, N1 and N2 correspond to the number of people in every cluster depending on the prevalence filter applied: N for no filtering, N1 for the 1% filter and N2 for the 2% filter

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.
Enseignement Superieur (ABES) .
bmjopen-2019-029594 on 30 August 2019. Downloaded from http://bmjopen.bmj.com/ on June 12, 2025 at Agence Bibliographique de l

STROBE Statement—Checklist of items that should be included in reports of ***cross-sectional studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	5
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants	5
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*6*
Bias	9	Describe any efforts to address potential sources of bias	7
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	6
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	6
		(*b*) Describe any methods used to examine subgroups and interactions	6
		(*c*) Explain how missing data were addressed	
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	
		(*e*) Describe any sensitivity analyses	7
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	Figure 1
		(c) Consider use of a flow diagram	Figure 1
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Table 1
		(b) Indicate number of participants with missing data for each variable of interest	Tables
Outcome data	15*	Report numbers of outcome events or summary measures	8-9

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	8-9 Tables
		(*b*) Report category boundaries when continuous variables were categorized	
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Additional File 1
Discussion			
Key results	18	Summarise key results with reference to study objectives	10
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	12
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	11
Generalisability	21	Discuss the generalisability (external validity) of the study results	12
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	14

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

For peer review only

Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies. Enseignement Superieur (ABES) .

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml