BMJ Open

Comparative analysis of methods for identifying multimorbidity patterns in a South Mediterranean European Region: a cross-sectional study.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018986
Article Type:	Research
Date Submitted by the Author:	03-Aug-2017
Complete List of Authors:	Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), ; Universitat Autònoma de Barcelona, Violan-Fors, Concepción; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Foguet-Boreu, Quintí; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol) Universitat Autònoma de Barcelona; Universitat Autonoma de Barcelona Rodriguez-Blanco, Teresa; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pons-Vigués, Mariona; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pujol-Ribera, Enriqueta; Institut Universitari d'Investigació en Atenció Primaria (IDIAP-Jordi Gol); Universitat Autonoma de Barcelona Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Research methods, General practice / Family practice
Keywords:	Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Comparative analysis of methods for identifying multimorbidity patterns in a South Mediterranean European Region: a cross-sectional study.

1. Albert Roso-Llorach. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

2. Concepción Violán. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

3. Quintí Foguet-Boreu. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: 42292qfb@comb.cat

4. Teresa Rodriguez-Blanco. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol). Universitat Autònoma de Barcelona, Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. E-mail: trodriguez@idiapjgol.org

5. Mariona Pons-Vigués. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain.
E-mail: mponsv@idiapjgol.org

6. Enriqueta Pujol-Ribera. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain. E-mail: epujol@idiapjgol.org

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

Corresponding author: Albert Roso-Llorach. IDIAP Jordi Gol
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: aroso@idiapjgol.org

Word count: 3,887

Comparative analysis of methods for identifying multimorbidity patterns in a South Mediterranean European Region: a cross-sectional study.

ABSTRACT

Objective. The aim was to compare multimorbidity patterns identified with the two most commonly used methods: hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA) in a large primary care database.

Design Cross-sectional study. Diagnoses were extracted using 263 blocks (ICD-10). Multimorbidity patterns were identified using HCA and EFA. Analysis was stratified by sex, and results compared for each method.

Setting and participants Electronic health records for 408,994 patients with multimorbidity aged 45-64 years in 274 primary health care teams from 2010 in Catalonia, Spain.

Results HCA identified 53 clusters for women, with just 12 clusters including at least 2 diagnoses, and 15 clusters for men, all of them including at least 2 diagnoses. EFA showed 9 factors for women and 10 factors for men. We observed differences by sex and method of analysis, although some patterns were consistent. Three combinations of diseases were observed consistently across sex groups and across both methods: hypertension and obesity, spondylopathies and deforming dorsopathies, and dermatitis eczema and mycosis.

Conclusions This study provides empirical evidence to demonstrate that multimorbidity patterns critically depend on the method of analysis used. The results suggest applications for each method of analysis used and add information about key aspects that must be considered in future studies on multimorbidity patterns.

Keywords: Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Strengths and limitations of this study

- This one of the first studies to compare the two methodologies most commonly used to obtain patterns of multimorbidity, Hierarchical Cluster Analysis and Exploratory Factor Analysis.

- The dual analysis was performed in a large, high-quality database of primary care electronic health records that have been shown to be representative of a much larger population, stratified by sex.

- Internal validation with bootstrap methods provided more robust evidence for the cluster analysis.

- The agglomerative hierarchical clustering forces every unit into a single cluster, is exploratory in nature and different clustering algorithms may produce different results.

- The study is cross-sectional and further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

INTRODUCTION

The reliable identification of patterns of multimorbidity is a critical step in developing health care services sensitive to the health needs of these patients.[1] Recent reviews of multimorbidity patterns have shown that individual studies differ widely in their design and choice of epidemiological and statistical methods, including sampling frameworks and selection criteria, coding systems, eligible diseases, and definition of disease clustering patterns.[2–4] This makes it difficult to draw firm conclusions based on the observations, but it also limits our ability to compare analyses head to head and to evaluate whether one approach may be better suited to the purpose. Therefore, it is difficult to know which are the multimorbidity patterns in order to provide adequate health services according to the population needs.

The most frequent methods used to date have been hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA), which offer very different approaches and solutions.[2,3] Both are descriptive methods to identify association of diagnoses and determine patterns of multimorbidity. HCA obtains the patterns of multimorbidity from the dissimilarities between diseases, while EFA is based on correlations between diagnoses to identify the patterns. The HCA clusters tend to contain diagnoses that are similar to each other (in terms of distances), but dissimilar from the diagnoses in other clusters; no diagnosis can be included in more than one cluster. In contrast, EFA permits inclusion of any diagnosis in more than one factor because there are significant correlations between EFA cluster variables that appear to explain the same factor. In addition, EFA cannot handle binary data properly; these data can be grouped in one factor because the distributions (rather than underlying relationships, as in HCA) are similar. Moreover, the association measure of EFA takes into account both positive and negative matches, while HCA allows for the possibility that one or more health problems can occur conditionally and does not consider the negative matches. After all, we have to bear

in mind that EFA is not designed for clustering purposes and it is essentially used for (visual) exploratory purposes, dimensionality reduction purposes or variables transformation.[5–8]

The purpose of this study was to compare multimorbidity patterns identified by HCA and EFA in adults with multimorbidity aged 45-64 years attended in primary health care in Catalonia (Spain), and stratified by sex.

METHODS

Design, setting and study population

A cross-sectional study was conducted in Catalonia (Spain), a Mediterranean region with 7,434,632 inhabitants, 81% of which live in urban municipalities (2010 census). The Spanish National Health Service (NHS) provides universal coverage, financed mainly by tax revenue. The Catalan Health Institute (CHI) manages primary health care teams (PHCTs) that serve 5,501,784 patients (274 PHCT), or 74% of the population; the remaining PHCTs are managed by other providers. The CHI's Information System for the Development of Research in Primary Care (SIDIAP) contains the coded clinical information recorded in electronic health records EHR by its 274 PHCTs since 2006. A subset of records meeting the highest quality criteria for clinical data (SIDIAP Q) includes 40% of the SIDIAP population (1,833,125 individuals), attended by the 1,365 general practitioners (GPs) whose data recording scores contain information on the majority of the population of Catalonia, and is highly representative for the whole region in terms of geography, age, gender and diseases.[9–11]

Prevalence of individual conditions varies with age and so does multimorbidity and their patterns. In order to obtain a more homogenous sample in terms of multimorbidity, we focussed on individuals aged 45 to 64 years.[12–15] We identified 408,944 individuals aged 45 to 64 years on 31 December 2010 with two or more diagnoses (Appendix 1).

Coding and selection of diseases

Diseases are coded in SIDIAP using International Classification of Diseases version 10 (ICD-10). For this study, we selected all active diagnoses recorded in EHR as of December 31, 2010, except for R codes (symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified) and Z codes (factors influencing health status and contact with health services).[16] Non-active diagnoses were excluded, based on the presence of an end date in the EHR. These diagnoses cover a broad list of acute diseases for which the system automatically assigns an end date (e.g., 60 days after the initial diagnosis).

To facilitate management of the diagnostic information, the diagnoses were extracted using the 263 blocks (disease categories) in the ICD-10 structure. These are homogeneous categories of very closely related specific diagnoses. For example, Hypertensive diseases include Essential (primary) hypertension, Hypertensive heart disease, Hypertensive renal disease, Hypertensive heart and renal disease and Secondary hypertension. To obtain consistent and clinically interpretable patterns of association, and to avoid spurious relationships that could bias the results, we considered only diagnoses with greater than 1% prevalence in each sex. All patients with multimorbidity (2 or more coexisting diagnoses recorded in the EHR on 31 December 2010) were included.

Variables

The variables considered were: diagnosis (values: 1 for present, 0 for absent), number of diseases (2, 3, 4, and 5 or more) and sex (women, men) were also recorded for each patient.

Statistical analysis

Data access: Data was obtained from SIDIAP after the study was authorized. All the project's authors could access the database. Cleaning methods: The analysis was limited to SIDIAP-Q, as the sample was representative of the population.[9–11] No missing values were handled as sex and age were recorded for all population. Wrong sex-specific diagnoses codes and diagnoses

with inconsistent dates were excluded. An individual with no disease diagnoses record was considered as disease free.

Analyses were stratified by sex. Descriptive statistics were used to summarize overall information. Categorical variables were expressed as frequencies (percentage) and continuous as mean (standard deviation, SD) or median (interquartile range, IQR). Two sample test of proportions and Mann-Whitney test were used to test differences by sex.

We identified disease patterns using two approaches: 1) hierarchical cluster analysis (HCA), and 2) exploratory factor analysis (EFA). Clinical criteria were used to evaluate the consistency and utility of the final HCA and EFA solutions, based on previously described patterns in the literature and a consensus opinion drawn from the clinical experience of the research team (4 family physicians, 1 epidemiologist). Clusters and factors in these analyses were considered as two sets of grouping solutions, which were then assigned to each individual patient. We considered patients to be associated with a given grouping solution if they had ≥1 diagnoses in that solution, allowing for the calculation of the prevalence of each solution in the sample. Patients could be associated with more than one solution in the same set. We also calculated prevalence, restricting the assignment of patients to those with ≥ 2 diagnoses in the same solution.

Hierarchical Cluster Analysis

The HCA approach assigns diagnoses to groups or clusters, so that diagnoses in the same cluster are more similar, based on a given measure, to one another than to diagnoses from different clusters. The Jaccard coefficient was used to measure similarity. This coefficient considers only the diagnoses that any two patients have and ignores the diagnoses that neither of them has.[5] As we do not know a priori the number of clusters to retain from the data, we used agglomerative hierarchical methods to identify possible clustering solutions: Average linkage, Ward, flexible beta and other methods with less bias, based on

nonparametric estimates, such as Single Linkage and Density Linkage. All but Ward and the flexible beta methods successively chained the observations into one cluster. Therefore, the Ward method, which minimizes the variance within clusters and produces clusters of similar sizes, was chosen as the primary method based on dendrograms analysis.[5] Data were randomly split into test and training datasets, equal in size and analysed separately. We ran the Ward method on both samples. The semi-partial R2, Calinski-Harabasz pseudo-F- and pseudo-T2-statistic criteria for different numbers of clusters were examined.[5] Clustering solutions were compared between the test and training datasets, taking into account the number of clusters, Adjusted Rand Index and clinical criteria. After checking algorithm stability, Ward method was run on the full data set, applying the same criteria to different numbers of clusters. Results were compared with flexible beta results, with beta values set at -0.25 and -0.5. The criteria for selecting the number of clusters were the highest adjusted Rand index with a high number of clusters and a high pseudo T2 statistic.[5] To assess internal cluster quality, we applied multiscale bootstrap resampling to obtain an approximately unbiased (AU) probability. This probability ('p-value') is the proportion of bootstrapped samples that contain the cluster; larger p-values indicate more support for the cluster.[17]

Multidimensional scaling (MDS) considering two dimensions was used to discover the underlying structure of distance measures between diseases in the cluster analysis. Essentially, MDS assigns observations to specific locations in a conceptual space such that the distances between points in the space match the given dissimilarities as closely as possible. We carried out classical MDS using the distance matrix obtained in the cluster analysis that considered the Jaccard coefficient as a dissimilarity measure. The conceptual map of the diseases distinguishes between the intra-disease cluster and the inter-disease cluster. Taking into account the final cluster's solution and the obtained groups, conceptual maps of the diseases were created. For a better interpretation of the conceptual map, prevalence of the disease was represented as the radius of the circle.[18]

Exploratory Factor Analysis

EFA reduces the observed set of diagnoses to a smaller number of latent factors that account for the correlations between them. As the study variables were dichotomous, the correlation matrix between the diagnoses was computed using tetrachoric correlations. The factorability of the matrix was tested using Bartlett Test of Sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. The extraction of the initial solution was carried out using the principal factors method with squared multiple correlations for the prior communality estimates. The optimal number of extracted factors for the final solution was determined with the Scree plot using the "elbow" rule and setting the percentage of variance equal to 100 percent. Factor loadings were analyzed to identify factor patterns. An oblique rotation, Oblimin, was performed to clarify the factor pattern in order to better interpret the nature of the factors, as we assumed that factors were allowed to be associated with each other. As a rule of thumb, factor loadings greater than or equal to 0.30 in absolute value were considered to be significant.[6]

Comparing multimorbidity patterns

We compared every cluster and factor solutions across sex groups agreement and the diagnoses included in it.

We considered grouping solutions (HCA vs EFA) from different sets (women vs men) to have the following degrees of similarity: a) perfect, when the solution included exactly the same diseases as another solution in the other comparison group (sex or statistical approach); b) partial, when the solution included a subset of diseases present in a solution in the other comparison group; and c) none, when each and every disease in the solution was part of a different solution in the other group and none was part of the same solution. These groups were named using the abbreviation of sex, method and number (For example, MC1: Men Cluster 1)

We further extracted the common subsets of diseases within partially similar solutions, which together with completely similar solutions gave a comprehensive picture of overlapping cluster and factor analysis solutions.

The analyses were performed using SPSS for Windows, version 18 (SPSS Inc., Chicago, IL, USA), SAS 9.2 for Windows (SAS Institute Inc., Cary, NC, USA) and R version 3.1.2 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Of 523,656 patients, 408,994 (78.1%) met the multimorbidity criteria; women had a higher multimorbidity prevalence than men (82.2% vs 73.9%, respectively; p<0.001) (Table 1). Both cluster and factor analyses included 79 diagnoses for women and 73 for men.

Table 1. Number of diseases, clusters, and factors identified in cluster and factorial analysis for patients 45-64 years-old, stratified by sex (N=523,656)

	Total, n (%)	Women, n (%)	Men, n (%)
≥2 Diagnoses*	408,994 (78.1%)	217,823 (82.2%)	191,171 (73.9%)
Number of diagnoses*			
2		26,106 (12.0%)	33,850 (17.7%)
3		28,243 (13.0%)	33,515 (17.5%)
4		28,274 (13.0%)	30,356 (15.9%)
≥5		135,200 (62.1%)	93,450 (48.9%)
Median number of diagnoses (IQR)**		5 (4-8)	4 (3-7)
Number of diagnoses included		79	73
Number of clusters		53	15
Number of clusters with ≥2 diagnoses		12	15
Median of diagnoses per clusters (IQR)***		2 (2-4)	5 (2.5-6)
Number of factors		9	10
Number of factors with ≥2 diagnoses		8	9
Median of diagnoses per factors (IQR)***		5.5 (2.75-7)	4 (4-5)

Abbreviations: IQR, Inter-quartile range.
** Two sample test of proportions; all p-values<0.001*
*** Mann-Whitney test; P<0.001*
****Median of clusters or factors with ≥2 diseases; P<0.001*

Hierarchical Cluster analysis

Using HCA with the Ward method, we obtained 53 clusters for women, with just 12 clusters grouping at least two diagnoses for women and 15 clusters for men (Table 1). We describe only the four most prevalent clusters (Table 2). For a complete description of the clusters and dendrograms, see Appendices 2-5.

Table 2. Four most prevalent clusters, by sex group (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2,%†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	AU p-value**
WOMEN					
WC1^ 82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	0.79 (0.74-0.85)
		F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	
		M70-M79:Other soft tissue disorders	27.0	32.8	
		N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	
		M20-M25:Other joint disorders	18.6	22.6	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	
		D10-D36:Benign neoplasms	16.2	19.7	
		M15-M19:Arthrosis	15.7	19.1	
WC2 55.8	23.0	E70-E90:Metabolic disorders	37.4	63.4	0.93 (0.86-1.00)
		I10-I15:Hypertensive diseases	25.6	45.8	
		E65-E68:Obesity and other hyperalimentation	19.0	34.0	
		E10-E14:Diabetes mellitus	7.7	13.7	
WC3 47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.7	39.4	0.78 (0.73-0.84)
		E00-E07:Disorders of thyroid gland	14.9	31.4	
		F30-F39:Mood [affective] disorders	14.6	30.8	
		M80-M85:Disorders of bone density and structure	11.3	23.9	
WC4 32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	0.71 (0.66-0.77)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	
		L20-L30:Dermatitis and eczema	9.3	28.8	
		B35-B49:Mycoses	5.7	17.8	
MEN					
MC1^^ 83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	0.69 (0.64-0.75)
		F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	
		I10-I15:Hypertensive diseases	32.5	38.8	
		M50-M54:Other dorsopathies	27.8	33.2	
		E65-E68:Obesity and other hyperalimentation	14.6	17.4	

		E10-E14:Diabetes mellitus	14.2	16.9	
MC2 57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	0.87 (0.84-0.90)
		N40-N51:Diseases of male genital organs	12.1	21.0	
		M20-M25:Other joint disorders	12.1	20.9	
		K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	
		K40-K46:Hernia	8.8	15.2	
		D10-D36:Benign neoplasms	8.6	14.9	
		M15-M19:Arthrosis	7.7	13.4	
		K55-K63:Other diseases of intestines	6.4	11.1	
MC3 54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	0.79 (0.74-0.84)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	
		J40-J47:Chronic lower respiratory diseases	9.3	17.2	
		J00-J06:Acute upper respiratory infections	8.9	16.4	
		J30-J39:Other diseases of upper respiratory tract	8.0	14.8	
		L20-L30:Dermatitis and eczema	7.5	13.9	
		G40-G47:Episodic and paroxysmal disorders	7.4	13.7	
		F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	
		F30-F39:Mood [affective] disorders	6.3	11.6	
		B35-B49:Mycoses	4.1	7.6	
MC4 25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	0.87 (0.83-0.91)
		H53-H54:Visual disturbances and blindness	3.9	15.5	
		B00-B09:Viral infections characterized by skin and mucous membrane lesions	3.5	13.9	
		L60-L75:Disorders of skin appendages	3.5	13.9	
		H10-H13:Disorders of conjunctiva	3.0	12.0	
		H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	
		L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	
		L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	
		H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	

*Individuals from the strata ≥1 diagnosis in the cluster/ † Individuals from the strata with ≥2 diagnosis in the cluster/‡Strata: same sex

**Approximately unbiased (AU) probability-value

^ Abbreviation of Sex, method and number (WC1: Women Cluster 1)

^^ Abbreviation of Sex, method and number (MC1: Men Cluster 1)

Twelve clusters with at least two diseases were identified for women, with prevalences ranging from 6.6% to 82.1% (median: 15.5%), (WC3; WC10). The clusters identified in men had prevalences ranging from 3.2% to 83.8% (median 10.1%). The most prevalent cluster included 8 diseases in women (WC1: musculoskeletal, psychiatric, circulatory, gynecological and neoplasms) and 6 in men (MC1: metabolic and circulatory); about half of all patients had at least two diagnoses (52.9% of women and 50.4% of men).

Two clusters were common to men and women, “Spondylopathies” and “Deforming dorsopathies” (WC11, MC13) and “Urolithiasis” and “Other diseases of the urinary systems” (WC9, MC12) (Box 1). The remaining clusters showed partial similarity in men and women, based on six subsets (Box 2), except for three clusters found only in women (WC3, WC7, WC10) and six only in men (MC4, MC5, MC8, MC10, MC14, MC15) (Appendices 4-5).

BOXES

Box 1. Combinations of diseases consistent in both men and women$

Clusters

Complete (whole) clusters

1. M45-M49:Spondylopathies*
   M40-M43:Deforming dorsopathies (WC11;MC13)#
2. N20-N23:Urolithiasis
   N30-N39:Other diseases of urinary system (WC9; MC12)

Subsets within clusters

1. E65-E68:Obesity and other hyperalimentation
   10-I15:Hypertensive diseases
   E10-E14:Diabetes mellitus (WC2; MC1)
2. M15-M19:Arthrosis
   M20-M25:Other joint disorders
   I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
   M70-M79:Other soft tissue disorders
   D10-D36:Benign neoplasms (WC1; MC2)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   K00-K14:Diseases of oral cavity, salivary glands and jaws
   J00-J06:Acute upper respiratory infections (WC4; MC3)
4. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12; MC7)
5. J30-J39:Other diseases of upper respiratory tract
   J40-J47:Chronic lower respiratory diseases (WC6; MC3)
6. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; MC2)
7. G50-G59:Nerve, nerve root and plexus disorders
   M65-M68:Disorders of synovium and tendon (WC8; MC6)

Factors*

Subgroups within factors

1. I10-I15:Hypertensive diseases
   I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I70-I79:Diseases of arteries, arterioles and capillaries (WF3; MF2)
2. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation (WF2;MF1)
3. J00-J06:Acute upper respiratory infections
   J20-J22:Other acute lower respiratory infections
   J09-J18:Influenza and pneumonia
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (WF4; MF6)
4. M15-M19:Arthrosis
   M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies
   M50-M54:Other dorsopathies (WF1;MF4)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   Q38-Q45:Other congenital malformations of the digestive system (WF6;MF8)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   H53-H54:Visual disturbances and blindness
   H10-H13:Disorders of conjunctiva
   L80-L99:Other disorders of the skin and subcutaneous tissue (WF5;MF5)
7. H25-H28:Disorders of lens
   H30-H36:Disorders of choroid and retina (WF3; MF7)

** Coincident disease in both sexes*

 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)

$ No two full factors were exactly the same for both sexes.

Underlined blocks of diagnosis represent coincident diseases in pattern.

Box 2. Combinations of diseases consistent across statistical methods (cluster and factor analysis)$

Women

1. I10-I15:Hypertensive diseases*
   E65-E68:Obesity and other hyperalimentation
   E10-E14:Diabetes mellitus (WC2; WF2)#
2. M15-M19:Arthrosis
   M50-M54:Other dorsopathies
   M70-M79:Other soft tissue disorders (WC1; WF1)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (WC4; WF5)
4. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (WC11; WF1)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; WF6)
6. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12 ;WF6)

Men

1. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation
   E70-E90:Metabolic disorders (MC1; MF1)
2. I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I60-I69:Cerebrovascular diseases
   I70-I79:Diseases of arteries, arterioles and capillaries
   N17-N19:Renal failure (MC5; MF2)
3. J09-J18:Influenza and pneumonia
   J20-J22:Other acute lower respiratory infections
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (MC10; MF6)
4. H10-H13:Disorders of conjunctiva
   H53-H54:Visual disturbances and blindness
   L80-L99:Other disorders of the skin and subcutaneous tissue (MC4; MF5)
5. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (MC13; MF4)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (MC3; MF5)
7. K70-K77:Diseases of liver
   B15-B19:Viral hepatitis (MC7; MF3)
8. T08-T14:Injuries to unspecified part of trunk, limb or body region
   S90-S99:Injuries to the ankle and foot (MC8; MF9)
9. H25-H28:Disorders of lens
   H40-H42:Glaucoma (MC14; MF7)

** Coincident disease in both methods*

 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)$ All subgroups of factors or clusters, no single cluster exactly the same as a factor.

Underlined blocks of diagnosis represent coincident diseases in pattern.

The top four clusters were reproduced in the graphical representation of coordinates using MDS. The most prevalent disease clusters were more clearly separated in women than they were in men, mostly due to the overlap of MC2 and MC3 (Figure 1).

Exploratory Factor Analysis

Using EFA, we obtained nine factors for women and 10 factors for men. In this analysis, the median number of diagnoses per factor was higher in women (Table 1).Two factors explained more than 50% of total variance (58.5% for women (WF2; WF3) and 50.7% for men (MF1; MF2)). All diseases were assigned to single factors except for three diseases, two in women only (J20-J22: Other acute lower respiratory infections and E10-E14: Diabetes mellitus) and one in both women and men (I10-I15: Hypertensive diseases); all three were assigned to two factors. The first four factors are described in Table 3; full factor solutions are shown in Appendices 6-7.

Table 3. Four most prevalent factors, by sex (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses**	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WOMEN						
WF1^ 59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
		M70-M79:Other soft tissue disorders	27.0	45.2		
		M15-M19:Arthrosis	15.7	26.2		
		G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
		M45-M49:Spondylopathies	4.3	7.3		
		M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
		E65-E68:Obesity and other hyperalimentation	19.0	50.2		
		E10-E14:Diabetes mellitus	7.7	20.3		
WF3 32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
		E10-E14:Diabetes mellitus	7.7	23.4		
		I30-I52:Other forms of heart disease	4.4	13.3		
		H25-H28:Disorders of lens	1.7	5.3		
		H30-H36:Disorders of choroid and retina	1.2	3.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
		I20-I25:Ischaemic heart diseases	1.0	3.1		

WF4 27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
		N30-N39:Other diseases of urinary system	5.9	21.3		
		H60-H62:Diseases of external ear	3.6	13.1		
		J20-J22:Other acute lower respiratory infections	3.4	12.2		
		A00-A09:Intestinal infectious diseases	2.7	10.0		
		H65-H75:Diseases of middle ear and mastoid	2.5	9.2		
		J09-J18:Influenza and pneumonia	1.7	6.1		
		B25-B34:Other viral diseases	1.3	4.8		
		M60-M63:Disorders of muscles	1.2	4.4		
MEN						
MF1^^ 61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
		I10-I15:Hypertensive diseases	32.6	52.7		
		E65-E68:Obesity and other hyperalimentation	14.6	23.6		
		M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
		I30-I52:Other forms of heart disease	6.9	17.6		
		I20-I25:Ischaemic heart diseases	5.0	12.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
		I60-I69:Cerebrovascular diseases	1.8	4.6		
		N17-N19:Renal failure	1.5	3.7		
MF3 38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
		K70-K77:Diseases of liver	5.2	13.6		
		B15-B19:Viral hepatitis	3.2	8.4		
		F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
		M15-M19:Arthrosis	7.7	22.2		
		M45-M49:Spondylopathies	3.1	8.8		
		M40-M43:Deforming dorsopathies	1.8	5.2		

**Individuals from the strata ≥1 diagnosis in the factor/ † Individuals from the strata with ≥2 diagnosis in the factor/‡Strata: same sex*

***KMO sampling adequacy index was 0.82 for women and 0.75 for men. Bartlett Test of Sphericity was statistically significant. (p<0.001) for both groups*

^ Abbreviation of Sex, method and number (WF1: Women Factor 1)

^^ Abbreviation of Sex, method and number (MC1: Men Factor 1)

Although no factor-based groupings were identical in men and women, almost all showed partial similarity by sex, based on seven subsets (Box 1), except for two groups found only in women (WF7, WF9) and one found only in men (MF9).

Multimorbidity patterns comparison across statistical approaches

The EFA multimorbidity patterns were more easily interpreted than the HCA groups, either because they made more sense from a clinical perspective or because of greater homogeneity in the diagnoses: about half of the factors containing at least 3 diagnoses corresponded to a maximum of 2 ICD-10 blocks, compared to about one fifth of the HCA clusters with at least 3 diseases (1/5 for women and 2/11 for men). No grouping solution for women or for men contained exactly the same set of diseases as a cluster and as a factor. Six clusters (WC3, WC6, WC7, WC8, WC9, WC10) and two factors (WF7, WF9) were only observed in women. However, six subsets of diseases were part of the same grouping in both a cluster and a factor (Box 2); all included two or three diagnoses, usually from the same ICD chapter. Five clusters and one factor were observed only in men (MC6, MC9, MC11, MC12, MC15 and MF6). Nine subsets of diseases were observed as part of the same grouping in both a cluster and a factor (Box 2). They included a range of diseases (2-5) and most frequently included diseases from different ICD chapters.

Three paired diseases were observed consistently in both men and women using both methods of analysis: 1) Hypertensive diseases and obesity/other hyperalimentation; 2) spondylopathies and deforming dorsopathies; and 3) dermatitis/eczema and mycoses.

DISCUSSION

In this study have observed differences in the groupings identified with the two most frequently used methods (HCA and EFA). No grouping solution contained exactly the same set of diseases as a cluster and as a factor, although some overlap was observed in both groups. Internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

The multimorbidity patterns obtained by HCA were identified graphically with MDS, allowing us to observe a given hierarchical structure. Nevertheless, internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

EFA, based on tetrachoric correlations where the dichotomous diseases were assumed to come from an underlying mechanism with a continuous variable, produced a wide range of multimorbidity patterns with several levels of correlations. Most of them seem to be highly consistent from a clinical perspective.

The HCA results would be useful in generating new hypotheses for intercluster and intracluster associations between diseases that could be applied to the analysis of multimorbidity, defined as the random coexistence of diseases or clusters that indicates significant associations between diseases without a causal explanation. In future studies, other non-hierarchical cluster analysis techniques will improve measurement of the observed distances and multiple interrelationships between different diseases in a given individual.[19] On the other hand, EFA could be more useful for analyzing multimorbidity patterns in the absence of causal comorbidity and for describing visual representation of diseases correlation with a pathophysiological relationship between them.

We obtained two perfect clusters that were common to both men and women: “spondylopathies and deforming dorsopathies” and “urolithiasis and other diseases of the urinary system”. In the first cluster, spondylosis is a degenerative disorder that may cause loss

of normal spinal structure and function and lead to scoliosis. Nevertheless, many individuals with untreated scoliosis will develop spondylosis; this may be one reason why these diseases were associated.[20] The second cluster can be explained by the complications produced by urolithiasis (such as urinary tract infection, persistent proteinuria, stress incontinence or other unspecified urinary incontinence) and those that have a pathophysiological explanation.[21]

EFA showed that the most frequent pattern in women was infectious diseases. This previously unreported pattern suggests that the multimorbidity patterns obtained in other studies are affected by the type of diseases included in each study.

Although the patterns obtained with both methods did not match exactly, finding matching pairs of diseases by both methods reinforces the idea that patterns of multimorbidity have a dominant disease that associates in some way with other diseases.

In general, it is difficult to compare our results with other studies because of variations in methods, data sources and structures, and populations and diseases studied. Six studies have been performed with HCA[7,19,22–25] and three using EFA.[26–28] Until now, very few analyses of multimorbidity patterns have used multiple methods to compare the same population.[19] The latter study included people aged 50 years and older, considering 11 diseases and using 2 different cluster methods, hierarchical (average linkage) and nonhierarchical (k-medoids), and one method for EFA (principal component analysis). The observed differences between this study and our results can be explained by differences in the underlying statistical formulae and diseases considered in both studies.

The major strength of this study is the dual analysis (HCA and EFA) of a large, high-quality database of primary care records that have been shown to be representative of a much larger population, stratified by sex. Admittedly, this analysis of almost all potential diagnoses may have added a complexity that will hinder interpretation of findings and comparison with other studies, particularly because the boundaries between chronic and acute disease are not always clear.[29,30] Whatever consistency (or discrepancy) we observed was validated by the findings

of two different approaches, which helps to identify the most appropriate use of each method in analyzing multimorbidity. We emphasize that the inclusion of chronic and acute diseases is a strength and not a weakness. Because, as we have shown, there are many chronic and acute diseases that coexist at a set time and this has implications for health care.

Internal validation with bootstrap methods, AU p-values, and MDS techniques provided more robust evidence for the cluster analysis. The KMO values obtained show the adequacy of fit of the factor analysis. These values were similar or higher than previous studies.[26,27]

A limitation of this study is our use of agglomerative hierarchical clustering, which forces every unit (i.e., diagnosis) into a single cluster. HCA is exploratory in nature, and different clustering algorithms may produce different results.[31] The final clustering solution presented here was obtained through a systematic and rigorous process: comparing the results from a randomly split dataset, testing different clustering algorithms, and using different objective numeric criteria to decide the number of clusters, internal validation, and graphical representation. In addition, a panel of experts applied subjective clinical criteria to assess the interpretability of the groupings in everyday practice. Due to the absence of a standard methodology to compare method solutions we have used ad hoc methodology. Finally, another limitation is our use of ICD-10 3-character codes as the unit of analysis, rather than the more specific individual diagnosis, but its use is justified to avoid spurious relationships that more than 10,000 individual codes of the ICD-10 could produce.

This is a cross-sectional study, based on EHR of The Catalan Health Institute and SIDIAP-Q is highly representative for the whole region in terms of both geography, age, gender and diseases, that avoid selection bias.

Multimorbidity can present a problem for health services delivery, affecting patients, health professionals and managers who are attempting to improve service delivery. Our study offers two methodological approaches to understanding the relationships between specific diseases, which is an essential step in improving our approach to this problem. Although we

demonstrated that different analytical methods can yield different results, we also showed that some associations were consistent in both analyses. This study illustrates the need to pay careful attention to the methods used to support policies and decision-making. Clinical guidelines tend to focus on a single disease rather than on multimorbidity, which includes not only diseases but also drug interactions and polypharmacy. The present study confirmed that multimorbidity patterns are a reality in the adult population, and do not apply only to chronic diseases. New guidelines are needed that incorporate multimorbidity into clinical recommendations.

This study was one of the first to compare the two most commonly used methodologies, HCA and EFA, in a large database that includes a large number of diseases. The findings reveal another limitation to be taken into account in comparing multimorbidity patterns between studies: in addition to the spectrum, number and type of diseases included, these patterns vary depending on the method of analysis used.

The results suggest that HCA can be useful to detect multimorbidity patterns and identify different associations between diseases, as the method allows for the possibility that one or more health problems can occur conditionally. On the other hand, EFA seems more applicable to clinical practice because places less restrictions in the diseases grouping, so may be better for generating hypotheses and is more sensitive in identifying clinical associations. Our results suggest that these aspects be considered in planning of future studies, including selection of diseases and methods of analysis.

Finally, our analysis of multimorbidity patterns only considered associations between diseases. Further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

Conclusions

This study offers empirical evidence to demonstrate that multimorbidity patterns critically depend on the method of analysis employed. The results suggest applications for each method of analysis used and add information about key aspects that must be considered in future studies on multimorbidity patterns.

Footnotes

Contributors: All authors contributed to the design of the study, revised the article, and approved the final version. CV and QFB obtained the funding. ARL and CV drafted the article. ARL, CV, QFB, TRB, MPV, EPR and JMV contributed to the analysis and interpretation of data. ARL and CV wrote the first draft, and all authors (ARL, CV, QFB, TRB, MPV, EPR and JMV) contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript. All authors read and approved the final manuscript.

Funding: The project has been funded by the *Instituto de Salud Carlos III* of the Ministry of Economy and Competitiveness (Spain) through the Network for Prevention and Health Promotion in Primary Health Care (redIAPP, RD12/0005), by a grant for research projects on health from ISCiii (PI12/00427) and co-financed with European Union ERDF funds). Jose M. Valderas was supported by the National Institute for Health Research Clinician Scientist Award NIHR/CS/010/024. The funders had no role in the study design, collection, analysis and interpretation of data, writing of the manuscript or decision to submit for publication. 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.

Ethical considerations: The study protocol was approved by the Committee on the Ethics of Clinical Research, IDIAP Jordi Gol (Protocol No: P12/28). All data were anonymized and EHR confidentiality was respected in accordance with national and international law.

Acknowledgment: We thank the Catalan Health Institute and especially the SIDIAP Unit, which provided the database for the study. The authors also appreciate the English language review by Elaine Lilly, PhD, and are grateful to Carmen Ibáñez for administrative work.

Availability of data and material: 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 organization is subject to periodic audits to ensure the validity and quality of the data.

References

1. Valderas J, Starfield B, Sibbald B, et al. Defining comorbidity: Implications for understanding health and health services. *Ann Fam Med* 2009;**7**: 357–363.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, et al. Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PLoS One* 2014;**9**: e102149.

3. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, et al. Multimorbidity patterns: A systematic review. *J Clin Epidemiol* 2014;**67**: 254–266.

4. Fortin M, Stewart M, Poitras ME, et al. A systematic review of prevalence studies on multimorbidity: Toward a more uniform methodology. *Ann Fam Med* 2012;**10**: 142–151.

5. Everitt B, Landau S, Leese M, et al. Cluster Analysis. 5th ed. Hoboken: Wiley; 2011.

6. Thompson B. Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington, DC: American Psychological Association; 2004.

7. Cornell JE, Pugh JA, Williams J, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters : clustering binary data from a large administrative medical database. *Appl Multivar Res* 2007;**12**: 163–182.

8. Nunnally J, Berstein I. Psychometric theory. 3rd ed. New York: McGraw-Hill, Inc; 1994.

9. Del Mar Garcia-Gil M, Hermosilla E, Prieto-Alhambra D, 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**: 135–145.

10. Prieto-Alhambra D, Judge A, Javaid MK, et al. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. *Ann Rheum Dis* 2014;**73**: 1659–64.

11. Ramos R, Balló E, Marrugat J, et al. Validity for use in research on vascular diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA

study. *Rev Esp Cardiol* (Engl Ed) 2012;**65**: 29–37.

12. Foguet-Boreu Q, Violán C, Rodriguez-Blanco T, et al. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. *PLoS One* 2015;**10**: e0141155.

13. Foguet-Boreu Q, Violan C, Roso-Llorach A, et al. Impact of multimorbidity: acute morbidity, area of residency and use of health services across the life span in a region of south Europe. *BMC Fam Pract* 2014;**15**: 55.

14. Violán C, Foguet-Boreu Q, Roso-Llorach A, 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**: 1–13.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, et al. Patrones de multimorbilidad en adultos jóvenes en Cataluña: un análisis de clústeres. *Aten Primaria* 2016; **48**:479-92.

16. World Health Organization. ICD-10 International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2007.

17. Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling 32: *Ann Stat* 2004;**32**: 2616–2641.

18. Borg I. Modern Multidimensional Scaling: theory and applications (2nd ed.). 2 nd. Springer-Verlag, editor. New York; 2005.

19. Islam M, Valderas J, Yen L, et al. Multimorbidity and comorbidity of chronic diseases among the senior Australians: prevalence and patterns. *PLoS One* 2014;**9**: e83783.

20. Pappou IP, Girardi FP, Sandhu HS, et al. Discordantly high spinal bone mineral density values in patients with adult lumbar scoliosis. *Spine* (Phila Pa 1976). 2006;**31**: 1614–1620.

21. Bartoletti R, Cai T, Mondaini N, et al. Epidemiology and risk factors in urolithiasis. *Urol Int* 2007;**79** Suppl 1: 3–7.

22. Formiga F, Ferrer A, Sanz H, et al. Patterns of comorbidity and multimorbidity in the oldest old: The Octabaix study. *Eur J Intern Med* 2013;**24**: 40–44.

23. John R, Kerby DS and Hennessy CH. Patterns and impact of comorbidity and multimorbidity among community-resident American Indian elders. *Gerontologist* 2003;**43**: 649–660.

24. Marengoni A, Rizzuto D, Wang HX, et al. Patterns of chronic multimorbidity in the elderly population. *J Am Geriatr Soc* 2009;**57**: 225–230.

25. Newcomer SR, Steiner JF and Bayliss EA. Identifying subgroups of complex patients with cluster analysis. *Am J Manag Care* 2011;**17**: e324–32.

26. Schäfer I, von Leitner EC, Schön G, et al. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS One* 2010;**5**: e15941.

27. Andre L, Prados-torres A, Poblador-plou B, et al. Multimorbidity Patterns in Primary Care : Interactions among Chronic Diseases Using Factor Analysis. *PLoS One* 2012;**7**(2): e32190.

28. Poblador-Plou B, van den Akker M, Vos R, et al. Similar multimorbidity patterns in primary care patients from two European regions: results of a factor analysis. *PLoS One* 2014;**9**: e100375.

29. O'Halloran J, Miller GC and Britt H. Defining chronic conditions for primary care with ICPC-2. *Fam Pract* 2004;**21**: 381–386.

30. Lamberts H and Wood M. International Classification of Primary Care. Oxford. Press OU, editor. Oxford; 2011.

31. Aldenderfer MS and Blashfield RK. Cluster Analysis: Quantitative Applications in the Social Sciences. 1984.

Figure 1a. Top four clusters in women (n= 217,823) aged 45-64 years, analyzed with multidimensional scaling

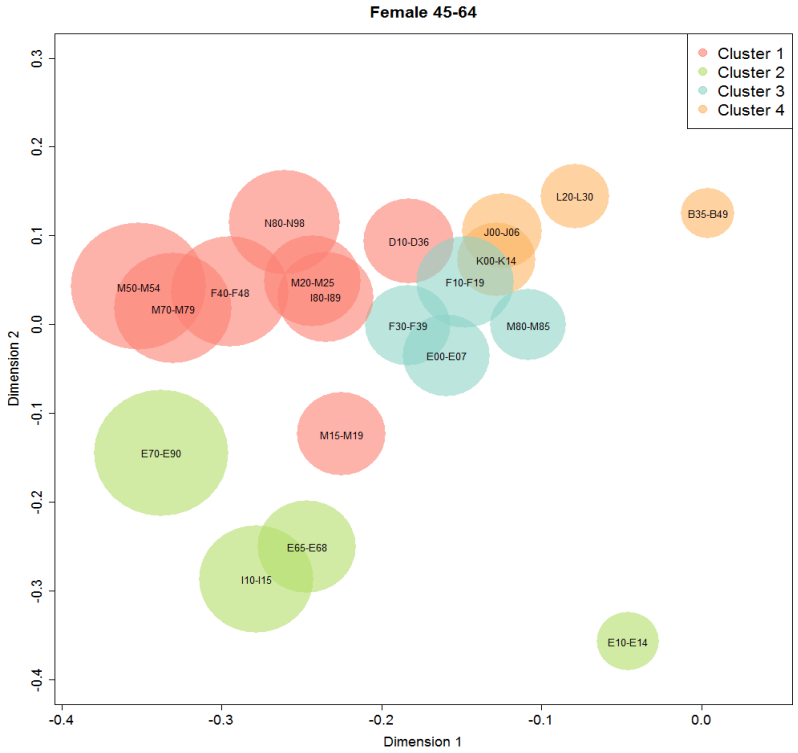
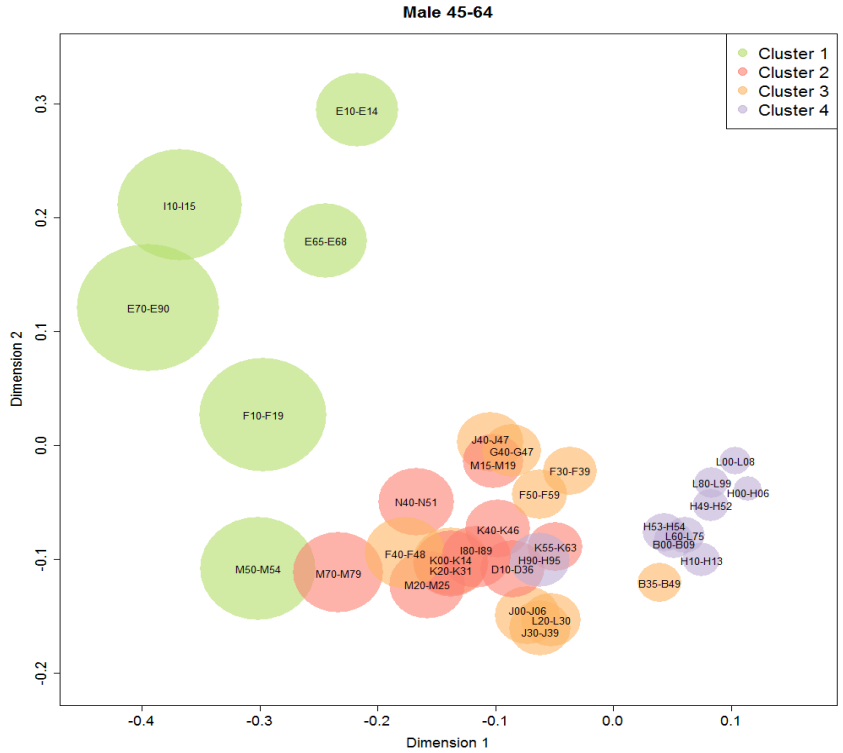


Figure 1 b. Top four clusters in men (n= 191,171) aged 45-64 years, analyzed with multidimensional scaling



M50-M54:Other dorsopathies
F40-F48:Neurotic, stress-related and somatoform disorder
M70-M79:Other soft tissue disorders
N80-N98:Noninflammatory disorders of female genital tract
M20-M25:Other joint disorders
I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
D10-D36:Benign neoplasms
M15-M19:Arthrosis

Cluster 2
E70-E90:Metabolic disorders
I10-I15:Hypertensive diseases
E65-E68:Obesity and other hyperalimentation
E10-E14:Diabetes mellitus

Cluster 3
F10-F19:Mental and behavioural disorders due to psychoactive substance use
E00-E07:Disorders of thyroid gland
F30-F39:Mood [affective] disorders
M80-M85:Disorders of bone density and structure

Cluster 4
J00-J06:Acute upper respiratory infections
K00-K14:Diseases of oral cavity, salivary glands and jaws
L20-L30:Dermatitis and eczema
B35-B49:Mycoses

Cluster 1
E70-E90:Metabolic disorders
F10-F19:Mental and behavioural disorders due to psychoactive substance use

I10-I15:Hypertensive diseases

M50-M54:Other dorsopathies
E65-E68:Obesity and other hyperalimentation

E10-E14:Diabetes mellitus

Cluster 3
F40-F48:Neurotic, stress-related and somatoform disorders
K00-K14:Diseases of oral cavity, salivary glands and jaws
J40-J47:Chronic lower respiratory diseases
J00-J06:Acute upper respiratory infections
J30-J39:Other diseases of upper respiratory tract
L20-L30:Dermatitis and eczema
G40-G47:Episodic and paroxysmal disorders

F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors
F30-F39:Mood [affective] disorders
B35-B49:Mycoses

Cluster 2
M70-M79:Other soft tissue disorders
N40-N51:Diseases of male genital organs
M20-M25:Other joint disorders

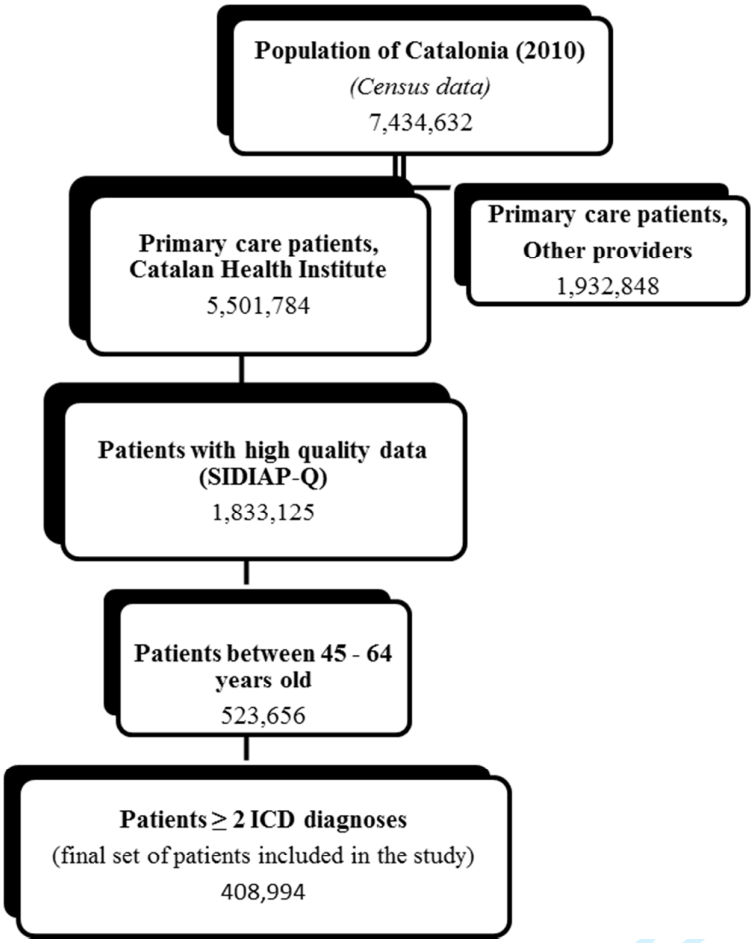
K20-K31:Diseases of oesophagus, stomach and duodenum
I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
K40-K46:Hernia
D10-D36:Benign neoplasms
M15-M19:Arthrosis
K55-K63:Other diseases of intestines
Cluster 4
H90-H95:Other disorders of ear
H53-H54:Visual disturbances and blindness
B00-B09:Viral infections characterized by skin and mucous membrane lesions
L60-L75:Disorders of skin appendages
H10-H13:Disorders of conjunctiva
H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction

L80-L99:Other disorders of the skin and subcutaneous tissue

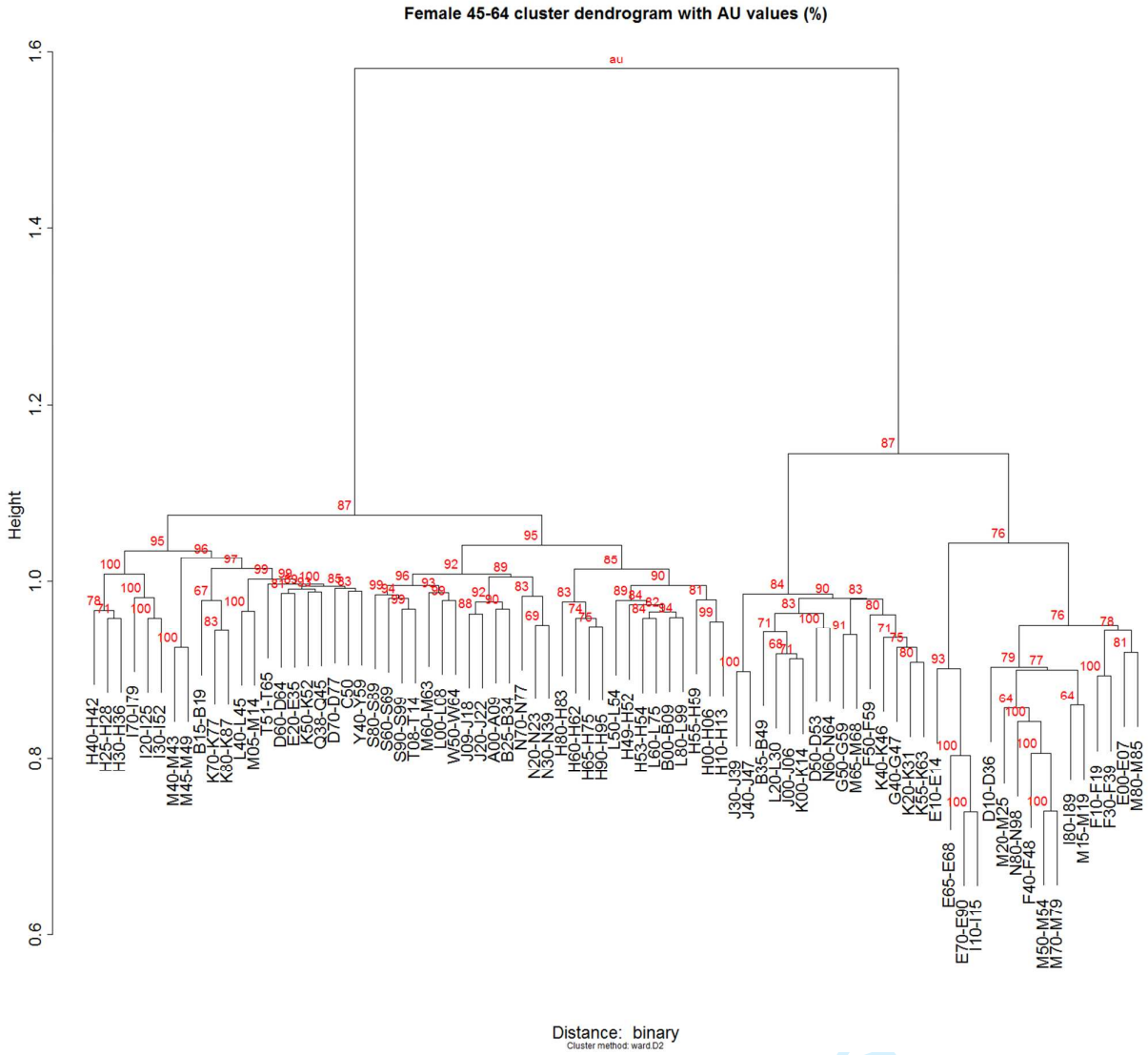
L00-L08:Infections of the skin and subcutaneous tissue
H00-H06:Disorders of eyelid, lacrimal system and orbit

Note: The cluster number are the same as expressed in table 2.

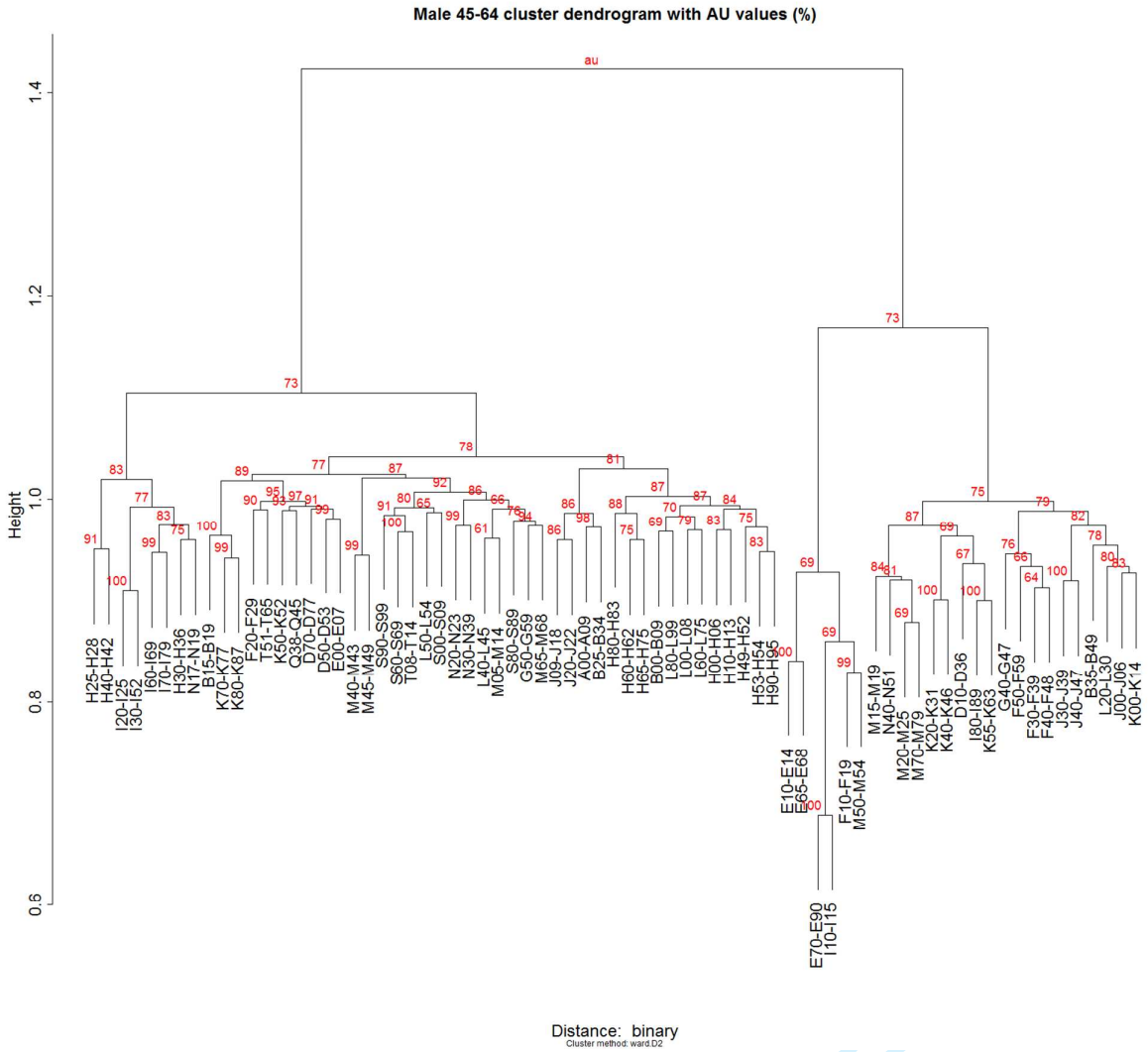
Appendix 1. Study Flow Chart.



Appendix 2. Cluster dendrogram in women aged 45-64 years.



Appendix 3. Cluster dendrogram in men aged 45-64 years, with approximately unbiased (AU) probability value.



Appendix 4. Prevalence and composition of diagnostic clusters in women aged 45-64 years (n= 217,823).

Cluster order number, n	Prevalence 1 (%)*	Prevalence 2 (%)†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value**
WC1^ 178,849	82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	M54 Dorsalgia	0.79 (0.74-0.85)
			F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	F41 Other anxiety disorders	
			M70-M79:Other soft tissue disorders	27.0	32.8	M79 Other soft tissue disorders, not elsewhere classified	
			N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	N95 Menopausal and other perimenopausal disorders	
			M20-M25:Other joint disorders	18.6	22.6	M25 Other joint disorders, not elsewhere classified	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	I83 Varicose veins of lower extremities	
			D10-D36:Benign neoplasms	16.2	19.7	D25 Leiomyoma of uterus	
			M15-M19:Arthrosis	15.7	19.1	M17 Gonarthrosis [arthrosis of knee]	
WC2 121,564	55.8	23.0	E70-E90:Metabolic disorders	35.4	63.4	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.93 (0.86-1.00)
			I10-I15:Hypertensive diseases	25.6	45.8	I10 Essential (primary) hypertension	
			E65-E68:Obesity and other hyperalimentation	19.0	34.0	E66 Obesity	
			E10-E14:Diabetes mellitus	7.7	13.7	E11 Non-insulin-	

						dependent diabetes mellitus	
WC3 103,147	47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	39.4	F17 Mental and behavioural disorders due to use of tobacco	0.78 (0.73-0.84)
			E00-E07:Disorders of thyroid gland	14.9	31.4	E03 Other hypothyroidism	
			F30-F39:Mood [affective] disorders	14.6	30.8	F32 Depressive episode	
			M80-M85:Disorders of bone density and structure	11.3	23.9	M81 Osteoporosis without pathological fracture	
WC4 70,294	32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	J00 Acute nasopharyngitis [common cold]	0.71 (0.66-0.77)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	K02 Dental caries	
			L20-L30:Dermatitis and eczema	9.3	28.8	L29 Pruritus	
			B35-B49:Mycoses	5.7	17.8	B35 Dermatophytosis	
WC5 63,792	29.3	6.0	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	38.9	K30 Dyspepsia	0.71 (0.66-0.76)
			G40-G47:Episodic and paroxysmal disorders	10.5	36.0	G43 Migraine	
			K55-K63:Other diseases of intestines	8.6	29.2	K59 Other functional intestinal disorders	
			K40-K46:Hernia	5.7	19.4	K44 Diaphragmatic hernia	
WC6 34,707	15.9	1.6	J30-J39:Other diseases of upper respiratory tract	9.4	59.0	J30 Vasomotor and allergic rhinitis	1.00
			J40-J47:Chronic lower respiratory diseases	8.2	51.2	J45 Asthma	
WC7 32,674	15.0	0.8	D50-D53:Nutritional anaemias	8.3	55.2	D50 Iron deficiency anaemia	1.00
			N60-N64:Disorders of breast	7.5	50.1	N60 Benign mammary	

						dysplasia	
WC8 27,150	12.5	0.8	G50-G59:Nerve, nerve root and plexus disorders	8.5	68.6	G56 Mononeuropathies of upper limb	0.91 (0.88-0.94)
			M65-M68:Disorders of synovium and tendon	4.7	37.5	M65 Synovitis and tenosynovitis	
WC9 19,199	8.8	0.4	N30-N39:Other diseases of urinary system	5.9	66.7	N39 Other disorders of urinary system	0.69 (0.62-0.75)
			N20-N23:Urolithiasis	3.4	38.3	N20 Calculus of kidney and ureter	
WC10 18,439	8.5	0.4	H90-H95:Other disorders of ear	6.3	75.0	H91 Other hearing loss	0.75 (0.70-0.80)
			H65-H75:Diseases of middle ear and mastoid	2.5	30.1	H65 Nonsuppurative otitis media	
WC11 16,535	7.6	0.6	M45-M49:Spondylopathies	4.3	57.1	M47 Spondylosis	1.00
			M40-M43:Deforming dorsopathies	3.8	50.4	M41 Scoliosis	
WC12 14, 335	6.6	0.4	K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	60.9	K80 Cholelithiasis	0.83*
			K70-K77:Diseases of liver	2.9	44.7	K76 Other diseases of liver	
WC13 12,320	5.7	0.0	F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	5.7	100.0	F51 Nonorganic sleep disorders	
WC14 10,357	4.8	0.0	L60-L75:Disorders of skin appendages	4.8	100.0	L72 Follicular cysts of skin and subcutaneous tissue	
WC15 9,555	4.4	0.0	H53-H54:Visual disturbances and blindness	4.4	100.0	H54 Blindness and low vision	
WC16 9,521	4.4	0.0	I30-I52:Other forms of heart disease	4.4	100.0	I34 Nonrheumatic mitral valve disorders	
WC17 9,442	4.3	0.0	B00-B09:Viral infections characterized by skin and mucous membrane lesions	4.3	100.0	B07 Viral warts	
WC18 8,541	3.9	0.0	T08-T14:Injuries to unspecified part of trunk, limb or body region	3.9	100.0	T14 Injury of unspecified body region	

WC19 8,268	3.8	0.0	H10-H13:Disorders of conjunctiva	3.8	100.0	H10 Conjunctivitis	
WC20 7,866	3.6	0.0	H60-H62:Diseases of external ear	3.6	100.0	H61 Other disorders of external ear	
WC21 7,714	3.5	0.0	H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	3.5	100.0	H52 Disorders of refraction and accommodation	
WC22 7,512	3.4	0.0	L50-L54:Urticaria and erythema	3.4	100.0	L50 Urticaria	
WC23 7,298	3.4	0.0	J20-J22:Other acute lower respiratory infections	3.4	100.0	J20 Acute bronchitis	
WC24 7,251	3.3	0.0	L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	100.0	L82 Seborrhoeic keratosis	
WC25 6,178	2.8	0.0	H00-H06:Disorders of eyelid, lacrimal system and orbit	2.8	100.0	H04 Disorders of lachrymal system	
WC26 5,993	2.8	0.0	H40-H42:Glaucoma	2.8	100.0	H40 Glaucoma	
WC27 5,986	2.7	0.0	A00-A09:Intestinal infectious diseases	2.7	100.0	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
WC28 5,890	2.7	0.0	M05-M14:Inflammatory polyarthropathies	2.7	100.0	M13 Other arthritis	
WC29 5,548	2.5	0.0	S80-S89:Injuries to the knee and lower leg	2.5	100.0	S83 Dislocation, sprain and strain of joints and ligaments of knee	
WC30 5,416	2.5	0.0	C50-C50:Malignant neoplasm of breast	2.5	100.0	C50 Malignant neoplasm of breast	
WC31 5,372	2.5	0.0	L40-L45:Papulosquamous disorders	2.5	100.0	L40 Psoriasis	
WC32 5,336	2.4	0.0	H80-H83:Diseases of inner ear	2.4	100.0	H81 Disorders of vestibular function	
WC33	2.0	0.0	S90-S99:Injuries to the ankle and	2.0	100.0	S93 Dislocation, sprain	

4,374			foot			and strain of joints and ligaments at ankle and foot level	
WC34 3,882	1.8	0.0	B15-B19:Viral hepatitis	1.8	100.0	B18 Chronic viral hepatitis	
WC35 3,758	1.7	0.0	H25-H28:Disorders of lens	1.7	100.0	H25 Senile cataract	
WC36 3,744	1.7	0.0	K50-K52:Noninfective enteritis and colitis	1.7	100.0	K52 Other non-infective gastro-enteritis and colitis	
WC37 3,689	1.7	0.0	N70-N77:Inflammatory diseases of female pelvic organs	1.7	100.0	N76 Other inflammation of vagina and vulva	
WC38 3,646	1.7	0.0	J09-J18:Influenza and pneumonia	1.7	100.0	J11 Influenza, virus not identified	
WC39 3,399	1.6	0.0	L00-L08:Infections of the skin and subcutaneous tissue	1.6	100.0	L03 Cellulitis	
WC40 3,284	1.5	0.0	Q38-Q45:Other congenital malformations of the digestive system	1.5	100.0	Q40 Other congenital malformations of upper alimentary tract	
WC41 3,180	1.5	0.0	D60-D64:Aplastic and other anaemias	1.5	100.0	D64 Other anaemias	
WC42 3,055	1.4	0.0	W50-W64:Exposure to animate mechanical forces	1.4	100.0	W57 Bitten or stung by non-venomous insect and other non-venomous arthropods	
WC43 2,926	1.3	0.0	Y40-Y59:Drugs, medicaments and biological substances causing adverse effects in therapeutic use	1.3	100.0	Y52 Agents primarily affecting the cardiovascular system	
WC44 2,884	1.3	0.0	B25-B34:Other viral diseases	1.3	100.0	B34 Viral infection of unspecified site	
WC45 2,788	1.3	0.0	E20-E35:Disorders of other endocrine glands	1.3	100.0	E28 Ovarian dysfunction	
WC46 2,765	1.3	0.0	S60-S69:Injuries to the wrist and hand	1.3	100.0	S61 Open wound of wrist and hand	

WC47 2,645	1.2	0.0	H55-H59:Other disorders of eye and adnexa	1.2	100.0	H57 Other disorders of eye and adnexa	
WC48 2,612	1.2	0.0	M60-M63:Disorders of muscles	1.2	100.0	M62 Other disorders of muscle	
WC49 2,600	1.2	0.0	H30-H36:Disorders of choroid and retina	1.2	100.0	H35 Other retinal disorders	
WC50 2,584	1.2	0.0	D70-D77:Other diseases of blood and blood-forming organs	1.2	100.0	D72 Other disorders of white blood cells	
WC51 2,508	1.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	1.2	100.0	T65 Toxic effect of other and unspecified substances	
WC52 2,309	1.1	0.0	I70-I79:Diseases of arteries, arterioles and capillaries	1.1	100.0	I73 Other peripheral vascular diseases	
WC53 2,241	1.03	0.0	I20-I25:Ischaemic heart diseases	1.0	100.0	I25 Chronic ischaemic heart disease	

* Standard error to large
**Approximately unbiased (AU) probability-value
^ Abbreviation of sex, method and number (WC1: Women Cluster 1)

Appendix 5. Prevalence and composition of diagnostic clusters in men aged 45-64 years (n= 191,171).

Cluster order number, n	Prevalence 1 .%*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value*
MC1^ 160,216	83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.69 (0.64-0.75)
			F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	F17 Mental and behavioural disorders due to use of tobacco	
			I10-I15:Hypertensive diseases	32.5	38.8	I10 Essential (primary) hypertension	
			M50-M54:Other dorsopathies	27.8	33.2	M54 Dorsalgia	
			E65-E68:Obesity and other hyperalimentation	14.6	17.4	E66 Obesity	
			E10-E14:Diabetes mellitus	14.2	16.9	E11 Non-insulin-dependent diabetes mellitus	
MC2 110,200	57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	M75 Shoulder lesions	0.87 (0.84-0.90)
			N40-N51:Diseases of male genital organs	12.1	21.0	N40 Hyperplasia of prostate	
			M20-M25:Other joint disorders	12.0	20.9	M25 Other joint disorders, not elsewhere classified	
			K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	K29 Gastritis and duodenitis	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	I84 Haemorrhoids	
			K40-K46:Hernia	8.8	15.2	K40 Inguinal hernia	

			D10-D36:Benign neoplasms	8.6	14.9	D17 Benign lipomatous neoplasm	
			M15-M19:Arthrosis	7.7	13.4	M17 Gonarthrosis [arthrosis of knee]	
			K55-K63:Other diseases of intestines	6.4	11.1	K62 Other diseases of anus and rectum	
MC3 103,365	54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	F41 Nutritional marasmus	0.79 (0.74-0.84)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	K02 Dental caries	
			J40-J47:Chronic lower respiratory diseases	9.3	17.2	J44 Other chronic obstructive pulmonary disease	
			J00-J06:Acute upper respiratory infections	8.9	16.4	J00 Acute nasopharyngitis [common cold]	
			J30-J39:Other diseases of upper respiratory tract	8.0	14.8	J30 Vasomotor and allergic rhinitis	
			L20-L30:Dermatitis and eczema	7.5	13.9	L21 Seborrhoeic dermatitis	
			G40-G47:Episodic and paroxysmal disorders	7.4	13.7	G47 Sleep disorders	
			F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	F52 Sexual dysfunction, not caused by organic disorder or disease	
			F30-F39:Mood [affective] disorders	6.3	11.6	F32 Depressive episode	
			B35-B49:Mycoses	4.1	7.6	B35 Dermatophytosis	
MC4 48,140	25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	H91 Disorders of sclera and cornea in diseases classified elsewhere	0.87 (0.83-0.91)
			H53-H54:Visual disturbances and blindness	3.9	15.5	H54 Blindness and low vision	
			B00-B09:Viral infections characterized by skin and	3.5	13.9	B07 Viral warts	

			mucous membrane lesions				
			L60-L75:Disorders of skin appendages	3.5	13.9	L72 Follicular cysts of skin and subcutaneous tissue	
			H10-H13:Disorders of conjunctiva	3.0	12.0	H10 Conjunctivitis	
			H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	H52 Disorders of refraction and accommodation	
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	L82 Seborrhoeic keratosis	
			L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	L02 Cutaneous abscess, furuncle and carbuncle	
			H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	H00 Hordeolum and chalazion	
MC5 30,315	15.9	2.9	I30-I52:Other forms of heart disease	6.9	43.7	I45 Other conduction disorders	0.77 (0.73-0.82)
			I20-I25:Ischaemic heart diseases	5.0	31.0	I25 Chronic ischaemic heart disease	
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	15.2	I73 Other peripheral vascular diseases	
			I60-I69:Cerebrovascular diseases	1.8	11.4	I64 Stroke, not specified as haemorrhage or infarction	
			H30-H36:Disorders of choroid and retina	1.7	11.0	H36 Retinal disorders in diseases classified elsewhere	
			N17-N19:Renal failure	1.5	9.2	N18 Chronic renal failure	
MC6 29,907	15.6	1.5	M05-M14:Inflammatory polyarthropathies	5.4	34.5	M10 Gout	0.66 (0.60-0.72)
			S80-S89:Injuries to the knee and lower leg	3.2	20.7	S83 Dislocation, sprain and strain of joints and ligaments of knee	

			L40-L45:Papulosquamous disorders	3.2	20.5	L40 Psoriasis	
			G50-G59:Nerve, nerve root and plexus disorders	2.7	17.4	G56 Mononeuropathies of upper limb	
			M65-M68:Disorders of synovium and tendon	2.6	16.8	M65 Synovitis and tenosynovitis	
MC7 19,313	10.1	0.9	K70-K77:Diseases of liver	5.2	51.9	K76 Other diseases of liver	1.00
			B15-B19:Viral hepatitis	3.2	32.0	B18 Chronic viral hepatitis	
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	2.6	25.8	K80 Cholelithiasis	
MC8 19,160	10.0	0.7	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	41.0	T14 Injury of unspecified body region	0.80 (0.75-0.86)
			L50-L54:Urticaria and erythema	2.1	20.8	L50 Urticaria	
			S60-S69:Injuries to the wrist and hand	1.9	18.5	S61 Open wound of wrist and hand	
			S90-S99:Injuries to the ankle and foot	1.4	13.9	S93 Dislocation, sprain and strain of joints and ligaments at ankle and foot level	
			S00-S09:Injuries to the head	1.3	13.3	S01 Open wound of head	
MC9 13,752	7.2	0.3	E00-E07:Disorders of thyroid gland	2.4	32.8	E03 Other hypothyroidism	0.97 (0.96-0.98)
			K50-K52:Noninfective enteritis and colitis	1.7	24.1	K52 Other non-infective gastro-enteritis and colitis	
			D50-D53:Nutritional anaemias	1.2	16.3	D50 Iron deficiency anaemia	
			Q38-Q45:Other congenital malformations of the digestive system	1.2	16.2	Q40 Other congenital malformations of upper alimentary tract	
			D70-D77:Other diseases of blood and blood-forming organs	1.1	15.3	D72 Other disorders of white blood cells	
MC10	7.1	0.5	J20-J22:Other acute lower	2.6	37.2	J20 Acute bronchitis	0.86 (0.82-0.91)

13,490			respiratory infections				
			A00-A09:Intestinal infectious diseases	2.4	34.2	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
			J09-J18:Influenza and pneumonia	1.6	22.0	J11 Influenza, virus not identified	
			B25-B34:Other viral diseases	1.0	14.6	B34 Viral infection of unspecified site	
MC11 13,434	7.0	0.4	H60-H62:Diseases of external ear	3.9	55.3	H61 Other disorders of external ear	0.88 (0.85-0.92)
			H65-H75:Diseases of middle ear and mastoid	2.1	30.5	H65 Nonsuppurative otitis media	
			H80-H83:Diseases of inner ear	1.4	19.9	H81 Disorders of vestibular function	
MC12 10,952	5.7	0.1	N20-N23:Urolithiasis	4.3	75.5	N20 Calculus of kidney and ureter	0.99 (0.98-1.00)
			N30-N39:Other diseases of urinary system	1.6	27.1	N39 Other disorders of urinary system	
MC13 8,794	4.6	0.3	M45-M49:Spondylopathies	3.1	66.7	M47 Spondylosis	0.99 (0.95-1.00)
			M40-M43:Deforming dorsopathies	1.8	38.9	M41 Scoliosis	
MC14 7,444	3.9	0.2	H40-H42:Glaucoma	2.3	59.6	H40 Glaucoma	0.91 (0.88-0.95)
			H25-H28:Disorders of lens	1.8	45.4	H25 Senile cataract	
MC15 6,161	3.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	2.1	65.8	T65 Toxic effect of other and unspecified substances	0.90 (0.87-0.93)
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	35.3	F20 Schizophrenia	

**Approximately unbiased (AU) probability-value

Appendix 6. Factors in women aged 45-64 years (n= 217,823).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WF1^ 130,072	59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
			M70-M79:Other soft tissue disorders	27.0	45.2		
			M15-M19:Arthrosis	15.7	26.2		
			G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
			M45-M49:Spondylopathies	4.3	7.3		
			M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 82,301	37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
			E65-E68:Obesity and other hyperalimentation	19.0	50.2		
			E10-E14:Diabetes mellitus	7.7	20.3		
WF3 71,436	32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
			E10-E14:Diabetes mellitus	7.7	23.4		
			I30-I52:Other forms of heart disease	4.4	13.3		
			H25-H28:Disorders of lens	1.7	5.3		
			H30-H36:Disorders of choroid and retina	1.2	3.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
			I20-I25:Ischaemic heart diseases	1.0	3.1		
WF4 60,027	27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
			N30-N39:Other diseases of urinary system	5.9	21.3		
			H60-H62:Diseases of external ear	3.6	13.1		
			J20-J22:Other acute lower respiratory infections	3.4	12.2		
			A00-A09:Intestinal infectious diseases	2.7	10.0		
			H65-H75:Diseases of middle ear and mastoid	2.5	9.2		

			J09-J18:Influenza and pneumonia	1.7	6.1		
			B25-B34:Other viral diseases	1.3	4.8		
			M60-M63:Disorders of muscles	1.2	4.4		
WF5 56,671	26.0	5.3	L20-L30:Dermatitis and eczema	9.3	35.7	8.4	77.5
			B35-B49:Mycoses	5.7	22.1		
			L60-L75:Disorders of skin appendages	4.8	18.3		
			H53-H54:Visual disturbances and blindness	4.4	16.9		
			H10-H13:Disorders of conjunctiva	3.8	14.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	12.8		
			H55-H59:Other disorders of eye and adnexa	1.2	4.7		
WF6 46,391	21.3	3.8	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	53.6	5.9	90.4
			K40-K46:Hernia	5.7	26.7		
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	18.8		
			K70-K77:Diseases of liver	2.9	13.8		
			Q38-Q45:Other congenital malformations of the digestive system	1.5	7.1		
WF7 41,492	19.0	0.6	M80-M85:Disorders of bone density and structure	11.3	59.5	5.1	95.5
			D50-D53:Nutritional anaemias	8.3	43.5		
WF8 40,619	18.6	0.0	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	100.0	4.3	99.8
WF9 23,729	10.9	0.6	J40-J47:Chronic lower respiratory diseases	8.2	74.9	4.1	103.9
			J20-J22:Other acute lower respiratory infections	3.4	30.8		

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor

^ Abbreviation of sex, method and number (WF1: Women Factor 1)

Appendix 7. Factors in men aged 45-64 years (n= 191,171).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in age, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
MF1^ 118,037	61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
			I10-I15:Hypertensive diseases	32.6	52.7		
			E65-E68:Obesity and other hyperalimentation	14.6	23.6		
			M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 75,315	39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
			I30-I52:Other forms of heart disease	6.9	17.6		
			I20-I25:Ischaemic heart diseases	5.0	12.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
			I60-I69:Cerebrovascular diseases	1.8	4.6		
			N17-N19:Renal failure	1.5	3.7		
MF3 73,638	38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
			K70-K77:Diseases of liver	5.2	13.6		
			B15-B19:Viral hepatitis	3.2	8.4		
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 66,303	34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
			M15-M19:Arthrosis	7.7	22.2		
			M45-M49:Spondylopathies	3.1	8.8		
			M40-M43:Deforming dorsopathies	1.8	5.2		
MF5 34,903	18.3	2.5	L20-L30:Dermatitis and eczema	7.5	41.3	22.2	50.7
			B35-B49:Mycoses	4.1	22.5		
			H53-H54:Visual disturbances and	3.9	21.3		

			blindness				
			H10-H13:Disorders of conjunctiva	3.0	16.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	13.8		
MF6 27,697	14.5	1.8	J00-J06:Acute upper respiratory infections	8.9	61.3	10.4	61.0
			J20-J22:Other acute lower respiratory infections	2.6	18.1		
			A00-A09:Intestinal infectious diseases	2.4	16.7		
			J09-J18:Influenza and pneumonia	1.6	10.7		
			B25-B34:Other viral diseases	1.0	7.1		
MF7 33,568	17.6	2.2	E10-E14:Diabetes mellitus	14.2	80.8	9.5	70.5
			H40-H42:Glaucoma	2.3	13.2		
			H25-H28:Disorders of lens	1.8	10.1		
			H30-H36:Disorders of choroid and retina	1.7	10.0		
MF8 25,121	13.1	0.7	K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	87.8	6.5	84.3
			D50-D53:Nutritional anaemias	1.2	8.9		
			Q38-Q45:Other congenital malformations of the digestive system	1.2	8.9		
MF9 15,974	8.4	0.4	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	49.2	5.0	99.8
			S80-S89:Injuries to the knee and lower leg	3.3	38.7		
			S90-S99:Injuries to the ankle and foot	1.4	16.7		
MF10 8,271	4.3	0.0	N20-N23:Urolithiasis	4.3	100.0	3.9	103.7

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor

^ Abbreviation of sex, method and number (MF1: Men Factor 1)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*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	6-8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	
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	6-9
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	6-9
		(*e*) Describe any sensitivity analyses	6-9
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	10
		(b) Give reasons for non-participation at each stage	10
		(c) Consider use of a flow diagram	Supplementary File Appendix 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	Tables
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	Tables 2 and 3.
		(*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	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	23
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	20-21
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	19
Generalisability	21	Discuss the generalisability (external validity) of the study results	20
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	23

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

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.

BMJ Open

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018986.R1
Article Type:	Research
Date Submitted by the Author:	30-Nov-2017
Complete List of Authors:	Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), ; Universitat Autònoma de Barcelona, Violan-Fors, Concepción; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Foguet-Boreu, Quintí; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol) Universitat Autònoma de Barcelona; Universitat Autonoma de Barcelona Rodriguez-Blanco, Teresa; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pons-Vigués, Mariona; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pujol-Ribera, Enriqueta; Institut Universitari d'Investigació en Atenció Primaria (IDIAP-Jordi Gol); Universitat Autonoma de Barcelona Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Research methods, General practice / Family practice, Anaesthesia
Keywords:	Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

1. Albert Roso-Llorach. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

2. Concepción Violán. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

3. Quintí Foguet-Boreu. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 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, Catalonia.
E-mail: 42292qfb@comb.cat

4. Teresa Rodriguez-Blanco. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol). Universitat Autònoma de Barcelona, Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. E-mail: trodriguez@idiapjgol.org

5. Mariona Pons-Vigués. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain.
E-mail: mponsv@idiapjgol.org

6. Enriqueta Pujol-Ribera. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain. E-mail: epujol@idiapjgol.org

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

Corresponding author: Albert Roso-Llorach. IDIAP Jordi Gol
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: aroso@idiapjgol.org

Word count: 4054

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

ABSTRACT

Objective The aim was to compare multimorbidity patterns identified with the two most commonly used methods: hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA) in a large primary care database. Specific objectives were 1) to determine how the choice of method may affect the composition of these patterns and 2) to consider the potential application of each method in the clinical setting.

Design Cross-sectional study. Diagnoses were based on the 263 corresponding blocks of the International Classification of Diseases (ICD-10) . Multimorbidity patterns were identified using HCA and EFA. Analysis was stratified by sex, and results compared for each method.

Setting and participants Electronic health records for 408,994 patients with multimorbidity aged 45-64 years in 274 primary health care teams from 2010 in Catalonia, Spain.

Results HCA identified 53 clusters for women, with just 12 clusters including at least 2 diagnoses, and 15 clusters for men, all of them including at least 2 diagnoses. EFA showed 9 factors for women and 10 factors for men. We observed differences by sex and method of analysis, although some patterns were consistent. Three combinations of diseases were observed consistently across sex groups and across both methods: hypertension and obesity, spondylopathies and deforming dorsopathies, and dermatitis eczema and mycosis.

Conclusions This study showed that multimorbidity patterns vary depending on the method of analysis used (HCA vs EFA) and provided new evidence about the known limitations of attempts to compare multimorbidity patterns between RWD studies. We found that EFA was useful in describing comorbidity relationships and HCA could be useful for in-depth study of multimorbidity. Our results suggest possible applications for each of these methods and add information about some aspects that must be considered in standardization of future studies: spectrum of diseases, data utilization, and methods of analysis.

Keywords: Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Strengths and limitations of this study

- This one of the first studies to compare the two methodologies most commonly used to obtain patterns of multimorbidity, Hierarchical Cluster Analysis and Exploratory Factor Analysis.
- The dual analysis was performed in a large, high-quality database of primary care electronic health records that have been shown to be representative of a much larger population, stratified by sex.
- Internal validation with bootstrap methods provided more robust evidence for the cluster analysis.
- The agglomerative hierarchical clustering forces every unit into a single cluster, is exploratory in nature and different clustering algorithms may produce different results.
- The study is cross-sectional and further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

INTRODUCTION

The reliable identification of patterns of multimorbidity is a critical step in developing health care services sensitive to the health needs of these patients.[1] Recent reviews of multimorbidity patterns have shown that individual studies differ widely in their design and choice of epidemiological and statistical methods, including sampling frameworks and selection criteria, coding systems, eligible diseases, and definition of disease clustering patterns.[2–4] These studies highlight the lack of consensus to measure patterns of comorbidity and multimorbidity. In recent years, the number of studies based on real-world data (RWD) [5] has increased significantly, which makes it even more difficult to establish a consensus on how to measure comorbidity and multimorbidity patterns. Although much more information is available, the different databases may not be comparable, making it difficult to arrive at observations and draw firm conclusions. It also limits our ability to compare analyses using RWD and to evaluate whether one approach may be better suited to the purpose. Therefore, it is difficult to identify multimorbidity patterns and provide adequate health services according to the population needs.

The most frequent methods used to date have been hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA), which offer very different approaches and solutions.[2,3] Both are descriptive methods to identify association of diagnoses and determine patterns of multimorbidity. HCA obtains the patterns of multimorbidity from the dissimilarities between diseases, while EFA is based on correlations between diagnoses to identify the patterns. The HCA clusters tend to contain diagnoses that are similar to each other (in terms of Euclidean distances), but dissimilar from the diagnoses in other clusters; no diagnosis can be included in more than one cluster. In contrast, EFA along with confirmatory factor analysis are primarily used to test hypothesized relationships between observed measures and latent constructs. In

addition, EFA allows for inclusion of any diagnosis in multiple factors as they can present significant correlations with more than one factor. [6-9]

For all these reasons, the aim was to compare multimorbidity patterns identified with the two most commonly used methods: hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA) in a large primary care database. Specific objectives were 1) to determine how the choice of method may affect the composition of these patterns and 2) to consider the potential application of each method in the clinical setting.

METHODS

Design, setting and study population

A cross-sectional study was conducted in Catalonia (Spain), a Mediterranean region with 7,434,632 inhabitants, 81% of which live in urban municipalities (2010 census). The Spanish National Health Service (NHS) provides universal coverage, financed mainly by tax revenue. The Catalan Health Institute (CHI) manages primary health care teams (PHCTs) that serve 5,501,784 patients (274 PHCT), or 74% of the population; the remaining PHCTs are managed by other providers. The CHI's Information System for the Development of Research in Primary Care (SIDIAP) contains the coded clinical information recorded in electronic health records EHR by its 274 PHCTs since 2006. A subset of records meeting the highest quality criteria for clinical data (SIDIAP Q) includes 40% of the SIDIAP population (1,833,125 individuals), attended by the 1,365 general practitioners (GPs) whose data recording scores contain information on the majority of the population of Catalonia, and is highly representative for the whole region in terms of geography, age, gender and diseases.[10–12]

Prevalence of individual conditions varies with age and so does multimorbidity and their patterns. In order to obtain a more homogenous sample in terms of multimorbidity, we

focussed on individuals aged 45 to 64 years.[13–16] We identified 408,944 individuals aged 45 to 64 years on 31 December 2010 with two or more diagnoses (Appendix 1).

Coding and selection of diseases

Diseases are coded in SIDIAP using International Classification of Diseases version 10 (ICD-10). For this study, we selected all active diagnoses recorded in EHR as of December 31, 2010, except for R codes (symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified) and Z codes (factors influencing health status and contact with health services).[17] Non-active diagnoses were excluded, based on the presence of an end date in the EHR. These diagnoses cover a broad list of acute diseases for which the system automatically assigns an end date (e.g., 60 days after the initial diagnosis).

To facilitate management of the diagnostic information, the diagnoses were extracted using the 263 blocks (disease categories) in the ICD-10 structure. These are homogeneous categories of very closely related specific diagnoses. For example, Hypertensive diseases include Essential (primary) hypertension, Hypertensive heart disease, Hypertensive renal disease, Hypertensive heart and renal disease and Secondary hypertension. To obtain consistent and clinically interpretable patterns of association, and to avoid spurious relationships that could bias the results, we considered only diagnoses with greater than 1% prevalence in each sex. All patients with multimorbidity (2 or more coexisting diagnoses recorded in the EHR on 31 December 2010) were included.

Variables

The variables considered were: diagnosis (values: 1 for present, 0 for absent), number of diseases (2, 3, 4, and 5 or more) and sex (women, men) were also recorded for each patient.

Statistical analysis

Data access: Data was obtained from SIDIAP after the study was authorized. All the project's authors could access the database. Cleaning methods: The analysis was limited to SIDIAP-Q, as

the sample was representative of the population.[10–12] No missing values were handled as sex and age were recorded for all population. Wrong sex-specific diagnoses codes and diagnoses with inconsistent dates were excluded. An individual with no disease diagnoses record was considered as disease free.

Analyses were stratified by sex. Descriptive statistics were used to summarize overall information. Categorical variables were expressed as frequencies (percentage) and continuous as mean (standard deviation, SD) or median (interquartile range, IQR). Two sample test of proportions and Mann-Whitney test were used to test differences by sex.

We identified disease patterns using two approaches: 1) hierarchical cluster analysis (HCA), and 2) exploratory factor analysis (EFA). Clinical criteria were used to evaluate the consistency and utility of the final HCA and EFA solutions, based on previously described patterns in the literature and a consensus opinion drawn from the clinical experience of the research team (4 family physicians, 1 epidemiologist). Clusters and factors in these analyses were considered as two sets of grouping solutions, which were then assigned to each individual patient. We considered patients to be associated with a given grouping solution if they had ≥1 diagnoses in that solution, allowing for the calculation of the prevalence of each solution in the sample. Patients could be associated with more than one solution in the same set. We also calculated prevalence, restricting the assignment of patients to those with ≥ 2 diagnoses in the same solution.

Hierarchical Cluster Analysis

The HCA approach assigns diagnoses to groups or clusters, so that diagnoses in the same cluster are more similar, based on a given measure, to one another than to diagnoses from different clusters. The Jaccard coefficient was used to measure similarity. This coefficient considers only the diagnoses that any two patients have and ignores the diagnoses that neither of them has.[6] As we do not know a priori the number of clusters to retain from the

data, we used agglomerative hierarchical methods to identify possible clustering solutions: Average linkage, Ward, flexible beta and other methods with less bias, based on nonparametric estimates, such as Single Linkage and Density Linkage. All but Ward and the flexible beta methods successively chained the observations into one cluster. Therefore, the Ward method, which minimizes the variance within clusters and produces clusters of similar sizes, was chosen as the primary method based on dendrograms analysis.[6] Data were randomly split into test and training datasets, equal in size and analysed separately. We ran the Ward method on both samples. The semi-partial R2, Calinski-Harabasz pseudo-F- and pseudo-T2-statistic criteria for different numbers of clusters were examined.[6] Clustering solutions were compared between the test and training datasets, taking into account the number of clusters, Adjusted Rand Index and clinical criteria. After checking algorithm stability, Ward method was run on the full data set, applying the same criteria to different numbers of clusters. Results were compared with flexible beta results, with beta values set at -0.25 and -0.5. The criteria for selecting the number of clusters were the highest adjusted Rand index with a high number of clusters and a high pseudo T2 statistic.[6] To assess internal cluster quality, we applied multiscale bootstrap resampling to obtain an approximately unbiased (AU) probability. This probability ('p-value') is the proportion of bootstrapped samples that contain the cluster; larger p-values indicate more support for the cluster.[18]

Multidimensional scaling (MDS) considering two dimensions was used to discover the underlying structure of distance measures between diseases in the cluster analysis. Essentially, MDS assigns observations to specific locations in a conceptual space such that the distances between points in the space match the given dissimilarities as closely as possible. We carried out classical MDS using the distance matrix obtained in the cluster analysis that considered the Jaccard coefficient as a dissimilarity measure. The conceptual map of the diseases distinguishes between the intra-disease cluster and the inter-disease cluster. Taking into account the final cluster's solution and the obtained groups, conceptual maps of the diseases

were created. For a better interpretation of the conceptual map, prevalence of the disease was represented as the radius of the circle.[19]

Exploratory Factor Analysis

EFA reduces the observed set of diagnoses to a smaller number of latent factors that account for the correlations between them. As the study variables were dichotomous, the correlation matrix between the diagnoses was computed using tetrachoric correlations. The factorability of the matrix was tested using Bartlett Test of Sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. The extraction of the initial solution was carried out using the principal factors method with squared multiple correlations for the prior communality estimates. The optimal number of extracted factors for the final solution was determined with the Scree plot using the "elbow" rule and setting the percentage of variance equal to 100 percent. Factor loadings were analyzed to identify factor patterns. An oblique rotation, Oblimin, was performed to clarify the factor pattern in order to better interpret the nature of the factors, as we assumed that factors were allowed to be associated with each other. As a rule of thumb, factor loadings greater than or equal to 0.30 in absolute value were considered to be significant.[7]

Comparing multimorbidity patterns

We compared every cluster and factor solutions across sex groups agreement and the diagnoses included in it.

We considered grouping solutions (HCA vs EFA) from different sets (women vs men) to have the following degrees of similarity: a) perfect, when the solution included exactly the same diseases as another solution in the other comparison group (sex or statistical approach); b) partial, when the solution included a subset of diseases present in a solution in the other comparison group; and c) none, when each and every disease in the solution was part of a different solution in the other group and none was part of the same solution. These groups

were named using the abbreviation of sex, method and number (For example, MC1: Men Cluster 1)

We further extracted the common subsets of diseases within partially similar solutions, which together with completely similar solutions gave a comprehensive picture of overlapping cluster and factor analysis solutions.

The analyses were performed using SPSS for Windows, version 18 (SPSS Inc., Chicago, IL, USA), SAS 9.2 for Windows (SAS Institute Inc., Cary, NC, USA) and R version 3.1.2 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Of 523,656 patients, 408,994 (78.1%) met the multimorbidity criteria; women had a higher multimorbidity prevalence than men (82.2% vs 73.9%, respectively; p<0.001) (Table 1). Both cluster and factor analyses included 79 diagnoses for women and 73 for men.

Table 1. Number of diseases, clusters, and factors identified in cluster and factorial analysis for patients 45-64 years-old, stratified by sex (N=523,656)

	Total, n (%)	Women, n (%)	Men, n (%)
≥2 Diagnoses*	408,994 (78.1%)	217,823 (82.2%)	191,171 (73.9%)
Number of diagnoses*			
2		26,106 (12.0%)	33,850 (17.7%)
3		28,243 (13.0%)	33,515 (17.5%)
4		28,274 (13.0%)	30,356 (15.9%)
≥5		135,200 (62.1%)	93,450 (48.9%)
Median number of diagnoses (IQR)**		5 (4-8)	4 (3-7)
Number of diagnoses included		79	73
Number of clusters		53	15
Number of clusters with ≥2 diagnoses		12	15
Median of diagnoses per clusters (IQR)***		2 (2-4)	5 (2.5-6)
Number of factors		9	10
Number of factors with ≥2 diagnoses		8	9
Median of diagnoses per factors (IQR)***		5.5 (2.75-7)	4 (4-5)

Abbreviations: IQR, Inter-quartile range.
** Two sample test of proportions; all p-values<0.001*
*** Mann-Whitney test; P<0.001*
****Median of clusters or factors with ≥2 diseases; P<0.001*

Hierarchical Cluster analysis

Using HCA with the Ward method, we obtained 53 clusters, with just 12 clusters grouping at least two diagnoses for women and 15 clusters for men (Table 1). We describe only the four most prevalent clusters (Table 2). For a complete description of the clusters and dendrograms, see Appendices 2-5.

Table 2. Four most prevalent clusters, by sex group (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2,%†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	AU p-value**
WOMEN					
WC1^ 82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	0.79 (0.74-0.85)
		F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	
		M70-M79:Other soft tissue disorders	27.0	32.8	
		N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	
		M20-M25:Other joint disorders	18.6	22.6	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	
		D10-D36:Benign neoplasms	16.2	19.7	
		M15-M19:Arthrosis	15.7	19.1	
WC2 55.8	23.0	E70-E90:Metabolic disorders	37.4	63.4	0.93 (0.86-1.00)
		I10-I15:Hypertensive diseases	25.6	45.8	
		E65-E68:Obesity and other hyperalimentation	19.0	34.0	
		E10-E14:Diabetes mellitus	7.7	13.7	
WC3 47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.7	39.4	0.78 (0.73-0.84)
		E00-E07:Disorders of thyroid gland	14.9	31.4	
		F30-F39:Mood [affective] disorders	14.6	30.8	
		M80-M85:Disorders of bone density and structure	11.3	23.9	
WC4 32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	0.71 (0.66-0.77)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	
		L20-L30:Dermatitis and eczema	9.3	28.8	
		B35-B49:Mycoses	5.7	17.8	
MEN					
MC1^^ 83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	0.69 (0.64-0.75)
		F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	
		I10-I15:Hypertensive diseases	32.5	38.8	

		M50-M54:Other dorsopathies	27.8	33.2	
		E65-E68:Obesity and other hyperalimentation	14.6	17.4	
		E10-E14:Diabetes mellitus	14.2	16.9	
MC2 57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	0.87 (0.84-0.90)
		N40-N51:Diseases of male genital organs	12.1	21.0	
		M20-M25:Other joint disorders	12.1	20.9	
		K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	
		K40-K46:Hernia	8.8	15.2	
		D10-D36:Benign neoplasms	8.6	14.9	
		M15-M19:Arthrosis	7.7	13.4	
		K55-K63:Other diseases of intestines	6.4	11.1	
MC3 54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	0.79 (0.74-0.84)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	
		J40-J47:Chronic lower respiratory diseases	9.3	17.2	
		J00-J06:Acute upper respiratory infections	8.9	16.4	
		J30-J39:Other diseases of upper respiratory tract	8.0	14.8	
		L20-L30:Dermatitis and eczema	7.5	13.9	
		G40-G47:Episodic and paroxysmal disorders	7.4	13.7	
		F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	
		F30-F39:Mood [affective] disorders	6.3	11.6	
		B35-B49:Mycoses	4.1	7.6	
MC4 25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	0.87 (0.83-0.91)
		H53-H54:Visual disturbances and blindness	3.9	15.5	
		B00-B09:Viral infections characterized by skin and mucous membrane lesions	3.5	13.9	
		L60-L75:Disorders of skin appendages	3.5	13.9	
		H10-H13:Disorders of conjunctiva	3.0	12.0	
		H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	
		L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	
		L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	
		H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	

*Individuals from the strata ≥1 diagnosis in the cluster/ † Individuals from the strata with ≥2 diagnosis in the cluster/‡Strata: same sex

**Approximately unbiased (AU) probability-value

^ Abbreviation of Sex, method and number (WC1: Women Cluster 1)

^^ Abbreviation of Sex, method and number (MC1: Men Cluster 1)

Twelve clusters with at least two diseases were identified for women, with prevalences ranging from 6.6% to 82.1% (median: 15.5%), (WC3; WC10). The clusters identified in men had prevalences ranging from 3.2% to 83.8% (median 10.1%). The most prevalent cluster included 8 diseases in women (WC1: musculoskeletal, psychiatric, circulatory, gynecological and neoplasms) and 6 in men (MC1: metabolic and circulatory); about half of all patients had at least two diagnoses (52.9% of women and 50.4% of men).

Two clusters were common to men and women, “Spondylopathies” and “Deforming dorsopathies” (WC11, MC13) and “Urolithiasis” and “Other diseases of the urinary systems” (WC9, MC12) (Box 1). The remaining clusters showed partial similarity in men and women, based on six subsets (Box 2), except for three clusters found only in women (WC3, WC7, WC10) and six only in men (MC4, MC5, MC8, MC10, MC14, MC15) (Appendices 4-5).

BOXES

Box 1. Combinations of diseases consistent in both men and women$

Clusters

Complete (whole) clusters

1. M45-M49:Spondylopathies*
   M40-M43:Deforming dorsopathies (WC11;MC13)#
2. N20-N23:Urolithiasis
   N30-N39:Other diseases of urinary system (WC9; MC12)

Subsets within clusters

1. E65-E68:Obesity and other hyperalimentation
   10-I15:Hypertensive diseases
   E10-E14:Diabetes mellitus (WC2; MC1)
2. M15-M19:Arthrosis
   M20-M25:Other joint disorders
   I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
   M70-M79:Other soft tissue disorders
   D10-D36:Benign neoplasms (WC1; MC2)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   K00-K14:Diseases of oral cavity, salivary glands and jaws
   J00-J06:Acute upper respiratory infections (WC4; MC3)
4. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12; MC7)
5. J30-J39:Other diseases of upper respiratory tract
   J40-J47:Chronic lower respiratory diseases (WC6; MC3)
6. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; MC2)
7. G50-G59:Nerve, nerve root and plexus disorders
   M65-M68:Disorders of synovium and tendon (WC8; MC6)

Factors*

Subgroups within factors

1. I10-I15:Hypertensive diseases
   I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I70-I79:Diseases of arteries, arterioles and capillaries (WF3; MF2)
2. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation (WF2;MF1)
3. J00-J06:Acute upper respiratory infections
   J20-J22:Other acute lower respiratory infections
   J09-J18:Influenza and pneumonia
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (WF4; MF6)
4. M15-M19:Arthrosis
   M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies
   M50-M54:Other dorsopathies (WF1;MF4)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   Q38-Q45:Other congenital malformations of the digestive system (WF6;MF8)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   H53-H54:Visual disturbances and blindness
   H10-H13:Disorders of conjunctiva
   L80-L99:Other disorders of the skin and subcutaneous tissue (WF5;MF5)
7. H25-H28:Disorders of lens
   H30-H36:Disorders of choroid and retina (WF3; MF7)

** Coincident disease in both sexes*

 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)

$ No two full factors were exactly the same for both sexes.

Underlined blocks of diagnosis represent coincident diseases in pattern.

Box 2. Combinations of diseases consistent across statistical methods (cluster and factor analysis)$

Women

1. I10-I15:Hypertensive diseases*
   E65-E68:Obesity and other hyperalimentation
   E10-E14:Diabetes mellitus (WC2; WF2)#
2. M15-M19:Arthrosis
   M50-M54:Other dorsopathies
   M70-M79:Other soft tissue disorders (WC1; WF1)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (WC4; WF5)
4. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (WC11; WF1)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; WF6)
6. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12 ;WF6)

Men

1. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation
   E70-E90:Metabolic disorders (MC1; MF1)
2. I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I60-I69:Cerebrovascular diseases
   I70-I79:Diseases of arteries, arterioles and capillaries
   N17-N19:Renal failure (MC5; MF2)
3. J09-J18:Influenza and pneumonia
   J20-J22:Other acute lower respiratory infections
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (MC10; MF6)
4. H10-H13:Disorders of conjunctiva
   H53-H54:Visual disturbances and blindness
   L80-L99:Other disorders of the skin and subcutaneous tissue (MC4; MF5)
5. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (MC13; MF4)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (MC3; MF5)
7. K70-K77:Diseases of liver
   B15-B19:Viral hepatitis (MC7; MF3)
8. T08-T14:Injuries to unspecified part of trunk, limb or body region
   S90-S99:Injuries to the ankle and foot (MC8; MF9)
9. H25-H28:Disorders of lens
   H40-H42:Glaucoma (MC14; MF7)

** Coincident disease in both methods*
 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)$ All subgroups of factors or clusters, no single cluster exactly the same as a factor.
Underlined blocks of diagnosis represent coincident diseases in pattern.

The top four clusters were reproduced in the graphical representation of coordinates using MDS. The most prevalent disease clusters were more clearly separated in women than they were in men, mostly due to the overlap of MC2 and MC3 (Figure 1).

Exploratory Factor Analysis

Using EFA, we obtained nine factors for women and 10 factors for men. In this analysis, the median number of diagnoses per factor was higher in women (Table 1).Two factors explained more than 50% of total variance (58.5% for women (WF2; WF3) and 50.7% for men (MF1; MF2)). All diseases were assigned to single factors except for three diseases, two in women only (J20-J22: Other acute lower respiratory infections and E10-E14: Diabetes mellitus) and one in both women and men (I10-I15: Hypertensive diseases); all three were assigned to two factors. The first four factors are described in Table 3; full factor solutions are shown in Appendices 6-7.

Table 3. Four most prevalent factors, by sex (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses**	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WOMEN						
WF1^ 59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
		M70-M79:Other soft tissue disorders	27.0	45.2		
		M15-M19:Arthrosis	15.7	26.2		
		G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
		M45-M49:Spondylopathies	4.3	7.3		
		M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
		E65-E68:Obesity and other hyperalimentation	19.0	50.2		
		E10-E14:Diabetes mellitus	7.7	20.3		
WF3 32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
		E10-E14:Diabetes mellitus	7.7	23.4		
		I30-I52:Other forms of heart disease	4.4	13.3		
		H25-H28:Disorders of lens	1.7	5.3		
		H30-H36:Disorders of choroid and retina	1.2	3.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
		I20-I25:Ischaemic heart diseases	1.0	3.1		
WF4 27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
		N30-N39:Other diseases of urinary system	5.9	21.3		
		H60-H62:Diseases of external ear	3.6	13.1		
		J20-J22:Other acute lower respiratory infections	3.4	12.2		
		A00-A09:Intestinal infectious diseases	2.7	10.0		

		H65-H75:Diseases of middle ear and mastoid	2.5	9.2		
		J09-J18:Influenza and pneumonia	1.7	6.1		
		B25-B34:Other viral diseases	1.3	4.8		
		M60-M63:Disorders of muscles	1.2	4.4		
MEN						
MF1^^ 61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
		I10-I15:Hypertensive diseases	32.6	52.7		
		E65-E68:Obesity and other hyperalimentation	14.6	23.6		
		M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
		I30-I52:Other forms of heart disease	6.9	17.6		
		I20-I25:Ischaemic heart diseases	5.0	12.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
		I60-I69:Cerebrovascular diseases	1.8	4.6		
		N17-N19:Renal failure	1.5	3.7		
MF3 38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
		K70-K77:Diseases of liver	5.2	13.6		
		B15-B19:Viral hepatitis	3.2	8.4		
		F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
		M15-M19:Arthrosis	7.7	22.2		
		M45-M49:Spondylopathies	3.1	8.8		
		M40-M43:Deforming dorsopathies	1.8	5.2		

**Individuals from the strata ≥1 diagnosis in the factor/ † Individuals from the strata with ≥2 diagnosis in the factor/‡Strata: same sex*
***KMO sampling adequacy index was 0.82 for women and 0.75 for men. Bartlett Test of Sphericity was statistically significant. (p<0.001) for both groups*
^ Abbreviation of Sex, method and number (WF1: Women Factor 1)
^^ Abbreviation of Sex, method and number (MC1: Men Factor 1)

Although no factor-based groupings were identical in men and women, almost all showed partial similarity by sex, based on seven subsets (Box 1), except for two groups found only in women (WF7, WF9) and one found only in men (MF9).

Multimorbidity patterns comparison across statistical approaches

The EFA multimorbidity patterns were more easily interpreted than the HCA groups, either because they made more sense from a clinical perspective or because of greater homogeneity

in the diagnoses: about half of the factors containing at least 3 diagnoses corresponded to a maximum of 2 ICD-10 blocks, compared to about one fifth of the HCA clusters with at least 3 diseases (1/5 for women and 2/11 for men). No grouping solution for women or for men contained exactly the same set of diseases as a cluster and as a factor. Six clusters (WC3, WC6, WC7, WC8, WC9, WC10) and two factors (WF7, WF9) were only observed in women. However, six subsets of diseases were part of the same grouping in both a cluster and a factor (Box 2); all included two or three diagnoses, usually from the same ICD chapter. Five clusters and one factor were observed only in men (MC6, MC9, MC11, MC12, MC15 and MF6). Nine subsets of diseases were observed as part of the same grouping in both a cluster and a factor (Box 2). They included a range of diseases (2-5) and most frequently included diseases from different ICD chapters.

Three paired diseases were observed consistently in both men and women using both methods of analysis: 1) Hypertensive diseases and obesity/other hyperalimentation; 2) spondylopathies and deforming dorsopathies; and 3) dermatitis/eczema and mycoses.

DISCUSSION

In this study have observed differences in the groupings identified with the two most frequently used methods (HCA and EFA). No grouping solution contained exactly the same set of diseases as a cluster and as a factor, although some overlap was observed in both groups. Internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

The multimorbidity patterns obtained by HCA were identified graphically with MDS, allowing us to observe a given hierarchical structure. Nevertheless, internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

EFA, based on tetrachoric correlations where the dichotomous diseases were assumed to come from an underlying mechanism with a continuous variable, produced a wide range of multimorbidity patterns with several levels of correlations. Most of them seem to be highly consistent from a clinical perspective. The multimorbidity patterns obtained with EFA show a main factor (a disease) that has a correlation with another disease, either coexisting or that may occur during the patient's clinical course. [20] Thus, EFA could be more useful for analyzing comorbidity and for describing the correlation between diseases that have a pathophysiological relationship. This approach also may help to answer the question of which condition should be considered the main disease and which the comorbidity.

The HCA results would be useful in generating new hypotheses for intercluster and intracluster associations between diseases that could be applied to the analysis of multimorbidity, defined as the random coexistence of diseases or clusters that indicates significant associations between diseases without a causal explanation. In future studies, other non-hierarchical cluster analysis techniques will improve measurement of the observed distances and multiple interrelationships between different diseases in a given individual.[21] On the other hand, EFA could be more useful for analyzing multimorbidity patterns in the absence of causal comorbidity and for describing visual representation of diseases correlation with a pathophysiological relationship between them.

We obtained two perfect clusters that were common to both men and women: "spondylopathies and deforming dorsopathies" and "urolithiasis and other diseases of the urinary system". In the first cluster, spondylosis is a degenerative disorder that may cause loss of normal spinal structure and function and lead to scoliosis. Nevertheless, many individuals with untreated scoliosis will develop spondylosis; this may be one reason why these diseases were associated.[22] The second cluster can be explained by the complications produced by urolithiasis (such as urinary tract infection, persistent proteinuria, stress incontinence or other unspecified urinary incontinence) and those that have a pathophysiological explanation. [23]

EFA showed that the most frequent pattern in women was infectious diseases. This previously unreported pattern suggests that the multimorbidity patterns obtained in other studies are affected by the type of diseases included in each study.

Although the patterns obtained with both methods did not match exactly, finding matching pairs of diseases by both methods reinforces the idea that patterns of multimorbidity have a dominant disease that associates in some way with other diseases.

In general, it is difficult to compare our results with other studies because of variations in methods, data sources and structures, and populations and diseases studied. Six studies have been performed with HCA[8,21,24–27] and three using EFA.[28–30] Until now, very few analyses of multimorbidity patterns have used multiple methods to compare the same population.[21] The latter study included people aged 50 years and older, considering 11 diseases and using 2 different cluster methods, hierarchical (average linkage) and nonhierarchical (k-medoids), and one method for EFA (principal component analysis). The observed differences between this study and our results can be explained by differences in the underlying statistical formulae and diseases considered in both studies.

The major strength of this study is the dual analysis (HCA and EFA) of a large, high-quality database of primary care records that have been shown to be representative of a much larger population, stratified by sex. Admittedly, this analysis of almost all potential diagnoses may have added a complexity that will hinder interpretation of findings and comparison with other studies, particularly because the boundaries between chronic and acute disease are not always clear.[31,32] Whatever consistency (or discrepancy) we observed was validated by the findings of two different approaches, which helps to identify the most appropriate use of each method in analyzing multimorbidity. We emphasize that the inclusion of chronic and acute diseases is a strength and not a weakness. Because, as we have shown, there are many chronic and acute diseases that coexist at a set time and this has implications for health care.

Internal validation with bootstrap methods, AU p-values, and MDS techniques provided more robust evidence for the cluster analysis. The KMO values obtained show the adequacy of fit of the factor analysis. These values were similar or higher than previous studies.[28,29]

A limitation of this study is our use of agglomerative hierarchical clustering, which forces every unit (i.e., diagnosis) into a single cluster. HCA is exploratory in nature, and different clustering algorithms may produce different results.[33] The final clustering solution presented here was obtained through a systematic and rigorous process: comparing the results from a randomly split dataset, testing different clustering algorithms, and using different objective numeric criteria to decide the number of clusters, internal validation, and graphical representation. In addition, a panel of experts applied subjective clinical criteria to assess the interpretability of the groupings in everyday practice. In addition, EFA is problematic for binary data, which can be grouped because of having similar distributions rather than any common underlying feature. On the other hand, in factor analysis the measure of association incorporates information on both positive and negative concordances [9]. In contrast, the analysis of clusters allows us to show that the occurrence of one or more health conditions can be conditioned by their co-occurrence, without considering negative concordances. [8] Due to the absence of a standard methodology to compare method solutions we have used ad hoc methodology. Finally, another limitation is our use of ICD-10 3-character codes as the unit of analysis, rather than the more specific individual diagnosis, but its use is justified to avoid spurious relationships that more than 10,000 individual codes of the ICD-10 could produce.

This is a cross-sectional study, based on EHR of The Catalan Health Institute and SIDIAP-Q is highly representative for the whole region in terms of both geography, age, gender and diseases, that avoid selection bias.

Multimorbidity can present a problem for health services delivery, affecting patients, health professionals and managers who are attempting to improve service delivery. Our study offers two methodological approaches to understanding the relationships between specific diseases,

which is an essential step in improving our approach to this problem. Although we demonstrated that different analytical methods can yield different results, we also showed that some associations were consistent in both analyses. This study illustrates the need to pay careful attention to the methods used to support policies and decision-making. Clinical guidelines tend to focus on a single disease rather than on multimorbidity, which includes not only diseases but also drug interactions and polypharmacy. The present study confirmed that multimorbidity patterns are a reality in the adult population, and do not apply only to chronic diseases. New guidelines are needed that incorporate multimorbidity into clinical recommendations.

This study was one of the first to compare the two most commonly used methodologies, HCA and EFA, in a large database that includes a large number of diseases. The findings reveal another limitation to be taken into account in comparing multimorbidity patterns between studies: in addition to the spectrum, number and type of diseases included, these patterns vary depending on the method of analysis used.

The results suggest that HCA can be useful to detect multimorbidity patterns and identify different associations between diseases, as the method allows for the possibility that one or more health problems can occur conditionally. On the other hand, EFA seems more applicable to clinical practice because places less restrictions in the diseases grouping, so may be better for generating hypotheses and is more sensitive in identifying clinical associations. Our results suggest that these aspects be considered in planning of future studies, including selection of diseases and methods of analysis.

Finally, our analysis of multimorbidity patterns only considered associations between diseases. Further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

Conclusions

The multimorbidity patterns obtained with EFA show a main factor (i.e., a disease) that has some correlation with the additional diseases in the pattern, suggesting a comorbidity relationship. Meanwhile, the HCA would be useful for in-depth study of multimorbidity pattern. We introduced new evidence about the known limitations of attempts to compare multimorbidity or comorbidity patterns between RWD studies, as our results add information about aspects that must be considered in standardization of future studies: spectrum of diseases, data utilization, and methods of analysis.

Footnotes

Contributors: All authors contributed to the design of the study, revised the article, and approved the final version. CV and QFB obtained the funding. ARL and CV drafted the article. ARL, CV, QFB, TRB, MPV, EPR and JMV contributed to the analysis and interpretation of data. ARL and CV wrote the first draft, and all authors (ARL, CV, QFB, TRB, MPV, EPR and JMV) contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript. All authors read and approved the final manuscript.

Funding: The project has been funded by the *Instituto de Salud Carlos III* of the Ministry of Economy and Competitiveness (Spain) through the Network for Prevention and Health Promotion in Primary Health Care (redIAPP, RD12/0005/0001; RD16/0007/001), and by European Union ERDF funds. Also, by a grant for research projects on health from ISCiii (PI12/00427) and co-financed with European Union ERDF funds). Jose M. Valderas was supported by the National Institute for Health Research Clinician Scientist Award NIHR/CS/010/024. The funders had no role in the study design, collection, analysis and interpretation of data, writing of the manuscript or decision to submit for publication. 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.

Ethical considerations: The study protocol was approved by the Committee on the Ethics of Clinical Research, IDIAP Jordi Gol (Protocol No: P12/28). All data were anonymized and EHR confidentiality was respected in accordance with national and international law.

Acknowledgment: We thank the Catalan Health Institute and especially the SIDIAP Unit, which provided the database for the study. The authors also appreciate the English language review by Elaine Lilly, PhD, and are grateful to Carmen Ibáñez for administrative work.

Availability of data and material: 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 organization is subject to periodic audits to ensure the validity and quality of the data.

References

1. Valderas J, Starfield B, Sibbald B, et al. Defining comorbidity: Implications for understanding health and health services. *Ann Fam Med* 2009;**7**: 357–363.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, et al. Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PLoS One* 2014;**9**: e102149.

3. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, et al. Multimorbidity patterns: A systematic review. *J Clin Epidemiol* 2014;**67**: 254–266.

4. Fortin M, Stewart M, Poitras ME, et al. A systematic review of prevalence studies on multimorbidity: Toward a more uniform methodology. *Ann Fam Med* 2012;**10**: 142–151.

5. Corrao, G., Mancia, G., Generating evidence from computerized health care utilization databases. *Hypertension* 2015; **65**(3):490-8.

6.. Everitt B, Landau S, Leese M, et al. Cluster Analysis. 5th ed. Hoboken: Wiley; 2011.

7. Thompson B. Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington, DC: American Psychological Association; 2004.

8. Cornell JE, Pugh JA, Williams J, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters : clustering binary data from a large administrative medical database. *Appl Multivar Res* 2007;**12**: 163–182.

9. Nunnally J, Berstein I. Psychometric theory. 3rd ed. New York: McGraw-Hill, Inc; 1994.

10.. Del Mar Garcia-Gil M, Hermosilla E, Prieto-Alhambra D, 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**: 135–145.

11. Prieto-Alhambra D, Judge A, Javaid MK, et al. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. *Ann Rheum Dis* 2014;**73**: 1659–64.

12. Ramos R, Balló E, Marrugat J, et al. Validity for use in research on vascular diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA study. *Rev Esp Cardiol* (Engl Ed) 2012;**65**: 29–37.

13. Foguet-Boreu Q, Violán C, Rodriguez-Blanco T, et al. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. *PLoS One* 2015;**10**: e0141155.

14. Foguet-Boreu Q, Violan C, Roso-Llorach A, et al. Impact of multimorbidity: acute morbidity, area of residency and use of health services across the life span in a region of south Europe. *BMC Fam Pract* 2014;**15**: 55.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, 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**: 1–13.

16. Violán C, Foguet-Boreu Q, Roso-Llorach A, et al. Patrones de multimorbilidad en adultos jóvenes en Cataluña: un análisis de clústeres. *Aten Primaria* 2016; **48**:479-92.

17. World Health Organization. ICD-10 International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2007.

18. Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling 32: *Ann Stat* 2004;**32**: 2616–2641.

19. Borg I. Modern Multidimensional Scaling: theory and applications (2nd ed.). 2 nd. Springer-Verlag, editor. New York; 2005.

20. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis* 1970; **23**(7):455-68.

21. Islam M, Valderas J, Yen L, et al. Multimorbidity and comorbidity of chronic diseases among the senior Australians: prevalence and patterns. *PLoS One* 2014;**9**: e83783.

22. Pappou IP, Girardi FP, Sandhu HS, et al. Discordantly high spinal bone mineral density values in patients with adult lumbar scoliosis. *Spine* (Phila Pa 1976). 2006;**31**: 1614–1620.

23. Bartoletti R, Cai T, Mondaini N, et al. Epidemiology and risk factors in urolithiasis. *Urol Int* 2007;**79** Suppl 1: 3–7.

24. Formiga F, Ferrer A, Sanz H, et al. Patterns of comorbidity and multimorbidity in the oldest old: The Octabaix study. *Eur J Intern Med* 2013;**24**: 40–44.

25. John R, Kerby DS and Hennessy CH. Patterns and impact of comorbidity and multimorbidity among community-resident American Indian elders. *Gerontologist* 2003;**43**: 649–660.

26. Marengoni A, Rizzuto D, Wang HX, et al. Patterns of chronic multimorbidity in the elderly population. *J Am Geriatr Soc* 2009;**57**: 225–230.

27. Newcomer SR, Steiner JF and Bayliss EA. Identifying subgroups of complex patients with cluster analysis. *Am J Manag Care* 2011;**17**: e324–32.

28. Schäfer I, von Leitner EC, Schön G, et al. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS One* 2010;**5**: e15941.

29. Andre L, Prados-torres A, Poblador-plou B, et al. Multimorbidity Patterns in Primary Care : Interactions among Chronic Diseases Using Factor Analysis. *PLoS One* 2012;**7**(2): e32190.

30. Poblador-Plou B, van den Akker M, Vos R, et al. Similar multimorbidity patterns in primary care patients from two European regions: results of a factor analysis. *PLoS One* 2014;**9**: e100375.

31. O'Halloran J, Miller GC and Britt H. Defining chronic conditions for primary care with ICPC-2. *Fam Pract* 2004;**21**: 381–386.

32. Lamberts H and Wood M. International Classification of Primary Care. Oxford. Press OU, editor. Oxford; 2011.

33. Aldenderfer MS and Blashfield RK. Cluster Analysis: Quantitative Applications in the Social Sciences. 1984.

Figure 1

Figure 1a. Top four clusters in women (n= 217,823) aged 45-64 years, analyzed with multidimensional scaling.

Figure 1b. Top four clusters in men (n= 191,171) aged 45-64 years, analyzed with multidimensional scaling.

Appendix

Appendix 1. Study Flow Chart.

Appendix 2. Cluster dendrogram in women aged 45-64 years.

Appendix 3. Cluster dendrogram in men aged 45-64 years, with approximately unbiased (AU) probability value.

Appendix 4. Prevalence and composition of diagnostic clusters in women aged 45-64 years (n= 217,823).

Appendix 5. Prevalence and composition of diagnostic clusters in men aged 45-64 years (n= 191,171).

Appendix 6. Factors in women aged 45-64 years (n= 217,823).

Appendix 7. Factors in men aged 45-64 years (n= 191,171).

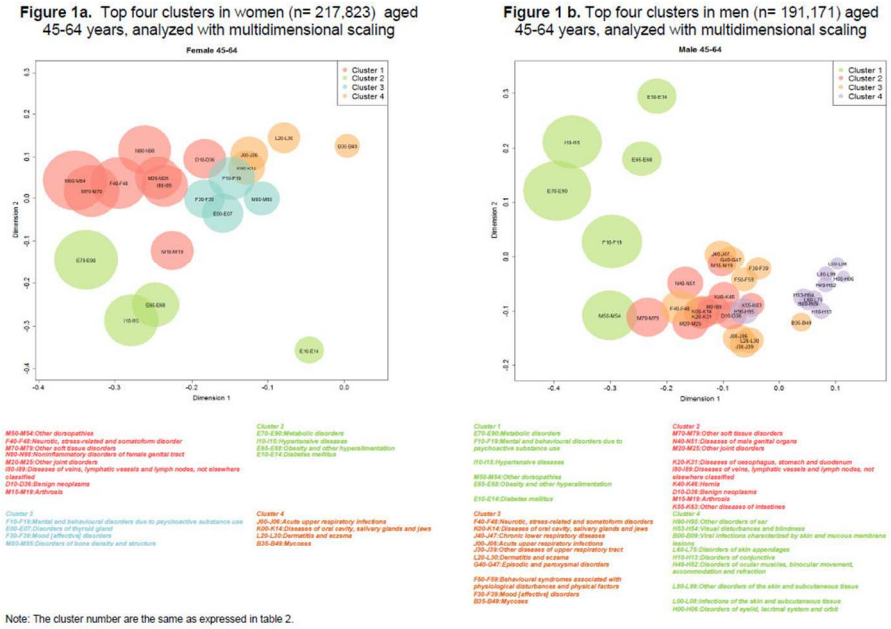


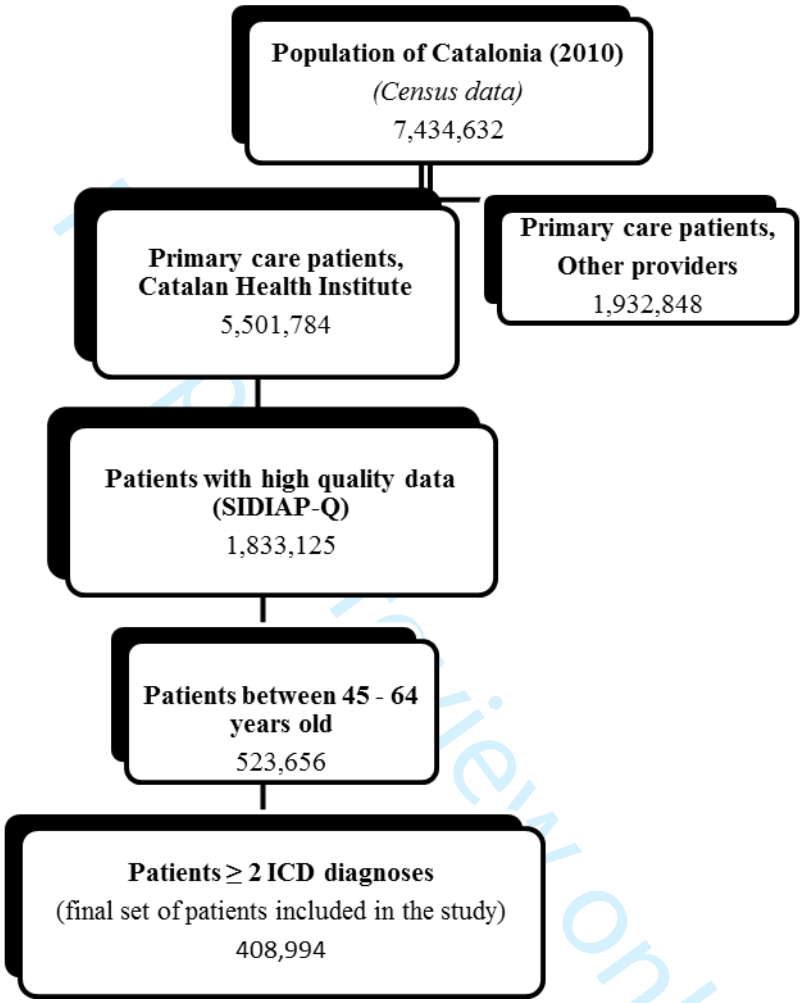
Figure 1

Figure 1a. Top four clusters in women (n= 217,823) aged 45-64 years, analyzed with multidimensional scaling.

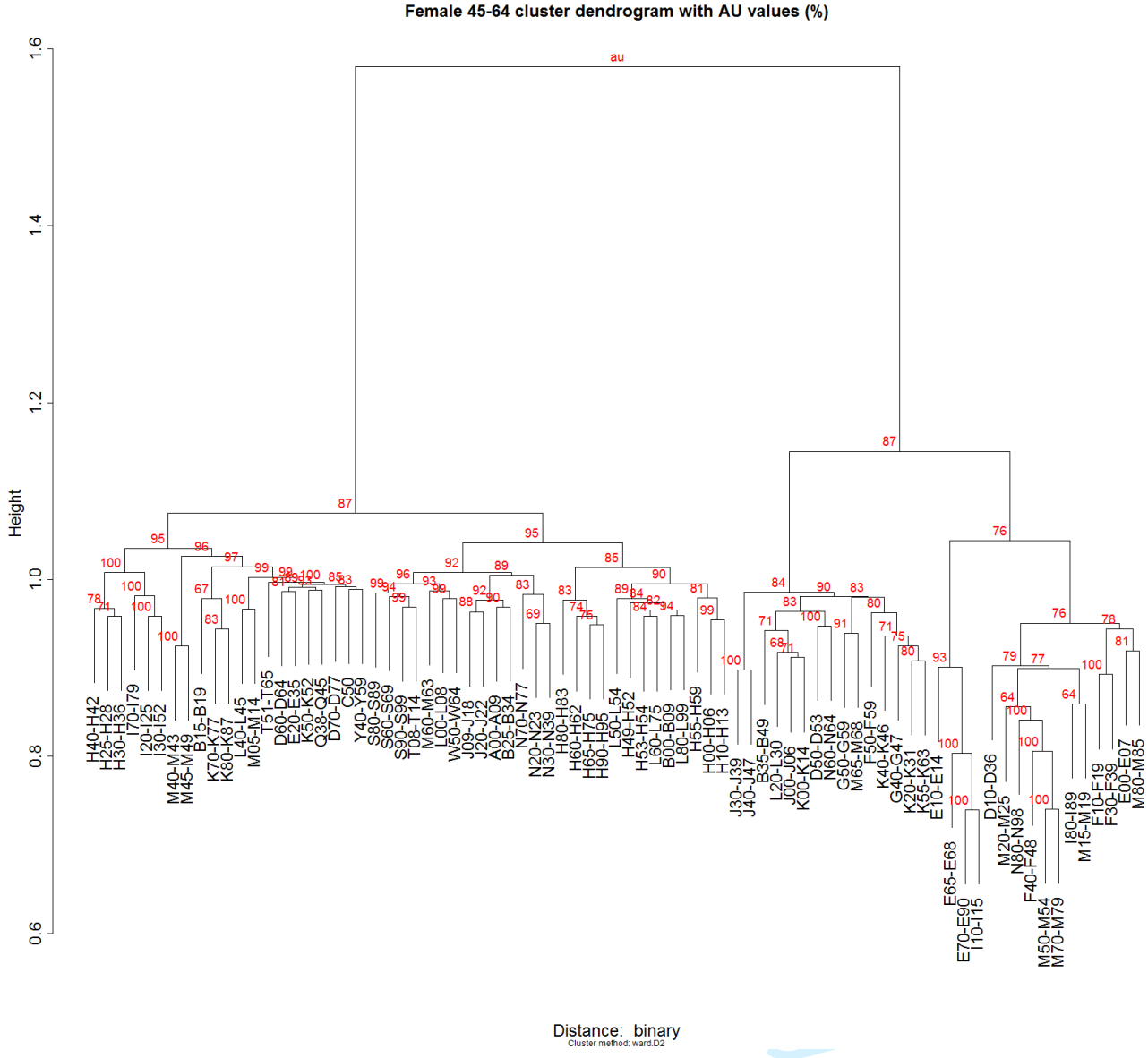
Figure 1b. Top four clusters in men (n= 191,171) aged 45-64 years, analyzed with multidimensional scaling.

97x64mm (300 x 300 DPI)

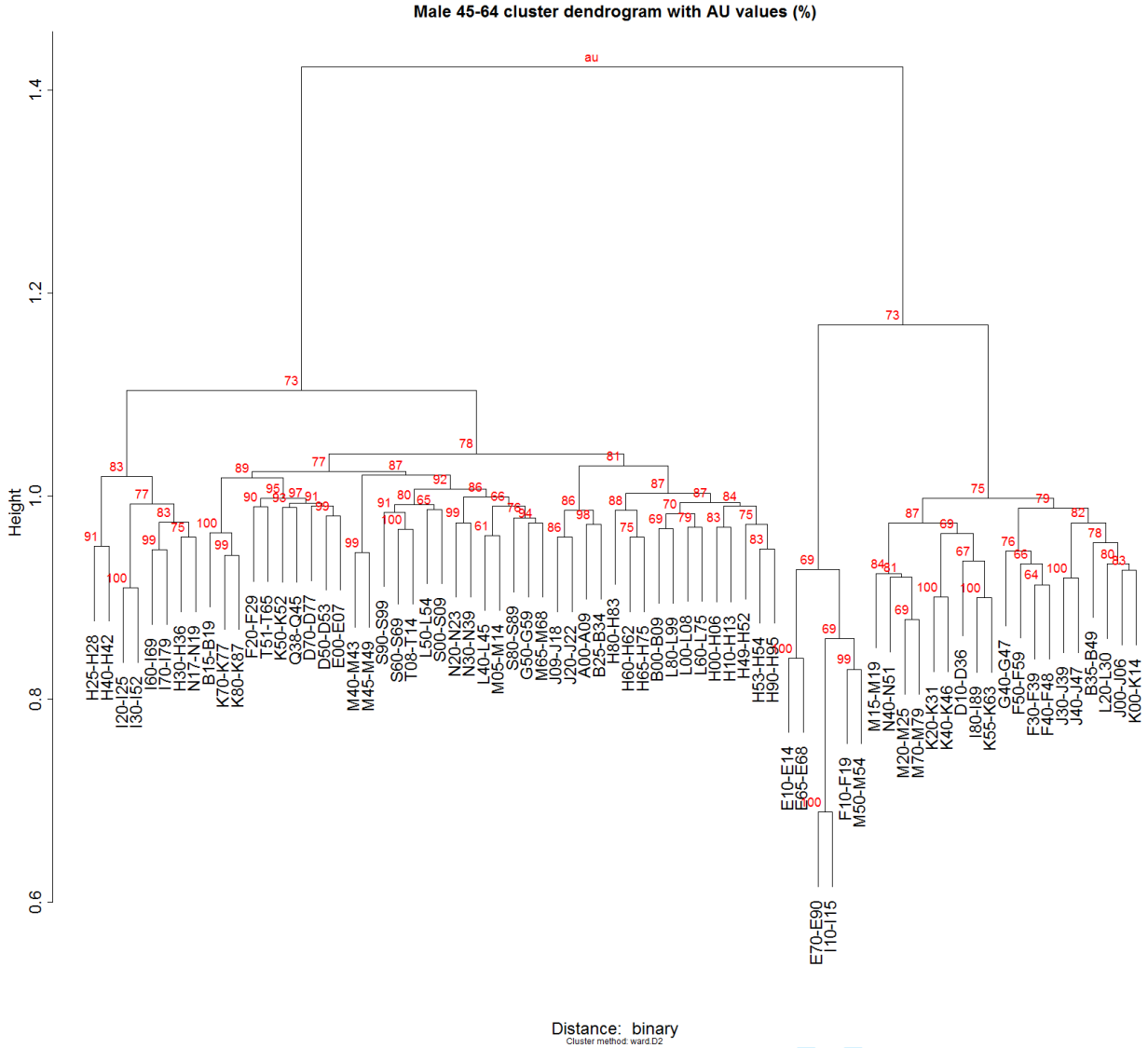
Appendix 1. Study Flow Chart.



Appendix 2. Cluster dendrogram in women aged 45-64 years.



Appendix 3. Cluster dendrogram in men aged 45-64 years, with approximately unbiased (AU) probability value.



Appendix 4. Prevalence and composition of diagnostic clusters in women aged 45-64 years (n= 217,823).

Cluster order number, n	Prevalence 1 (%)*	Prevalence 2 (%)†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value**
WC1^ 178,849	82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	M54 Dorsalgia	0.79 (0.74-0.85)
			F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	F41 Other anxiety disorders	
			M70-M79:Other soft tissue disorders	27.0	32.8	M79 Other soft tissue disorders, not elsewhere classified	
			N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	N95 Menopausal and other perimenopausal disorders	
			M20-M25:Other joint disorders	18.6	22.6	M25 Other joint disorders, not elsewhere classified	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	I83 Varicose veins of lower extremities	
			D10-D36:Benign neoplasms	16.2	19.7	D25 Leiomyoma of uterus	
			M15-M19:Arthrosis	15.7	19.1	M17 Gonarthrosis [arthrosis of knee]	
WC2 121,564	55.8	23.0	E70-E90:Metabolic disorders	35.4	63.4	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.93 (0.86-1.00)
			I10-I15:Hypertensive diseases	25.6	45.8	I10 Essential (primary) hypertension	
			E65-E68:Obesity and other hyperalimentation	19.0	34.0	E66 Obesity	
			E10-E14:Diabetes mellitus	7.7	13.7	E11 Non-insulin-	

						dependent diabetes mellitus	
WC3 103,147	47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	39.4	F17 Mental and behavioural disorders due to use of tobacco	0.78 (0.73-0.84)
			E00-E07:Disorders of thyroid gland	14.9	31.4	E03 Other hypothyroidism	
			F30-F39:Mood [affective] disorders	14.6	30.8	F32 Depressive episode	
			M80-M85:Disorders of bone density and structure	11.3	23.9	M81 Osteoporosis without pathological fracture	
WC4 70,294	32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	J00 Acute nasopharyngitis [common cold]	0.71 (0.66-0.77)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	K02 Dental caries	
			L20-L30:Dermatitis and eczema	9.3	28.8	L29 Pruritus	
			B35-B49:Mycoses	5.7	17.8	B35 Dermatophytosis	
WC5 63,792	29.3	6.0	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	38.9	K30 Dyspepsia	0.71 (0.66-0.76)
			G40-G47:Episodic and paroxysmal disorders	10.5	36.0	G43 Migraine	
			K55-K63:Other diseases of intestines	8.6	29.2	K59 Other functional intestinal disorders	
			K40-K46:Hernia	5.7	19.4	K44 Diaphragmatic hernia	
WC6 34,707	15.9	1.6	J30-J39:Other diseases of upper respiratory tract	9.4	59.0	J30 Vasomotor and allergic rhinitis	1.00
			J40-J47:Chronic lower respiratory diseases	8.2	51.2	J45 Asthma	
WC7 32,674	15.0	0.8	D50-D53:Nutritional anaemias	8.3	55.2	D50 Iron deficiency anaemia	1.00
			N60-N64:Disorders of breast	7.5	50.1	N60 Benign mammary	

						dysplasia	
WC8 27,150	12.5	0.8	G50-G59:Nerve, nerve root and plexus disorders	8.5	68.6	G56 Mononeuropathies of upper limb	0.91 (0.88-0.94)
			M65-M68:Disorders of synovium and tendon	4.7	37.5	M65 Synovitis and tenosynovitis	
WC9 19,199	8.8	0.4	N30-N39:Other diseases of urinary system	5.9	66.7	N39 Other disorders of urinary system	0.69 (0.62-0.75)
			N20-N23:Urolithiasis	3.4	38.3	N20 Calculus of kidney and ureter	
WC10 18,439	8.5	0.4	H90-H95:Other disorders of ear	6.3	75.0	H91 Other hearing loss	0.75 (0.70-0.80)
			H65-H75:Diseases of middle ear and mastoid	2.5	30.1	H65 Nonsuppurative otitis media	
WC11 16,535	7.6	0.6	M45-M49:Spondylopathies	4.3	57.1	M47 Spondylosis	1.00
			M40-M43:Deforming dorsopathies	3.8	50.4	M41 Scoliosis	
WC12 14, 335	6.6	0.4	K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	60.9	K80 Cholelithiasis	0.83*
			K70-K77:Diseases of liver	2.9	44.7	K76 Other diseases of liver	
WC13 12,320	5.7	0.0	F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	5.7	100.0	F51 Nonorganic sleep disorders	
WC14 10,357	4.8	0.0	L60-L75:Disorders of skin appendages	4.8	100.0	L72 Follicular cysts of skin and subcutaneous tissue	
WC15 9,555	4.4	0.0	H53-H54:Visual disturbances and blindness	4.4	100.0	H54 Blindness and low vision	
WC16 9,521	4.4	0.0	I30-I52:Other forms of heart disease	4.4	100.0	I34 Nonrheumatic mitral valve disorders	
WC17 9,442	4.3	0.0	B00-B09:Viral infections characterized by skin and mucous membrane lesions	4.3	100.0	B07 Viral warts	
WC18 8,541	3.9	0.0	T08-T14:Injuries to unspecified part of trunk, limb or body region	3.9	100.0	T14 Injury of unspecified body region	

WC19 8,268	3.8	0.0	H10-H13:Disorders of conjunctiva	3.8	100.0	H10 Conjunctivitis	
WC20 7,866	3.6	0.0	H60-H62:Diseases of external ear	3.6	100.0	H61 Other disorders of external ear	
WC21 7,714	3.5	0.0	H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	3.5	100.0	H52 Disorders of refraction and accommodation	
WC22 7,512	3.4	0.0	L50-L54:Urticaria and erythema	3.4	100.0	L50 Urticaria	
WC23 7,298	3.4	0.0	J20-J22:Other acute lower respiratory infections	3.4	100.0	J20 Acute bronchitis	
WC24 7,251	3.3	0.0	L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	100.0	L82 Seborrhoeic keratosis	
WC25 6,178	2.8	0.0	H00-H06:Disorders of eyelid, lacrimal system and orbit	2.8	100.0	H04 Disorders of lachrymal system	
WC26 5,993	2.8	0.0	H40-H42:Glaucoma	2.8	100.0	H40 Glaucoma	
WC27 5,986	2.7	0.0	A00-A09:Intestinal infectious diseases	2.7	100.0	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
WC28 5,890	2.7	0.0	M05-M14:Inflammatory polyarthropathies	2.7	100.0	M13 Other arthritis	
WC29 5,548	2.5	0.0	S80-S89:Injuries to the knee and lower leg	2.5	100.0	S83 Dislocation, sprain and strain of joints and ligaments of knee	
WC30 5,416	2.5	0.0	C50-C50:Malignant neoplasm of breast	2.5	100.0	C50 Malignant neoplasm of breast	
WC31 5,372	2.5	0.0	L40-L45:Papulosquamous disorders	2.5	100.0	L40 Psoriasis	
WC32 5,336	2.4	0.0	H80-H83:Diseases of inner ear	2.4	100.0	H81 Disorders of vestibular function	
WC33	2.0	0.0	S90-S99:Injuries to the ankle and	2.0	100.0	S93 Dislocation, sprain	

4,374			foot			and strain of joints and ligaments at ankle and foot level	
WC34 3,882	1.8	0.0	B15-B19:Viral hepatitis	1.8	100.0	B18 Chronic viral hepatitis	
WC35 3,758	1.7	0.0	H25-H28:Disorders of lens	1.7	100.0	H25 Senile cataract	
WC36 3,744	1.7	0.0	K50-K52:Noninfective enteritis and colitis	1.7	100.0	K52 Other non-infective gastro-enteritis and colitis	
WC37 3,689	1.7	0.0	N70-N77:Inflammatory diseases of female pelvic organs	1.7	100.0	N76 Other inflammation of vagina and vulva	
WC38 3,646	1.7	0.0	J09-J18:Influenza and pneumonia	1.7	100.0	J11 Influenza, virus not identified	
WC39 3,399	1.6	0.0	L00-L08:Infections of the skin and subcutaneous tissue	1.6	100.0	L03 Cellulitis	
WC40 3,284	1.5	0.0	Q38-Q45:Other congenital malformations of the digestive system	1.5	100.0	Q40 Other congenital malformations of upper alimentary tract	
WC41 3,180	1.5	0.0	D60-D64:Aplastic and other anaemias	1.5	100.0	D64 Other anaemias	
WC42 3,055	1.4	0.0	W50-W64:Exposure to animate mechanical forces	1.4	100.0	W57 Bitten or stung by non-venomous insect and other non-venomous arthropods	
WC43 2,926	1.3	0.0	Y40-Y59:Drugs, medicaments and biological substances causing adverse effects in therapeutic use	1.3	100.0	Y52 Agents primarily affecting the cardiovascular system	
WC44 2,884	1.3	0.0	B25-B34:Other viral diseases	1.3	100.0	B34 Viral infection of unspecified site	
WC45 2,788	1.3	0.0	E20-E35:Disorders of other endocrine glands	1.3	100.0	E28 Ovarian dysfunction	
WC46 2,765	1.3	0.0	S60-S69:Injuries to the wrist and hand	1.3	100.0	S61 Open wound of wrist and hand	

WC47 2,645	1.2	0.0	H55-H59:Other disorders of eye and adnexa	1.2	100.0	H57 Other disorders of eye and adnexa	
WC48 2,612	1.2	0.0	M60-M63:Disorders of muscles	1.2	100.0	M62 Other disorders of muscle	
WC49 2,600	1.2	0.0	H30-H36:Disorders of choroid and retina	1.2	100.0	H35 Other retinal disorders	
WC50 2,584	1.2	0.0	D70-D77:Other diseases of blood and blood-forming organs	1.2	100.0	D72 Other disorders of white blood cells	
WC51 2,508	1.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	1.2	100.0	T65 Toxic effect of other and unspecified substances	
WC52 2,309	1.1	0.0	I70-I79:Diseases of arteries, arterioles and capillaries	1.1	100.0	I73 Other peripheral vascular diseases	
WC53 2,241	1.03	0.0	I20-I25:Ischaemic heart diseases	1.0	100.0	I25 Chronic ischaemic heart disease	

* Standard error to large
**Approximately unbiased (AU) probability-value
^ Abbreviation of sex, method and number (WC1: Women Cluster 1)

Appendix 5. Prevalence and composition of diagnostic clusters in men aged 45-64 years (n= 191,171).

Cluster order number, n	Prevalence 1 .%*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value*
MC1^ 160,216	83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.69 (0.64-0.75)
			F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	F17 Mental and behavioural disorders due to use of tobacco	
			I10-I15:Hypertensive diseases	32.5	38.8	I10 Essential (primary) hypertension	
			M50-M54:Other dorsopathies	27.8	33.2	M54 Dorsalgia	
			E65-E68:Obesity and other hyperalimentation	14.6	17.4	E66 Obesity	
			E10-E14:Diabetes mellitus	14.2	16.9	E11 Non-insulin-dependent diabetes mellitus	
MC2 110,200	57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	M75 Shoulder lesions	0.87 (0.84-0.90)
			N40-N51:Diseases of male genital organs	12.1	21.0	N40 Hyperplasia of prostate	
			M20-M25:Other joint disorders	12.0	20.9	M25 Other joint disorders, not elsewhere classified	
			K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	K29 Gastritis and duodenitis	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	I84 Haemorrhoids	
			K40-K46:Hernia	8.8	15.2	K40 Inguinal hernia	

			D10-D36:Benign neoplasms	8.6	14.9	D17 Benign lipomatous neoplasm	
			M15-M19:Arthrosis	7.7	13.4	M17 Gonarthrosis [arthrosis of knee]	
			K55-K63:Other diseases of intestines	6.4	11.1	K62 Other diseases of anus and rectum	
MC3 103,365	54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	F41 Nutritional marasmus	0.79 (0.74-0.84)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	K02 Dental caries	
			J40-J47:Chronic lower respiratory diseases	9.3	17.2	J44 Other chronic obstructive pulmonary disease	
			J00-J06:Acute upper respiratory infections	8.9	16.4	J00 Acute nasopharyngitis [common cold]	
			J30-J39:Other diseases of upper respiratory tract	8.0	14.8	J30 Vasomotor and allergic rhinitis	
			L20-L30:Dermatitis and eczema	7.5	13.9	L21 Seborrhoeic dermatitis	
			G40-G47:Episodic and paroxysmal disorders	7.4	13.7	G47 Sleep disorders	
			F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	F52 Sexual dysfunction, not caused by organic disorder or disease	
			F30-F39:Mood [affective] disorders	6.3	11.6	F32 Depressive episode	
			B35-B49:Mycoses	4.1	7.6	B35 Dermatophytosis	
MC4 48,140	25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	H91 Disorders of sclera and cornea in diseases classified elsewhere	0.87 (0.83-0.91)
			H53-H54:Visual disturbances and blindness	3.9	15.5	H54 Blindness and low vision	
			B00-B09:Viral infections characterized by skin and	3.5	13.9	B07 Viral warts	

			mucous membrane lesions				
			L60-L75:Disorders of skin appendages	3.5	13.9	L72 Follicular cysts of skin and subcutaneous tissue	
			H10-H13:Disorders of conjunctiva	3.0	12.0	H10 Conjunctivitis	
			H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	H52 Disorders of refraction and accommodation	
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	L82 Seborrhoeic keratosis	
			L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	L02 Cutaneous abscess, furuncle and carbuncle	
			H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	H00 Hordeolum and chalazion	
MC5 30,315	15.9	2.9	I30-I52:Other forms of heart disease	6.9	43.7	I45 Other conduction disorders	0.77 (0.73-0.82)
			I20-I25:Ischaemic heart diseases	5.0	31.0	I25 Chronic ischaemic heart disease	
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	15.2	I73 Other peripheral vascular diseases	
			I60-I69:Cerebrovascular diseases	1.8	11.4	I64 Stroke, not specified as haemorrhage or infarction	
			H30-H36:Disorders of choroid and retina	1.7	11.0	H36 Retinal disorders in diseases classified elsewhere	
			N17-N19:Renal failure	1.5	9.2	N18 Chronic renal failure	
MC6 29,907	15.6	1.5	M05-M14:Inflammatory polyarthropathies	5.4	34.5	M10 Gout	0.66 (0.60-0.72)
			S80-S89:Injuries to the knee and lower leg	3.2	20.7	S83 Dislocation, sprain and strain of joints and ligaments of knee	

			L40-L45:Papulosquamous disorders	3.2	20.5	L40 Psoriasis	
			G50-G59:Nerve, nerve root and plexus disorders	2.7	17.4	G56 Mononeuropathies of upper limb	
			M65-M68:Disorders of synovium and tendon	2.6	16.8	M65 Synovitis and tenosynovitis	
MC7 19,313	10.1	0.9	K70-K77:Diseases of liver	5.2	51.9	K76 Other diseases of liver	1.00
			B15-B19:Viral hepatitis	3.2	32.0	B18 Chronic viral hepatitis	
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	2.6	25.8	K80 Cholelithiasis	
MC8 19,160	10.0	0.7	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	41.0	T14 Injury of unspecified body region	0.80 (0.75-0.86)
			L50-L54:Urticaria and erythema	2.1	20.8	L50 Urticaria	
			S60-S69:Injuries to the wrist and hand	1.9	18.5	S61 Open wound of wrist and hand	
			S90-S99:Injuries to the ankle and foot	1.4	13.9	S93 Dislocation, sprain and strain of joints and ligaments at ankle and foot level	
			S00-S09:Injuries to the head	1.3	13.3	S01 Open wound of head	
MC9 13,752	7.2	0.3	E00-E07:Disorders of thyroid gland	2.4	32.8	E03 Other hypothyroidism	0.97 (0.96-0.98)
			K50-K52:Noninfective enteritis and colitis	1.7	24.1	K52 Other non-infective gastro-enteritis and colitis	
			D50-D53:Nutritional anaemias	1.2	16.3	D50 Iron deficiency anaemia	
			Q38-Q45:Other congenital malformations of the digestive system	1.2	16.2	Q40 Other congenital malformations of upper alimentary tract	
			D70-D77:Other diseases of blood and blood-forming organs	1.1	15.3	D72 Other disorders of white blood cells	
MC10	7.1	0.5	J20-J22:Other acute lower	2.6	37.2	J20 Acute bronchitis	0.86 (0.82-0.91)

13,490			respiratory infections				
			A00-A09:Intestinal infectious diseases	2.4	34.2	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
			J09-J18:Influenza and pneumonia	1.6	22.0	J11 Influenza, virus not identified	
			B25-B34:Other viral diseases	1.0	14.6	B34 Viral infection of unspecified site	
MC11 13,434	7.0	0.4	H60-H62:Diseases of external ear	3.9	55.3	H61 Other disorders of external ear	0.88 (0.85-0.92)
			H65-H75:Diseases of middle ear and mastoid	2.1	30.5	H65 Nonsuppurative otitis media	
			H80-H83:Diseases of inner ear	1.4	19.9	H81 Disorders of vestibular function	
MC12 10,952	5.7	0.1	N20-N23:Urolithiasis	4.3	75.5	N20 Calculus of kidney and ureter	0.99 (0.98-1.00)
			N30-N39:Other diseases of urinary system	1.6	27.1	N39 Other disorders of urinary system	
MC13 8,794	4.6	0.3	M45-M49:Spondylopathies	3.1	66.7	M47 Spondylosis	0.99 (0.95-1.00)
			M40-M43:Deforming dorsopathies	1.8	38.9	M41 Scoliosis	
MC14 7,444	3.9	0.2	H40-H42:Glaucoma	2.3	59.6	H40 Glaucoma	0.91 (0.88-0.95)
			H25-H28:Disorders of lens	1.8	45.4	H25 Senile cataract	
MC15 6,161	3.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	2.1	65.8	T65 Toxic effect of other and unspecified substances	0.90 (0.87-0.93)
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	35.3	F20 Schizophrenia	

**Approximately unbiased (AU) probability-value

Appendix 6. Factors in women aged 45-64 years (n= 217,823).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WF1^ 130,072	59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
			M70-M79:Other soft tissue disorders	27.0	45.2		
			M15-M19:Arthrosis	15.7	26.2		
			G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
			M45-M49:Spondylopathies	4.3	7.3		
			M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 82,301	37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
			E65-E68:Obesity and other hyperalimentation	19.0	50.2		
			E10-E14:Diabetes mellitus	7.7	20.3		
WF3 71,436	32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
			E10-E14:Diabetes mellitus	7.7	23.4		
			I30-I52:Other forms of heart disease	4.4	13.3		
			H25-H28:Disorders of lens	1.7	5.3		
			H30-H36:Disorders of choroid and retina	1.2	3.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
			I20-I25:Ischaemic heart diseases	1.0	3.1		
WF4 60,027	27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
			N30-N39:Other diseases of urinary system	5.9	21.3		
			H60-H62:Diseases of external ear	3.6	13.1		
			J20-J22:Other acute lower respiratory infections	3.4	12.2		
			A00-A09:Intestinal infectious diseases	2.7	10.0		
			H65-H75:Diseases of middle ear and mastoid	2.5	9.2		

			J09-J18:Influenza and pneumonia	1.7	6.1		
			B25-B34:Other viral diseases	1.3	4.8		
			M60-M63:Disorders of muscles	1.2	4.4		
WF5 56,671	26.0	5.3	L20-L30:Dermatitis and eczema	9.3	35.7	8.4	77.5
			B35-B49:Mycoses	5.7	22.1		
			L60-L75:Disorders of skin appendages	4.8	18.3		
			H53-H54:Visual disturbances and blindness	4.4	16.9		
			H10-H13:Disorders of conjunctiva	3.8	14.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	12.8		
			H55-H59:Other disorders of eye and adnexa	1.2	4.7		
WF6 46,391	21.3	3.8	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	53.6	5.9	90.4
			K40-K46:Hernia	5.7	26.7		
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	18.8		
			K70-K77:Diseases of liver	2.9	13.8		
			Q38-Q45:Other congenital malformations of the digestive system	1.5	7.1		
WF7 41,492	19.0	0.6	M80-M85:Disorders of bone density and structure	11.3	59.5	5.1	95.5
			D50-D53:Nutritional anaemias	8.3	43.5		
WF8 40,619	18.6	0.0	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	100.0	4.3	99.8
WF9 23,729	10.9	0.6	J40-J47:Chronic lower respiratory diseases	8.2	74.9	4.1	103.9
			J20-J22:Other acute lower respiratory infections	3.4	30.8		

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor
^ Abbreviation of sex, method and number (WF1: Women Factor 1)

Appendix 7. Factors in men aged 45-64 years (n= 191,171).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in age, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
MF1^ 118,037	61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
			I10-I15:Hypertensive diseases	32.6	52.7		
			E65-E68:Obesity and other hyperalimentation	14.6	23.6		
			M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 75,315	39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
			I30-I52:Other forms of heart disease	6.9	17.6		
			I20-I25:Ischaemic heart diseases	5.0	12.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
			I60-I69:Cerebrovascular diseases	1.8	4.6		
			N17-N19:Renal failure	1.5	3.7		
MF3 73,638	38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
			K70-K77:Diseases of liver	5.2	13.6		
			B15-B19:Viral hepatitis	3.2	8.4		
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 66,303	34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
			M15-M19:Arthrosis	7.7	22.2		
			M45-M49:Spondylopathies	3.1	8.8		
			M40-M43:Deforming dorsopathies	1.8	5.2		
MF5 34,903	18.3	2.5	L20-L30:Dermatitis and eczema	7.5	41.3	22.2	50.7
			B35-B49:Mycoses	4.1	22.5		
			H53-H54:Visual disturbances and	3.9	21.3		

			blindness				
			H10-H13:Disorders of conjunctiva	3.0	16.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	13.8		
MF6 27,697	14.5	1.8	J00-J06:Acute upper respiratory infections	8.9	61.3	10.4	61.0
			J20-J22:Other acute lower respiratory infections	2.6	18.1		
			A00-A09:Intestinal infectious diseases	2.4	16.7		
			J09-J18:Influenza and pneumonia	1.6	10.7		
			B25-B34:Other viral diseases	1.0	7.1		
MF7 33,568	17.6	2.2	E10-E14:Diabetes mellitus	14.2	80.8	9.5	70.5
			H40-H42:Glaucoma	2.3	13.2		
			H25-H28:Disorders of lens	1.8	10.1		
			H30-H36:Disorders of choroid and retina	1.7	10.0		
MF8 25,121	13.1	0.7	K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	87.8	6.5	84.3
			D50-D53:Nutritional anaemias	1.2	8.9		
			Q38-Q45:Other congenital malformations of the digestive system	1.2	8.9		
MF9 15,974	8.4	0.4	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	49.2	5.0	99.8
			S80-S89:Injuries to the knee and lower leg	3.3	38.7		
			S90-S99:Injuries to the ankle and foot	1.4	16.7		
MF10 8,271	4.3	0.0	N20-N23:Urolithiasis	4.3	100.0	3.9	103.7

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor

^ Abbreviation of sex, method and number (MF1: Men Factor 1)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*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	6-8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	
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	6-9
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	6-9
		(*e*) Describe any sensitivity analyses	6-9
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	10
		(b) Give reasons for non-participation at each stage	10
		(c) Consider use of a flow diagram	Supplementary File Appendix 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	Tables
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	Tables 2 and 3.
		(*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	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	23
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	20-21
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	19
Generalisability	21	Discuss the generalisability (external validity) of the study results	20
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	23

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

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.

BMJ Open

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018986.R2
Article Type:	Research
Date Submitted by the Author:	16-Jan-2018
Complete List of Authors:	Roso-Llorach, Albert; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), ; Universitat Autònoma de Barcelona, Violan-Fors, Concepción; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Foguet-Boreu, Quintí; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol) Universitat Autònoma de Barcelona; Universitat Autonoma de Barcelona Rodriguez-Blanco, Teresa; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pons-Vigués, Mariona; Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol); Universitat Autonoma de Barcelona Pujol-Ribera, Enriqueta; Institut Universitari d'Investigació en Atenció Primaria (IDIAP-Jordi Gol); Universitat Autonoma de Barcelona Valderas, Jose; University of Exeter Medical School, Health Services & Policy Research Group, Academic Collaboration for Primary Care
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Research methods, General practice / Family practice, Anaesthesia
Keywords:	Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

1. Albert Roso-Llorach. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: aroso@idiapjgol.org

2. Concepción Violán. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain.
E-mail: cviolan@idiapjgol.org

3. Quintí Foguet-Boreu. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 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, Catalonia.
E-mail: 42292qfb@comb.cat

4. Teresa Rodriguez-Blanco. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol). Universitat Autònoma de Barcelona, Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. E-mail: trodriguez@idiapjgol.org

5. Mariona Pons-Vigués. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain.
E-mail: mponsv@idiapjgol.info

6. Enriqueta Pujol-Ribera. 1. Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Gran Via Corts Catalanes, 587 àtic, 08007 Barcelona, Spain. 2. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain. 3. Faculty of Nursing, University of Girona, Emili Grahit, 77, 17071 Girona, Spain. E-mail: epujol@idiapjgol.org

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

Corresponding author: Albert Roso-Llorach. IDIAP Jordi Gol
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: aroso@idiapjgol.org

Word count: 4088

Comparative analysis of methods for identifying multimorbidity patterns: A study of 'real world' data.

ABSTRACT

Objective The aim was to compare multimorbidity patterns identified with the two most commonly used methods: hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA) in a large primary care database. Specific objectives were 1) to determine whether choice of method affects the composition of these patterns and 2) to consider the potential application of each method in the clinical setting.

Design Cross-sectional study. Diagnoses were based on the 263 corresponding blocks of the International Classification of Diseases (ICD-10) . Multimorbidity patterns were identified using HCA and EFA. Analysis was stratified by sex, and results compared for each method.

Setting and participants Electronic health records for 408,994 patients with multimorbidity aged 45-64 years in 274 primary health care teams from 2010 in Catalonia, Spain.

Results HCA identified 53 clusters for women, with just 12 clusters including at least 2 diagnoses, and 15 clusters for men, all of them including at least 2 diagnoses. EFA showed 9 factors for women and 10 factors for men. We observed differences by sex and method of analysis, although some patterns were consistent. Three combinations of diseases were observed consistently across sex groups and across both methods: hypertension and obesity, spondylopathies and deforming dorsopathies, and dermatitis eczema and mycosis.

Conclusions This study showed that multimorbidity patterns vary depending on the method of analysis used (HCA vs EFA) and provided new evidence about the known limitations of attempts to compare multimorbidity patterns in Real World Data studies. We found that EFA was useful in describing comorbidity relationships and HCA could be useful for in-depth study of multimorbidity. Our results suggest possible applications for each of these methods in clinical and research settings and add information about some aspects that must be considered in standardization of future studies: spectrum of diseases, data utilization, and methods of analysis.

Keywords: Multimorbidity, Cluster analysis, Factor analysis, Primary health care, Electronic health records, diseases

Strengths and limitations of this study

- This one of the first studies to compare the two methodologies most commonly used to obtain patterns of multimorbidity, Hierarchical Cluster Analysis and Exploratory Factor Analysis.
- The dual analysis was performed in a large, high-quality database of primary care electronic health records that have been shown to be representative of a much larger population, stratified by sex.
- Internal validation with bootstrap methods provided more robust evidence for the cluster analysis.
- The agglomerative hierarchical clustering forces every unit into a single cluster, is exploratory in nature and different clustering algorithms may produce different results.
- The study is cross-sectional and further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

INTRODUCTION

The reliable identification of patterns of multimorbidity is a critical step in developing health care services sensitive to the health needs of these patients.[1] Recent reviews of multimorbidity patterns have shown that individual studies differ widely in their design and choice of epidemiological and statistical methods, including sampling frameworks and selection criteria, coding systems, eligible diseases, and definition of disease clustering patterns.[2–4] These studies highlight the lack of consensus to measure patterns of comorbidity and multimorbidity. In recent years, the number of studies based on real-world data (RWD) [5] has increased significantly, which makes it even more difficult to establish a consensus on how to measure comorbidity and multimorbidity patterns. Although much more information is available, the different databases may not be comparable, making it difficult to arrive at observations and draw firm conclusions. It also limits our ability to compare analyses using RWD and to evaluate whether one approach may be better suited to the purpose. Therefore, it is difficult to identify multimorbidity patterns and provide adequate health services according to the population needs.

The most frequent methods used to date have been hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA), which offer very different approaches and solutions.[2,3] Both are descriptive methods to identify association of diagnoses and determine patterns of multimorbidity. HCA obtains the patterns of multimorbidity from the dissimilarities between diseases, while EFA is based on correlations between diagnoses to identify the patterns. The HCA clusters tend to contain diagnoses that are similar to each other (in terms of Euclidean distances), but dissimilar from the diagnoses in other clusters; no diagnosis can be included in more than one cluster. In contrast, EFA along with confirmatory factor analysis are primarily used to test hypothesized relationships between observed measures and latent constructs. In

addition, EFA allows for inclusion of any diagnosis in multiple factors as they can present significant correlations with more than one factor. [6-9]

For all these reasons, the aim was to compare multimorbidity patterns identified with the two most commonly used methods: hierarchical cluster analysis (HCA) and exploratory factor analysis (EFA) in a large primary care database. Specific objectives were 1) to determine whether choice of method affects the composition of these paterns and 2) to consider the potential application of each method in the clinical setting.

METHODS

Design, setting and study population

A cross-sectional study was conducted in Catalonia (Spain), a Mediterranean region with 7,434,632 inhabitants, 81% of which live in urban municipalities (2010 census). The Spanish National Health Service (NHS) provides universal coverage, financed mainly by tax revenue. The Catalan Health Institute (CHI) manages primary health care teams (PHCTs) that serve 5,501,784 patients (274 PHCT), or 74% of the population; the remaining PHCTs are managed by other providers. The CHI's Information System for the Development of Research in Primary Care (SIDIAP) contains the coded clinical information recorded in electronic health records EHR by its 274 PHCTs since 2006. A subset of records meeting the highest quality criteria for clinical data (SIDIAP Q) includes 40% of the SIDIAP population (1,833,125 individuals), attended by the 1,365 general practitioners (GPs) whose data recording scores contain information on the majority of the population of Catalonia, and is highly representative for the whole region in terms of geography, age, gender and diseases.[10–12]

Prevalence of individual conditions varies with age and so does multimorbidity and their patterns. In order to obtain a more homogenous sample in terms of multimorbidity, we

focussed on individuals aged 45 to 64 years.[13–16] We identified 408,944 individuals aged 45 to 64 years on 31 December 2010 with two or more diagnoses (Appendix 1).

Coding and selection of diseases

Diseases are coded in SIDIAP using International Classification of Diseases version 10 (ICD-10). For this study, we selected all active diagnoses recorded in EHR as of December 31, 2010, except for R codes (symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified) and Z codes (factors influencing health status and contact with health services).[17] Non-active diagnoses were excluded, based on the presence of an end date in the EHR. These diagnoses cover a broad list of acute diseases for which the system automatically assigns an end date (e.g., 60 days after the initial diagnosis).

To facilitate management of the diagnostic information, the diagnoses were extracted using the 263 blocks (disease categories) in the ICD-10 structure. These are homogeneous categories of very closely related specific diagnoses. For example, Hypertensive diseases include Essential (primary) hypertension, Hypertensive heart disease, Hypertensive renal disease, Hypertensive heart and renal disease and Secondary hypertension. To obtain consistent and clinically interpretable patterns of association, and to avoid spurious relationships that could bias the results, we considered only diagnoses with greater than 1% prevalence in each sex. All patients with multimorbidity (2 or more coexisting diagnoses recorded in the EHR on 31 December 2010) were included.

Variables

The variables considered were: diagnosis (values: 1 for present, 0 for absent), number of diseases (2, 3, 4, and 5 or more) and sex (women, men) were also recorded for each patient.

Statistical analysis

Data access: Data was obtained from SIDIAP after the study was authorized. All the project's authors could access the database. Cleaning methods: The analysis was limited to SIDIAP-Q, as

the sample was representative of the population.[10–12] No missing values were handled as sex and age were recorded for all population. Wrong sex-specific diagnoses codes and diagnoses with inconsistent dates were excluded. An individual with no disease diagnoses record was considered as disease free.

Analyses were stratified by sex. Descriptive statistics were used to summarize overall information. Categorical variables were expressed as frequencies (percentage) and continuous as mean (standard deviation, SD) or median (interquartile range, IQR). Two sample test of proportions and Mann-Whitney test were used to test differences by sex.

We identified disease patterns using two approaches: 1) hierarchical cluster analysis (HCA), and 2) exploratory factor analysis (EFA). Clinical criteria were used to evaluate the consistency and utility of the final HCA and EFA solutions, based on previously described patterns in the literature and a consensus opinion drawn from the clinical experience of the research team (4 family physicians, 1 epidemiologist). Clusters and factors in these analyses were considered as two sets of grouping solutions, which were then assigned to each individual patient. We considered patients to be associated with a given grouping solution if they had ≥1 diagnoses in that solution, allowing for the calculation of the prevalence of each solution in the sample. Patients could be associated with more than one solution in the same set. We also calculated prevalence, restricting the assignment of patients to those with ≥ 2 diagnoses in the same solution.

Hierarchical Cluster Analysis

The HCA approach assigns diagnoses to groups or clusters, so that diagnoses in the same cluster are more similar, based on a given measure, to one another than to diagnoses from different clusters. The Jaccard coefficient was used to measure similarity. This coefficient considers only the diagnoses that any two patients have and ignores the diagnoses that neither of them has.[6] As we do not know a priori the number of clusters to retain from the

data, we used agglomerative hierarchical methods to identify possible clustering solutions: Average linkage, Ward, flexible beta and other methods with less bias, based on nonparametric estimates, such as Single Linkage and Density Linkage. All but Ward and the flexible beta methods successively chained the observations into one cluster. Therefore, the Ward method, which minimizes the variance within clusters and produces clusters of similar sizes, was chosen as the primary method based on dendrograms analysis.[6] Data were randomly split into test and training datasets, equal in size and analysed separately. We ran the Ward method on both samples. The semi-partial R2, Calinski-Harabasz pseudo-F- and pseudo-T2-statistic criteria for different numbers of clusters were examined.[6] Clustering solutions were compared between the test and training datasets, taking into account the number of clusters, Adjusted Rand Index and clinical criteria. After checking algorithm stability, Ward method was run on the full data set, applying the same criteria to different numbers of clusters. Results were compared with flexible beta results, with beta values set at -0.25 and -0.5. The criteria for selecting the number of clusters were the highest adjusted Rand index with a high number of clusters and a high pseudo T2 statistic.[6] To assess internal cluster quality, we applied multiscale bootstrap resampling to obtain an approximately unbiased (AU) probability. This probability ('p-value') is the proportion of bootstrapped samples that contain the cluster; larger p-values indicate more support for the cluster.[18]

Multidimensional scaling (MDS) considering two dimensions was used to discover the underlying structure of distance measures between diseases in the cluster analysis. Essentially, MDS assigns observations to specific locations in a conceptual space such that the distances between points in the space match the given dissimilarities as closely as possible. We carried out classical MDS using the distance matrix obtained in the cluster analysis that considered the Jaccard coefficient as a dissimilarity measure. The conceptual map of the diseases distinguishes between the intra-disease cluster and the inter-disease cluster. Taking into account the final cluster's solution and the obtained groups, conceptual maps of the diseases

were created. For a better interpretation of the conceptual map, prevalence of the disease was represented as the radius of the circle.[19]

Exploratory Factor Analysis

EFA reduces the observed set of diagnoses to a smaller number of latent factors that account for the correlations between them. As the study variables were dichotomous, the correlation matrix between the diagnoses was computed using tetrachoric correlations. The factorability of the matrix was tested using Bartlett Test of Sphericity and Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy. The extraction of the initial solution was carried out using the principal factors method with squared multiple correlations for the prior communality estimates. The optimal number of extracted factors for the final solution was determined with the Scree plot using the "elbow" rule and setting the percentage of variance equal to 100 percent. Factor loadings were analyzed to identify factor patterns. An oblique rotation, Oblimin, was performed to clarify the factor pattern in order to better interpret the nature of the factors, as we assumed that factors were allowed to be associated with each other. As a rule of thumb, factor loadings greater than or equal to 0.30 in absolute value were considered to be significant.[7]

Comparing multimorbidity patterns

We compared every cluster and factor solutions across sex groups agreement and the diagnoses included in it.

We considered grouping solutions (HCA vs EFA) from different sets (women vs men) to have the following degrees of similarity: a) perfect, when the solution included exactly the same diseases as another solution in the other comparison group (sex or statistical approach); b) partial, when the solution included a subset of diseases present in a solution in the other comparison group; and c) none, when each and every disease in the solution was part of a different solution in the other group and none was part of the same solution. These groups

were named using the abbreviation of sex, method and number (For example, MC1: Men Cluster 1)

We further extracted the common subsets of diseases within partially similar solutions, which together with completely similar solutions gave a comprehensive picture of overlapping cluster and factor analysis solutions.

The analyses were performed using SPSS for Windows, version 18 (SPSS Inc., Chicago, IL, USA), SAS 9.2 for Windows (SAS Institute Inc., Cary, NC, USA) and R version 3.1.2 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Of 523,656 patients, 408,994 (78.1%) met the multimorbidity criteria; women had a higher multimorbidity prevalence than men (82.2% vs 73.9%, respectively; p<0.001) (Table 1). Both cluster and factor analyses included 79 diagnoses for women and 73 for men.

Table 1. Number of diseases, clusters, and factors identified in cluster and factorial analysis for patients 45-64 years-old, stratified by sex (N=523,656)

	Total, n (%)	Women, n (%)	Men, n (%)
≥2 Diagnoses*	408,994 (78.1%)	217,823 (82.2%)	191,171 (73.9%)
Number of diagnoses*			
2		26,106 (12.0%)	33,850 (17.7%)
3		28,243 (13.0%)	33,515 (17.5%)
4		28,274 (13.0%)	30,356 (15.9%)
≥5		135,200 (62.1%)	93,450 (48.9%)
Median number of diagnoses (IQR)**		5 (4-8)	4 (3-7)
Number of diagnoses included		79	73
Number of clusters		53	15
Number of clusters with ≥2 diagnoses		12	15
Median of diagnoses per clusters (IQR)***		2 (2-4)	5 (2.5-6)
Number of factors		9	10
Number of factors with ≥2 diagnoses		8	9
Median of diagnoses per factors (IQR)***		5.5 (2.75-7)	4 (4-5)

Abbreviations: IQR, Inter-quartile range.
** Two sample test of proportions; all p-values<0.001*
*** Mann-Whitney test; P<0.001*
****Median of clusters or factors with ≥2 diseases; P<0.001*

Hierarchical Cluster analysis

Using HCA with the Ward method, we obtained 53 clusters, with just 12 clusters grouping at least two diagnoses for women and 15 clusters for men (Table 1). We describe only the four most prevalent clusters (Table 2). For a complete description of the clusters and dendrograms, see Appendices 2-5.

Table 2. Four most prevalent clusters, by sex group (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2,%†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	AU p-value**
WOMEN					
WC1^ 82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	0.79 (0.74-0.85)
		F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	
		M70-M79:Other soft tissue disorders	27.0	32.8	
		N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	
		M20-M25:Other joint disorders	18.6	22.6	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	
		D10-D36:Benign neoplasms	16.2	19.7	
		M15-M19:Arthrosis	15.7	19.1	
WC2 55.8	23.0	E70-E90:Metabolic disorders	37.4	63.4	0.93 (0.86-1.00)
		I10-I15:Hypertensive diseases	25.6	45.8	
		E65-E68:Obesity and other hyperalimentation	19.0	34.0	
		E10-E14:Diabetes mellitus	7.7	13.7	
WC3 47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.7	39.4	0.78 (0.73-0.84)
		E00-E07:Disorders of thyroid gland	14.9	31.4	
		F30-F39:Mood [affective] disorders	14.6	30.8	
		M80-M85:Disorders of bone density and structure	11.3	23.9	
WC4 32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	0.71 (0.66-0.77)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	
		L20-L30:Dermatitis and eczema	9.3	28.8	
		B35-B49:Mycoses	5.7	17.8	
MEN					
MC1^^ 83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	0.69 (0.64-0.75)
		F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	
		I10-I15:Hypertensive diseases	32.5	38.8	

		M50-M54:Other dorsopathies	27.8	33.2	
		E65-E68:Obesity and other hyperalimentation	14.6	17.4	
		E10-E14:Diabetes mellitus	14.2	16.9	
MC2 57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	0.87 (0.84-0.90)
		N40-N51:Diseases of male genital organs	12.1	21.0	
		M20-M25:Other joint disorders	12.1	20.9	
		K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	
		I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	
		K40-K46:Hernia	8.8	15.2	
		D10-D36:Benign neoplasms	8.6	14.9	
		M15-M19:Arthrosis	7.7	13.4	
		K55-K63:Other diseases of intestines	6.4	11.1	
MC3 54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	0.79 (0.74-0.84)
		K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	
		J40-J47:Chronic lower respiratory diseases	9.3	17.2	
		J00-J06:Acute upper respiratory infections	8.9	16.4	
		J30-J39:Other diseases of upper respiratory tract	8.0	14.8	
		L20-L30:Dermatitis and eczema	7.5	13.9	
		G40-G47:Episodic and paroxysmal disorders	7.4	13.7	
		F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	
		F30-F39:Mood [affective] disorders	6.3	11.6	
		B35-B49:Mycoses	4.1	7.6	
MC4 25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	0.87 (0.83-0.91)
		H53-H54:Visual disturbances and blindness	3.9	15.5	
		B00-B09:Viral infections characterized by skin and mucous membrane lesions	3.5	13.9	
		L60-L75:Disorders of skin appendages	3.5	13.9	
		H10-H13:Disorders of conjunctiva	3.0	12.0	
		H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	
		L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	
		L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	
		H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	

*Individuals from the strata ≥1 diagnosis in the cluster/ † Individuals from the strata with ≥2 diagnosis in the cluster/‡Strata: same sex

**Approximately unbiased (AU) probability-value

^ Abbreviation of Sex, method and number (WC1: Women Cluster 1)

^^ Abbreviation of Sex, method and number (MC1: Men Cluster 1)

Twelve clusters with at least two diseases were identified for women, with prevalences ranging from 6.6% to 82.1% (median: 15.5%), (WC3; WC10). The clusters identified in men had prevalences ranging from 3.2% to 83.8% (median 10.1%). The most prevalent cluster included 8 diseases in women (WC1: musculoskeletal, psychiatric, circulatory, gynecological and neoplasms) and 6 in men (MC1: metabolic and circulatory); about half of all patients had at least two diagnoses (52.9% of women and 50.4% of men).

Two clusters were common to men and women, “Spondylopathies” and “Deforming dorsopathies” (WC11, MC13) and “Urolithiasis” and “Other diseases of the urinary systems” (WC9, MC12) (Box 1). The remaining clusters showed partial similarity in men and women, based on six subsets (Box 2), except for three clusters found only in women (WC3, WC7, WC10) and six only in men (MC4, MC5, MC8, MC10, MC14, MC15) (Appendices 4-5).

BOXES

Box 1. Combinations of diseases consistent in both men and women$

Clusters

Complete (whole) clusters

1. M45-M49:Spondylopathies*
   M40-M43:Deforming dorsopathies (WC11;MC13)#
2. N20-N23:Urolithiasis
   N30-N39:Other diseases of urinary system (WC9; MC12)

Subsets within clusters

1. E65-E68:Obesity and other hyperalimentation
   10-I15:Hypertensive diseases
   E10-E14:Diabetes mellitus (WC2; MC1)
2. M15-M19:Arthrosis
   M20-M25:Other joint disorders
   I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
   M70-M79:Other soft tissue disorders
   D10-D36:Benign neoplasms (WC1; MC2)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   K00-K14:Diseases of oral cavity, salivary glands and jaws
   J00-J06:Acute upper respiratory infections (WC4; MC3)
4. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12; MC7)
5. J30-J39:Other diseases of upper respiratory tract
   J40-J47:Chronic lower respiratory diseases (WC6; MC3)
6. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; MC2)
7. G50-G59:Nerve, nerve root and plexus disorders
   M65-M68:Disorders of synovium and tendon (WC8; MC6)

Factors*

Subgroups within factors

1. I10-I15:Hypertensive diseases
   I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I70-I79:Diseases of arteries, arterioles and capillaries (WF3; MF2)
2. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation (WF2;MF1)
3. J00-J06:Acute upper respiratory infections
   J20-J22:Other acute lower respiratory infections
   J09-J18:Influenza and pneumonia
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (WF4; MF6)
4. M15-M19:Arthrosis
   M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies
   M50-M54:Other dorsopathies (WF1;MF4)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   Q38-Q45:Other congenital malformations of the digestive system (WF6;MF8)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses
   H53-H54:Visual disturbances and blindness
   H10-H13:Disorders of conjunctiva
   L80-L99:Other disorders of the skin and subcutaneous tissue (WF5;MF5)
7. H25-H28:Disorders of lens
   H30-H36:Disorders of choroid and retina (WF3; MF7)

** Coincident disease in both sexes*

 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)

$ No two full factors were exactly the same for both sexes.

Underlined blocks of diagnosis represent coincident diseases in pattern.

Box 2. Combinations of diseases consistent across statistical methods (cluster and factor analysis)$

Women

1. I10-I15:Hypertensive diseases*
   E65-E68:Obesity and other hyperalimentation
   E10-E14:Diabetes mellitus (WC2; WF2)#
2. M15-M19:Arthrosis
   M50-M54:Other dorsopathies
   M70-M79:Other soft tissue disorders (WC1; WF1)
3. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (WC4; WF5)
4. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (WC11; WF1)
5. K20-K31:Diseases of oesophagus, stomach and duodenum
   K40-K46:Hernia (WC5; WF6)
6. K70-K77:Diseases of liver
   K80-K87:Disorders of gallbladder, biliary tract and pancreas (WC12 ;WF6)

Men

1. I10-I15:Hypertensive diseases
   E65-E68:Obesity and other hyperalimentation
   E70-E90:Metabolic disorders (MC1; MF1)
2. I20-I25:Ischaemic heart diseases
   I30-I52:Other forms of heart disease
   I60-I69:Cerebrovascular diseases
   I70-I79:Diseases of arteries, arterioles and capillaries
   N17-N19:Renal failure (MC5; MF2)
3. J09-J18:Influenza and pneumonia
   J20-J22:Other acute lower respiratory infections
   B25-B34:Other viral diseases
   A00-A09:Intestinal infectious diseases (MC10; MF6)
4. H10-H13:Disorders of conjunctiva
   H53-H54:Visual disturbances and blindness
   L80-L99:Other disorders of the skin and subcutaneous tissue (MC4; MF5)
5. M45-M49:Spondylopathies
   M40-M43:Deforming dorsopathies (MC13; MF4)
6. L20-L30:Dermatitis and eczema
   B35-B49:Mycoses (MC3; MF5)
7. K70-K77:Diseases of liver
   B15-B19:Viral hepatitis (MC7; MF3)
8. T08-T14:Injuries to unspecified part of trunk, limb or body region
   S90-S99:Injuries to the ankle and foot (MC8; MF9)
9. H25-H28:Disorders of lens
   H40-H42:Glaucoma (MC14; MF7)

** Coincident disease in both methods*
 # Abbreviation of Sex (W: Women; M: Men), method (C: Hierarchical Cluster Analysis; F: Factor Analysis) and number (e.q, WC1: Women Cluster 1)$ All subgroups of factors or clusters, no single cluster exactly the same as a factor.
Underlined blocks of diagnosis represent coincident diseases in pattern.

The top four clusters were reproduced in the graphical representation of coordinates using MDS. The most prevalent disease clusters were more clearly separated in women than they were in men, mostly due to the overlap of MC2 and MC3 (Figure 1).

Exploratory Factor Analysis

Using EFA, we obtained nine factors for women and 10 factors for men. In this analysis, the median number of diagnoses per factor was higher in women (Table 1).Two factors explained more than 50% of total variance (58.5% for women (WF2; WF3) and 50.7% for men (MF1; MF2)). All diseases were assigned to single factors except for three diseases, two in women only (J20-J22: Other acute lower respiratory infections and E10-E14: Diabetes mellitus) and one in both women and men (I10-I15: Hypertensive diseases); all three were assigned to two factors. The first four factors are described in Table 3; full factor solutions are shown in Appendices 6-7.

Table 3. Four most prevalent factors, by sex (N(women)=217,823; N(men)=191,171)

Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses**	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WOMEN						
WF1^ 59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
		M70-M79:Other soft tissue disorders	27.0	45.2		
		M15-M19:Arthrosis	15.7	26.2		
		G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
		M45-M49:Spondylopathies	4.3	7.3		
		M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
		E65-E68:Obesity and other hyperalimentation	19.0	50.2		
		E10-E14:Diabetes mellitus	7.7	20.3		
WF3 32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
		E10-E14:Diabetes mellitus	7.7	23.4		
		I30-I52:Other forms of heart disease	4.4	13.3		
		H25-H28:Disorders of lens	1.7	5.3		
		H30-H36:Disorders of choroid and retina	1.2	3.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
		I20-I25:Ischaemic heart diseases	1.0	3.1		
WF4 27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
		N30-N39:Other diseases of urinary system	5.9	21.3		
		H60-H62:Diseases of external ear	3.6	13.1		
		J20-J22:Other acute lower respiratory infections	3.4	12.2		
		A00-A09:Intestinal infectious diseases	2.7	10.0		

		H65-H75:Diseases of middle ear and mastoid	2.5	9.2		
		J09-J18:Influenza and pneumonia	1.7	6.1		
		B25-B34:Other viral diseases	1.3	4.8		
		M60-M63:Disorders of muscles	1.2	4.4		
MEN						
MF1^^ 61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
		I10-I15:Hypertensive diseases	32.6	52.7		
		E65-E68:Obesity and other hyperalimentation	14.6	23.6		
		M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
		I30-I52:Other forms of heart disease	6.9	17.6		
		I20-I25:Ischaemic heart diseases	5.0	12.6		
		I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
		I60-I69:Cerebrovascular diseases	1.8	4.6		
		N17-N19:Renal failure	1.5	3.7		
MF3 38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
		K70-K77:Diseases of liver	5.2	13.6		
		B15-B19:Viral hepatitis	3.2	8.4		
		F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
		M15-M19:Arthrosis	7.7	22.2		
		M45-M49:Spondylopathies	3.1	8.8		
		M40-M43:Deforming dorsopathies	1.8	5.2		

**Individuals from the strata ≥1 diagnosis in the factor/ † Individuals from the strata with ≥2 diagnosis in the factor/‡Strata: same sex*
***KMO sampling adequacy index was 0.82 for women and 0.75 for men. Bartlett Test of Sphericity was statistically significant. (p<0.001) for both groups*
^ Abbreviation of Sex, method and number (WF1: Women Factor 1)
^^ Abbreviation of Sex, method and number (MC1: Men Factor 1)

Although no factor-based groupings were identical in men and women, almost all showed partial similarity by sex, based on seven subsets (Box 1), except for two groups found only in women (WF7, WF9) and one found only in men (MF9).

Multimorbidity patterns comparison across statistical approaches

The EFA multimorbidity patterns were more easily interpreted than the HCA groups, either because they made more sense from a clinical perspective or because of greater homogeneity

in the diagnoses: about half of the factors containing at least 3 diagnoses corresponded to a maximum of 2 ICD-10 blocks, compared to about one fifth of the HCA clusters with at least 3 diseases (1/5 for women and 2/11 for men). No grouping solution for women or for men contained exactly the same set of diseases as a cluster and as a factor. Six clusters (WC3, WC6, WC7, WC8, WC9, WC10) and two factors (WF7, WF9) were only observed in women. However, six subsets of diseases were part of the same grouping in both a cluster and a factor (Box 2); all included two or three diagnoses, usually from the same ICD chapter. Five clusters and one factor were observed only in men (MC6, MC9, MC11, MC12, MC15 and MF6). Nine subsets of diseases were observed as part of the same grouping in both a cluster and a factor (Box 2). They included a range of diseases (2-5) and most frequently included diseases from different ICD chapters.

Three paired diseases were observed consistently in both men and women using both methods of analysis: 1) Hypertensive diseases and obesity/other hyperalimentation; 2) spondylopathies and deforming dorsopathies; and 3) dermatitis/eczema and mycoses.

DISCUSSION

In this study have observed differences in the groupings identified with the two most frequently used methods (HCA and EFA). No grouping solution contained exactly the same set of diseases as a cluster and as a factor, although some overlap was observed in both groups. Internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

The multimorbidity patterns obtained by HCA were identified graphically with MDS, allowing us to observe a given hierarchical structure. Nevertheless, internal quality validation with AU p-values showed strong evidence of the multimorbidity patterns in the data.

EFA, based on tetrachoric correlations where the dichotomous diseases were assumed to come from an underlying mechanism with a continuous variable, produced a wide range of multimorbidity patterns with several levels of correlations. Most of them seem to be highly consistent from a clinical perspective. The multimorbidity patterns obtained with EFA show a main factor (a disease) that has a correlation with another disease, either coexisting or that may occur during the patient's clinical course. [20] Thus, EFA could be more useful for analyzing comorbidity and for describing the correlation between diseases that have a pathophysiological relationship. This approach also may help to answer the question of which condition should be considered the main disease and which the comorbidity.

The HCA results would be useful in generating new hypotheses for intercluster and intracluster associations between diseases that could be applied to the analysis of multimorbidity, defined as the random coexistence of diseases or clusters that indicates significant associations between diseases without a causal explanation. In future studies, other non-hierarchical cluster analysis techniques will improve measurement of the observed distances and multiple interrelationships between different diseases in a given individual.[21] On the other hand, EFA could be more useful for analyzing multimorbidity patterns in the absence of causal comorbidity and for describing visual representation of diseases correlation with a pathophysiological relationship between them.

We obtained two perfect clusters that were common to both men and women: "spondylopathies and deforming dorsopathies" and "urolithiasis and other diseases of the urinary system". In the first cluster, spondylosis is a degenerative disorder that may cause loss of normal spinal structure and function and lead to scoliosis. Nevertheless, many individuals with untreated scoliosis will develop spondylosis; this may be one reason why these diseases were associated.[22] The second cluster can be explained by the complications produced by urolithiasis (such as urinary tract infection, persistent proteinuria, stress incontinence or other unspecified urinary incontinence) and those that have a pathophysiological explanation. [23]

EFA showed that the most frequent pattern in women was infectious diseases. This previously unreported pattern suggests that the multimorbidity patterns obtained in other studies are affected by the type of diseases included in each study.

Although the patterns obtained with both methods did not match exactly, finding matching pairs of diseases by both methods reinforces the idea that patterns of multimorbidity have a dominant disease that associates in some way with other diseases.

In general, it is difficult to compare our results with other studies because of variations in methods, data sources and structures, and populations and diseases studied. Six studies have been performed with HCA[8,21,24–27] and three using EFA.[28–30] Until now, very few analyses of multimorbidity patterns have used multiple methods to compare the same population.[21] The latter study included people aged 50 years and older, considering 11 diseases and using 2 different cluster methods, hierarchical (average linkage) and nonhierarchical (k-medoids), and one method for EFA (principal component analysis). The observed differences between this study and our results can be explained by differences in the underlying statistical formulae and diseases considered in both studies.

The major strength of this study is the dual analysis (HCA and EFA) of a large, high-quality database of primary care records that have been shown to be representative of a much larger population, stratified by sex. Admittedly, this analysis of almost all potential diagnoses may have added a complexity that will hinder interpretation of findings and comparison with other studies, particularly because the boundaries between chronic and acute disease are not always clear.[31,32] Whatever consistency (or discrepancy) we observed was validated by the findings of two different approaches, which helps to identify the most appropriate use of each method in analyzing multimorbidity. We emphasize that the inclusion of chronic and acute diseases is a strength and not a weakness. Because, as we have shown, there are many chronic and acute diseases that coexist at a set time and this has implications for health care.

Internal validation with bootstrap methods, AU p-values, and MDS techniques provided more robust evidence for the cluster analysis. The KMO values obtained show the adequacy of fit of the factor analysis. These values were similar or higher than previous studies.[28,29]

A limitation of this study is our use of agglomerative hierarchical clustering, which forces every unit (i.e., diagnosis) into a single cluster. HCA is exploratory in nature, and different clustering algorithms may produce different results.[33] The final clustering solution presented here was obtained through a systematic and rigorous process: comparing the results from a randomly split dataset, testing different clustering algorithms, and using different objective numeric criteria to decide the number of clusters, internal validation, and graphical representation. In addition, a panel of experts applied subjective clinical criteria to assess the interpretability of the groupings in everyday practice. In addition, EFA is problematic for binary data, which can be grouped because of having similar distributions rather than any common underlying feature. On the other hand, in factor analysis the measure of association incorporates information on both positive and negative concordances [9]. In contrast, the analysis of clusters allows us to show that the occurrence of one or more health conditions can be conditioned by their co-occurrence, without considering negative concordances. [8] Due to the absence of a standard methodology to compare method solutions we have used ad hoc methodology. Finally, another limitation is our use of ICD-10 3-character codes as the unit of analysis, rather than the more specific individual diagnosis, but its use is justified to avoid spurious relationships that more than 10,000 individual codes of the ICD-10 could produce.

This is a cross-sectional study, based on EHR of The Catalan Health Institute and SIDIAP-Q is highly representative for the whole region in terms of both geography, age, gender and diseases, that avoid selection bias.

Multimorbidity can present a problem for health services delivery, affecting patients, health professionals and managers who are attempting to improve service delivery. Our study offers two methodological approaches to understanding the relationships between specific diseases,

which is an essential step in improving our approach to this problem. Although we demonstrated that different analytical methods can yield different results, we also showed that some associations were consistent in both analyses. This study illustrates the need to pay careful attention to the methods used to support policies and decision-making. Clinical guidelines tend to focus on a single disease rather than on multimorbidity, which includes not only diseases but also drug interactions and polypharmacy. The present study confirmed that multimorbidity patterns are a reality in the adult population, and do not apply only to chronic diseases. New guidelines are needed that incorporate multimorbidity into clinical recommendations.

This study was one of the first to compare the two most commonly used methodologies, HCA and EFA, in a large database that includes a large number of diseases. The findings reveal another limitation to be taken into account in comparing multimorbidity patterns between studies: in addition to the spectrum, number and type of diseases included, these patterns vary depending on the method of analysis used. Nevertheless, it would be necessary to carry out a simulation study to determine how the choice of method may affect the patterns, as it allows us to test the obtained patterns in all kinds of situations.

The results suggest that HCA can be useful to detect multimorbidity patterns and identify different associations between diseases, as the method allows for the possibility that one or more health problems can occur conditionally. On the other hand, EFA seems more applicable to clinical practice because places less restrictions in the diseases grouping, so may be better for generating hypotheses and is more sensitive in identifying clinical associations. Our results suggest that these aspects be considered in planning of future studies, including selection of diseases and methods of analysis.

Finally, our analysis of multimorbidity patterns only considered associations between diseases. Further studies are needed to analyze the patterns that develop longitudinally as individual patients acquire subsequent comorbidities.

Conclusions

The multimorbidity patterns obtained with EFA show a main factor (i.e., a disease) that has some correlation with the additional diseases in the pattern, suggesting a comorbidity relationship. Meanwhile, the HCA would be useful for in-depth study of multimorbidity pattern. We introduced new evidence about the known limitations of attempts to compare multimorbidity or comorbidity patterns between RWD studies, as our results add information about aspects that must be considered in standardization of future studies: spectrum of diseases, data utilization, and methods of analysis.

Footnotes

Contributors: All authors contributed to the design of the study, revised the article, and approved the final version. CV and QFB obtained the funding. ARL and CV drafted the article. ARL, CV, QFB, TRB, MPV, EPR and JMV contributed to the analysis and interpretation of data. ARL and CV wrote the first draft, and all authors (ARL, CV, QFB, TRB, MPV, EPR and JMV) contributed ideas, interpreted the findings and reviewed rough drafts of the manuscript. All authors read and approved the final manuscript.

Funding: The project has been funded by the *Instituto de Salud Carlos III* of the Ministry of Economy and Competitiveness (Spain) through the Network for Prevention and Health Promotion in Primary Health Care (redIAPP, RD12/0005/0001; RD16/0007/001), and by European Union ERDF funds. Also, by a grant for research projects on health from ISCiii (PI12/00427) and co-financed with European Union ERDF funds). Jose M. Valderas was supported by the National Institute for Health Research Clinician Scientist Award NIHR/CS/010/024. The funders had no role in the study design, collection, analysis and interpretation of data, writing of the manuscript or decision to submit for publication. 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.

Ethical considerations: The study protocol was approved by the Committee on the Ethics of Clinical Research, IDIAP Jordi Gol (Protocol No: P12/28). All data were anonymized and EHR confidentiality was respected in accordance with national and international law.

Acknowledgment: We thank the Catalan Health Institute and especially the SIDIAP Unit, which provided the database for the study. The authors also appreciate the English language review by Elaine Lilly, PhD, and are grateful to Carmen Ibáñez for administrative work.

Data Sharing: 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 organization is subject to periodic audits to ensure the validity and quality of the data.

References

1. Valderas J, Starfield B, Sibbald B, et al. Defining comorbidity: Implications for understanding health and health services. *Ann Fam Med* 2009;**7**: 357–363.

2. Violan C, Foguet-Boreu Q, Flores-Mateo G, et al. Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PLoS One* 2014;**9**: e102149.

3. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, et al. Multimorbidity patterns: A systematic review. *J Clin Epidemiol* 2014;**67**: 254–266.

4. Fortin M, Stewart M, Poitras ME, et al. A systematic review of prevalence studies on multimorbidity: Toward a more uniform methodology. *Ann Fam Med* 2012;**10**: 142–151.

5. Corrao, G., Mancia, G., Generating evidence from computerized health care utilization databases. *Hypertension* 2015; **65**(3):490-8.

6.. Everitt B, Landau S, Leese M, et al. Cluster Analysis. 5th ed. Hoboken: Wiley; 2011.

7. Thompson B. Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. Washington, DC: American Psychological Association; 2004.

8. Cornell JE, Pugh JA, Williams J, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters : clustering binary data from a large administrative medical database. *Appl Multivar Res* 2007;**12**: 163–182.

9. Nunnally J, Berstein I. Psychometric theory. 3rd ed. New York: McGraw-Hill, Inc; 1994.

10.. Del Mar Garcia-Gil M, Hermosilla E, Prieto-Alhambra D, 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**: 135–145.

11. Prieto-Alhambra D, Judge A, Javaid MK, et al. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. *Ann Rheum Dis* 2014;**73**: 1659–64.

12. Ramos R, Balló E, Marrugat J, et al. Validity for use in research on vascular diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA study. *Rev Esp Cardiol* (Engl Ed) 2012;**65**: 29–37.

13. Foguet-Boreu Q, Violán C, Rodriguez-Blanco T, et al. Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis. *PLoS One* 2015;**10**: e0141155.

14. Foguet-Boreu Q, Violan C, Roso-Llorach A, et al. Impact of multimorbidity: acute morbidity, area of residency and use of health services across the life span in a region of south Europe. *BMC Fam Pract* 2014;**15**: 55.

15. Violán C, Foguet-Boreu Q, Roso-Llorach A, 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**: 1–13.

16. Violán C, Foguet-Boreu Q, Roso-Llorach A, et al. Patrones de multimorbilidad en adultos jóvenes en Cataluña: un análisis de clústeres. *Aten Primaria* 2016; **48**:479-92.

17. World Health Organization. ICD-10 International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2007.

18. Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling 32: *Ann Stat* 2004;**32**: 2616–2641.

19. Borg I. Modern Multidimensional Scaling: theory and applications (2nd ed.). 2 nd. Springer-Verlag, editor. New York; 2005.

20. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis* 1970; **23**(7):455-68.

21. Islam M, Valderas J, Yen L, et al. Multimorbidity and comorbidity of chronic diseases among the senior Australians: prevalence and patterns. *PLoS One* 2014;**9**: e83783.

22. Pappou IP, Girardi FP, Sandhu HS, et al. Discordantly high spinal bone mineral density values in patients with adult lumbar scoliosis. *Spine* (Phila Pa 1976). 2006;**31**: 1614–1620.

23. Bartoletti R, Cai T, Mondaini N, et al. Epidemiology and risk factors in urolithiasis. *Urol Int* 2007;**79** Suppl 1: 3–7.

24. Formiga F, Ferrer A, Sanz H, et al. Patterns of comorbidity and multimorbidity in the oldest old: The Octabaix study. *Eur J Intern Med* 2013;**24**: 40–44.

25. John R, Kerby DS and Hennessy CH. Patterns and impact of comorbidity and multimorbidity among community-resident American Indian elders. *Gerontologist* 2003;**43**: 649–660.

26. Marengoni A, Rizzuto D, Wang HX, et al. Patterns of chronic multimorbidity in the elderly population. *J Am Geriatr Soc* 2009;**57**: 225–230.

27. Newcomer SR, Steiner JF and Bayliss EA. Identifying subgroups of complex patients with cluster analysis. *Am J Manag Care* 2011;**17**: e324–32.

28. Schäfer I, von Leitner EC, Schön G, et al. Multimorbidity patterns in the elderly: a new approach of disease clustering identifies complex interrelations between chronic conditions. *PLoS One* 2010;**5**: e15941.

29. Andre L, Prados-torres A, Poblador-plou B, et al. Multimorbidity Patterns in Primary Care : Interactions among Chronic Diseases Using Factor Analysis. *PLoS One* 2012;**7**(2): e32190.

30. Poblador-Plou B, van den Akker M, Vos R, et al. Similar multimorbidity patterns in primary care patients from two European regions: results of a factor analysis. *PLoS One* 2014;**9**: e100375.

31. O'Halloran J, Miller GC and Britt H. Defining chronic conditions for primary care with ICPC-2. *Fam Pract* 2004;**21**: 381–386.

32. Lamberts H and Wood M. International Classification of Primary Care. Oxford. Press OU, editor. Oxford; 2011.

33. Aldenderfer MS and Blashfield RK. Cluster Analysis: Quantitative Applications in the Social Sciences. 1984.

Figure 1

Figure 1a. Top four clusters in women (n= 217,823) aged 45-64 years, analyzed with multidimensional scaling.

Figure 1b. Top four clusters in men (n= 191,171) aged 45-64 years, analyzed with multidimensional scaling.

Appendix

Appendix 1. Study Flow Chart.

Appendix 2. Cluster dendrogram in women aged 45-64 years.

Appendix 3. Cluster dendrogram in men aged 45-64 years, with approximately unbiased (AU) probability value.

Appendix 4. Prevalence and composition of diagnostic clusters in women aged 45-64 years (n= 217,823).

Appendix 5. Prevalence and composition of diagnostic clusters in men aged 45-64 years (n= 191,171).

Appendix 6. Factors in women aged 45-64 years (n= 217,823).

Appendix 7. Factors in men aged 45-64 years (n= 191,171).

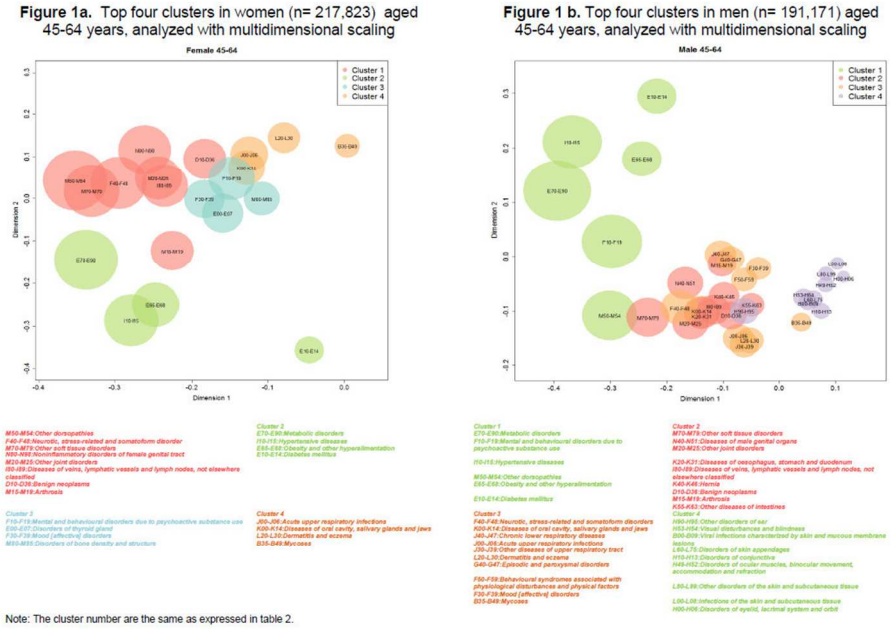


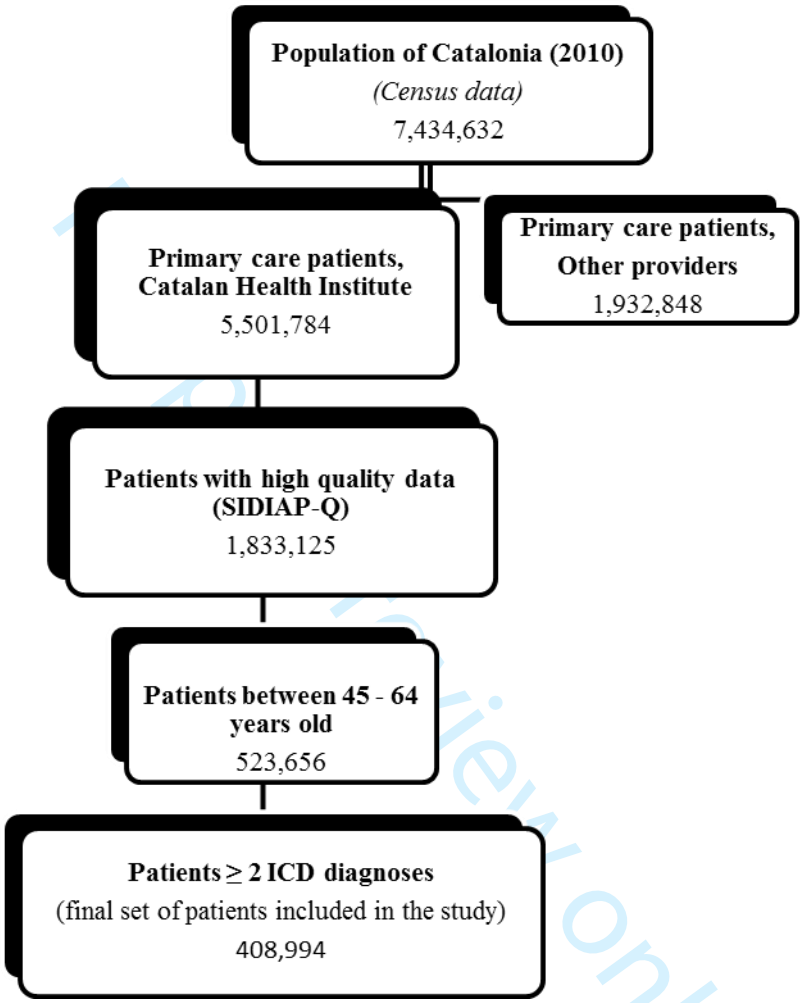
Figure 1

Figure 1a. Top four clusters in women (n= 217,823) aged 45-64 years, analyzed with multidimensional scaling.

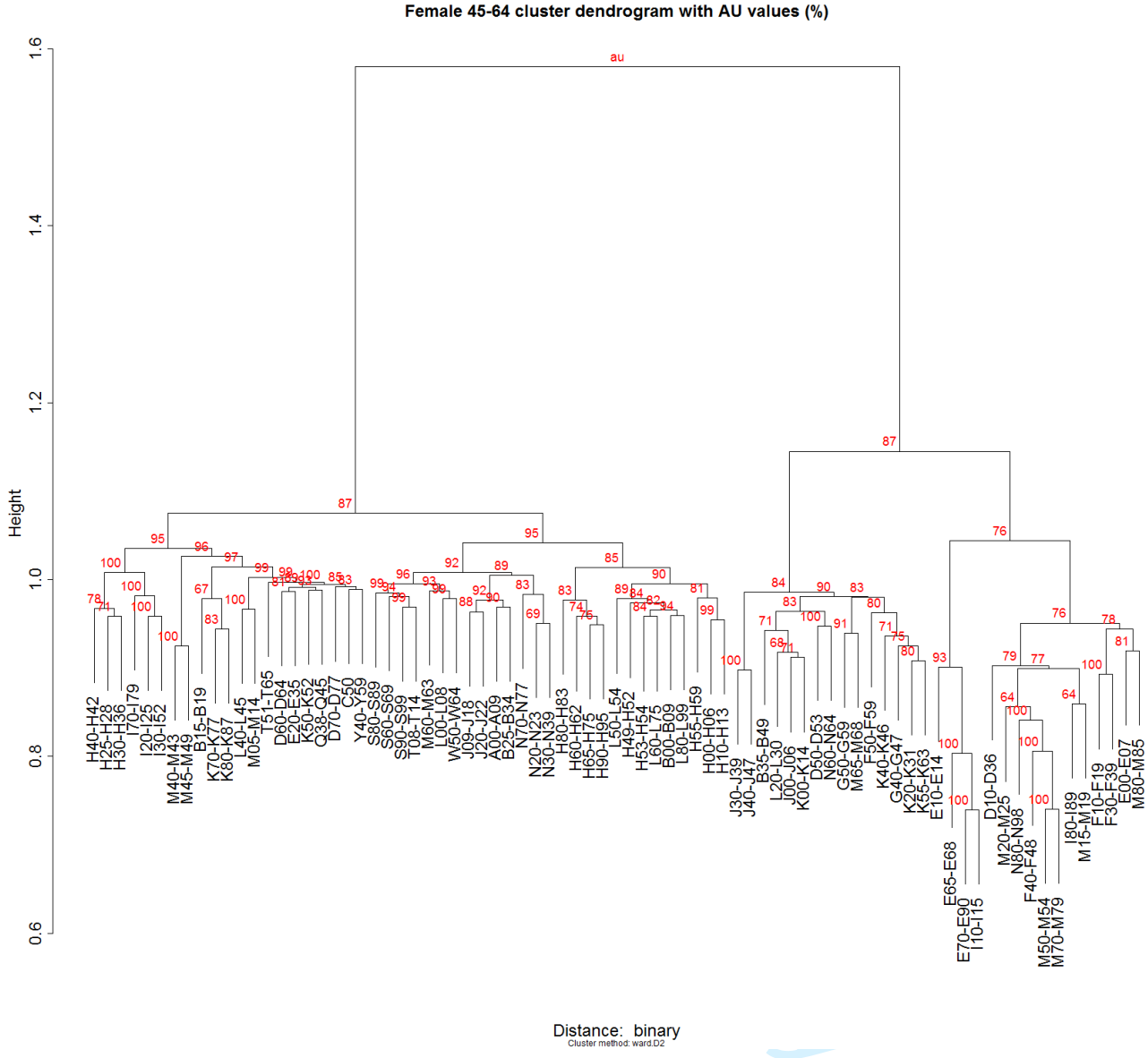
Figure 1b. Top four clusters in men (n= 191,171) aged 45-64 years, analyzed with multidimensional scaling.

97x64mm (300 x 300 DPI)

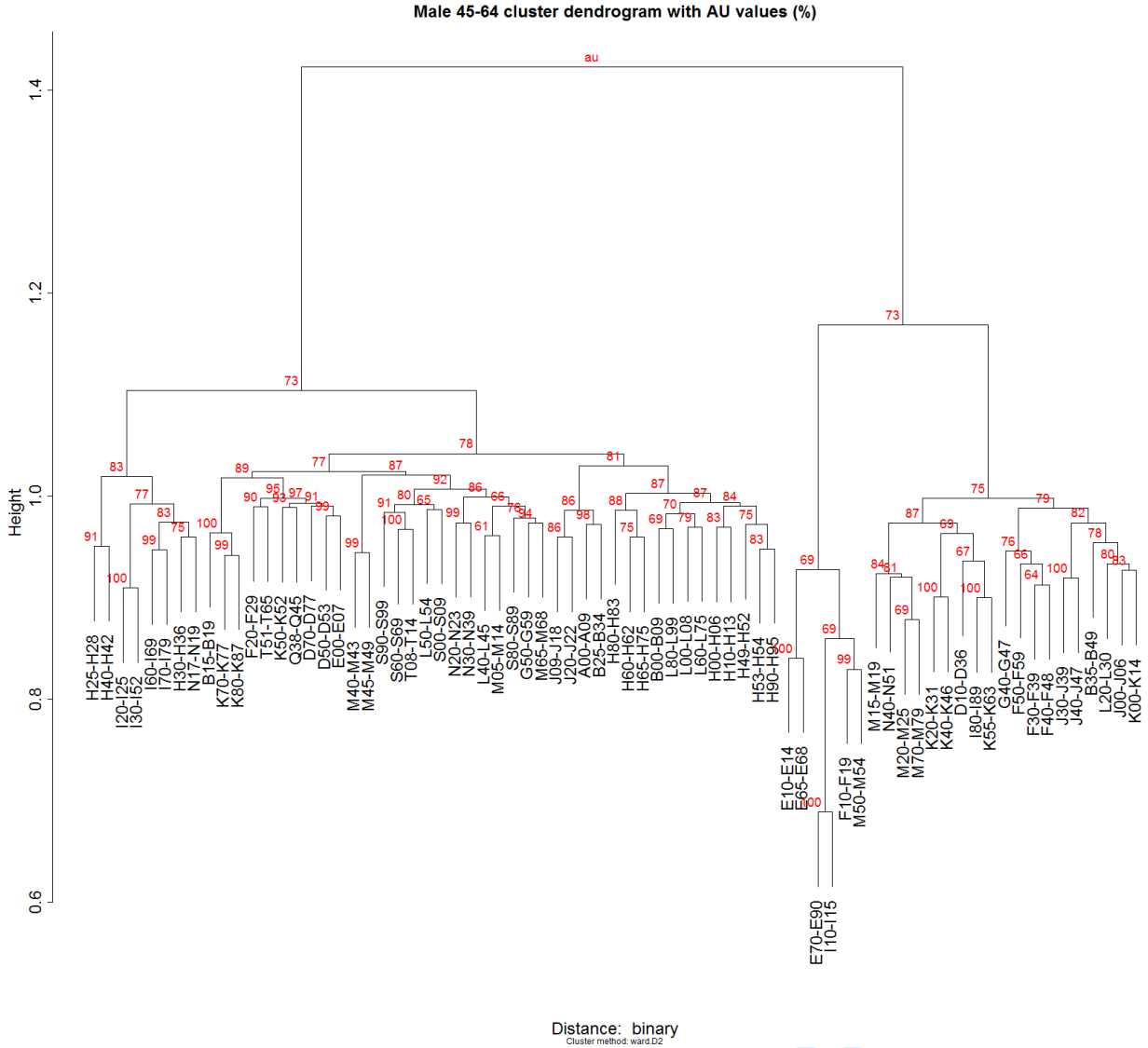
Appendix 1. Study Flow Chart.



Appendix 2. Cluster dendrogram in women aged 45-64 years.



Appendix 3. Cluster dendrogram in men aged 45-64 years, with approximately unbiased (AU) probability value.



Appendix 4. Prevalence and composition of diagnostic clusters in women aged 45-64 years (n= 217,823).

Cluster order number, n	Prevalence 1 (%)*	Prevalence 2 (%)†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value**
WC1^ 178,849	82.1	52.9	M50-M54:Other dorsopathies	35.8	43.5	M54 Dorsalgia	0.79 (0.74-0.85)
			F40-F48:Neurotic, stress-related and somatoform disorders	27.3	33.2	F41 Other anxiety disorders	
			M70-M79:Other soft tissue disorders	27.0	32.8	M79 Other soft tissue disorders, not elsewhere classified	
			N80-N98:Noninflammatory disorders of female genital tract	24.2	29.5	N95 Menopausal and other perimenopausal disorders	
			M20-M25:Other joint disorders	18.6	22.6	M25 Other joint disorders, not elsewhere classified	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	18.3	22.2	I83 Varicose veins of lower extremities	
			D10-D36:Benign neoplasms	16.2	19.7	D25 Leiomyoma of uterus	
			M15-M19:Arthrosis	15.7	19.1	M17 Gonarthrosis [arthrosis of knee]	
WC2 121,564	55.8	23.0	E70-E90:Metabolic disorders	35.4	63.4	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.93 (0.86-1.00)
			I10-I15:Hypertensive diseases	25.6	45.8	I10 Essential (primary) hypertension	
			E65-E68:Obesity and other hyperalimentation	19.0	34.0	E66 Obesity	
			E10-E14:Diabetes mellitus	7.7	13.7	E11 Non-insulin-	

						dependent diabetes mellitus	
WC3 103,147	47.4	10.8	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	39.4	F17 Mental and behavioural disorders due to use of tobacco	0.78 (0.73-0.84)
			E00-E07:Disorders of thyroid gland	14.9	31.4	E03 Other hypothyroidism	
			F30-F39:Mood [affective] disorders	14.6	30.8	F32 Depressive episode	
			M80-M85:Disorders of bone density and structure	11.3	23.9	M81 Osteoporosis without pathological fracture	
WC4 70,294	32.3	6.4	J00-J06:Acute upper respiratory infections	12.6	39.1	J00 Acute nasopharyngitis [common cold]	0.71 (0.66-0.77)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.1	37.3	K02 Dental caries	
			L20-L30:Dermatitis and eczema	9.3	28.8	L29 Pruritus	
			B35-B49:Mycoses	5.7	17.8	B35 Dermatophytosis	
WC5 63,792	29.3	6.0	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	38.9	K30 Dyspepsia	0.71 (0.66-0.76)
			G40-G47:Episodic and paroxysmal disorders	10.5	36.0	G43 Migraine	
			K55-K63:Other diseases of intestines	8.6	29.2	K59 Other functional intestinal disorders	
			K40-K46:Hernia	5.7	19.4	K44 Diaphragmatic hernia	
WC6 34,707	15.9	1.6	J30-J39:Other diseases of upper respiratory tract	9.4	59.0	J30 Vasomotor and allergic rhinitis	1.00
			J40-J47:Chronic lower respiratory diseases	8.2	51.2	J45 Asthma	
WC7 32,674	15.0	0.8	D50-D53:Nutritional anaemias	8.3	55.2	D50 Iron deficiency anaemia	1.00
			N60-N64:Disorders of breast	7.5	50.1	N60 Benign mammary	

						dysplasia	
WC8 27,150	12.5	0.8	G50-G59:Nerve, nerve root and plexus disorders	8.5	68.6	G56 Mononeuropathies of upper limb	0.91 (0.88-0.94)
			M65-M68:Disorders of synovium and tendon	4.7	37.5	M65 Synovitis and tenosynovitis	
WC9 19,199	8.8	0.4	N30-N39:Other diseases of urinary system	5.9	66.7	N39 Other disorders of urinary system	0.69 (0.62-0.75)
			N20-N23:Urolithiasis	3.4	38.3	N20 Calculus of kidney and ureter	
WC10 18,439	8.5	0.4	H90-H95:Other disorders of ear	6.3	75.0	H91 Other hearing loss	0.75 (0.70-0.80)
			H65-H75:Diseases of middle ear and mastoid	2.5	30.1	H65 Nonsuppurative otitis media	
WC11 16,535	7.6	0.6	M45-M49:Spondylopathies	4.3	57.1	M47 Spondylosis	1.00
			M40-M43:Deforming dorsopathies	3.8	50.4	M41 Scoliosis	
WC12 14, 335	6.6	0.4	K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	60.9	K80 Cholelithiasis	0.83*
			K70-K77:Diseases of liver	2.9	44.7	K76 Other diseases of liver	
WC13 12,320	5.7	0.0	F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	5.7	100.0	F51 Nonorganic sleep disorders	
WC14 10,357	4.8	0.0	L60-L75:Disorders of skin appendages	4.8	100.0	L72 Follicular cysts of skin and subcutaneous tissue	
WC15 9,555	4.4	0.0	H53-H54:Visual disturbances and blindness	4.4	100.0	H54 Blindness and low vision	
WC16 9,521	4.4	0.0	I30-I52:Other forms of heart disease	4.4	100.0	I34 Nonrheumatic mitral valve disorders	
WC17 9,442	4.3	0.0	B00-B09:Viral infections characterized by skin and mucous membrane lesions	4.3	100.0	B07 Viral warts	
WC18 8,541	3.9	0.0	T08-T14:Injuries to unspecified part of trunk, limb or body region	3.9	100.0	T14 Injury of unspecified body region	

WC19 8,268	3.8	0.0	H10-H13:Disorders of conjunctiva	3.8	100.0	H10 Conjunctivitis	
WC20 7,866	3.6	0.0	H60-H62:Diseases of external ear	3.6	100.0	H61 Other disorders of external ear	
WC21 7,714	3.5	0.0	H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	3.5	100.0	H52 Disorders of refraction and accommodation	
WC22 7,512	3.4	0.0	L50-L54:Urticaria and erythema	3.4	100.0	L50 Urticaria	
WC23 7,298	3.4	0.0	J20-J22:Other acute lower respiratory infections	3.4	100.0	J20 Acute bronchitis	
WC24 7,251	3.3	0.0	L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	100.0	L82 Seborrhoeic keratosis	
WC25 6,178	2.8	0.0	H00-H06:Disorders of eyelid, lacrimal system and orbit	2.8	100.0	H04 Disorders of lachrymal system	
WC26 5,993	2.8	0.0	H40-H42:Glaucoma	2.8	100.0	H40 Glaucoma	
WC27 5,986	2.7	0.0	A00-A09:Intestinal infectious diseases	2.7	100.0	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
WC28 5,890	2.7	0.0	M05-M14:Inflammatory polyarthropathies	2.7	100.0	M13 Other arthritis	
WC29 5,548	2.5	0.0	S80-S89:Injuries to the knee and lower leg	2.5	100.0	S83 Dislocation, sprain and strain of joints and ligaments of knee	
WC30 5,416	2.5	0.0	C50-C50:Malignant neoplasm of breast	2.5	100.0	C50 Malignant neoplasm of breast	
WC31 5,372	2.5	0.0	L40-L45:Papulosquamous disorders	2.5	100.0	L40 Psoriasis	
WC32 5,336	2.4	0.0	H80-H83:Diseases of inner ear	2.4	100.0	H81 Disorders of vestibular function	
WC33	2.0	0.0	S90-S99:Injuries to the ankle and	2.0	100.0	S93 Dislocation, sprain	

4,374			foot			and strain of joints and ligaments at ankle and foot level	
WC34 3,882	1.8	0.0	B15-B19:Viral hepatitis	1.8	100.0	B18 Chronic viral hepatitis	
WC35 3,758	1.7	0.0	H25-H28:Disorders of lens	1.7	100.0	H25 Senile cataract	
WC36 3,744	1.7	0.0	K50-K52:Noninfective enteritis and colitis	1.7	100.0	K52 Other non-infective gastro-enteritis and colitis	
WC37 3,689	1.7	0.0	N70-N77:Inflammatory diseases of female pelvic organs	1.7	100.0	N76 Other inflammation of vagina and vulva	
WC38 3,646	1.7	0.0	J09-J18:Influenza and pneumonia	1.7	100.0	J11 Influenza, virus not identified	
WC39 3,399	1.6	0.0	L00-L08:Infections of the skin and subcutaneous tissue	1.6	100.0	L03 Cellulitis	
WC40 3,284	1.5	0.0	Q38-Q45:Other congenital malformations of the digestive system	1.5	100.0	Q40 Other congenital malformations of upper alimentary tract	
WC41 3,180	1.5	0.0	D60-D64:Aplastic and other anaemias	1.5	100.0	D64 Other anaemias	
WC42 3,055	1.4	0.0	W50-W64:Exposure to animate mechanical forces	1.4	100.0	W57 Bitten or stung by non-venomous insect and other non-venomous arthropods	
WC43 2,926	1.3	0.0	Y40-Y59:Drugs, medicaments and biological substances causing adverse effects in therapeutic use	1.3	100.0	Y52 Agents primarily affecting the cardiovascular system	
WC44 2,884	1.3	0.0	B25-B34:Other viral diseases	1.3	100.0	B34 Viral infection of unspecified site	
WC45 2,788	1.3	0.0	E20-E35:Disorders of other endocrine glands	1.3	100.0	E28 Ovarian dysfunction	
WC46 2,765	1.3	0.0	S60-S69:Injuries to the wrist and hand	1.3	100.0	S61 Open wound of wrist and hand	

WC47 2,645	1.2	0.0	H55-H59:Other disorders of eye and adnexa	1.2	100.0	H57 Other disorders of eye and adnexa	
WC48 2,612	1.2	0.0	M60-M63:Disorders of muscles	1.2	100.0	M62 Other disorders of muscle	
WC49 2,600	1.2	0.0	H30-H36:Disorders of choroid and retina	1.2	100.0	H35 Other retinal disorders	
WC50 2,584	1.2	0.0	D70-D77:Other diseases of blood and blood-forming organs	1.2	100.0	D72 Other disorders of white blood cells	
WC51 2,508	1.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	1.2	100.0	T65 Toxic effect of other and unspecified substances	
WC52 2,309	1.1	0.0	I70-I79:Diseases of arteries, arterioles and capillaries	1.1	100.0	I73 Other peripheral vascular diseases	
WC53 2,241	1.03	0.0	I20-I25:Ischaemic heart diseases	1.0	100.0	I25 Chronic ischaemic heart disease	

* Standard error to large
**Approximately unbiased (AU) probability-value
^ Abbreviation of sex, method and number (WC1: Women Cluster 1)

Appendix 5. Prevalence and composition of diagnostic clusters in men aged 45-64 years (n= 191,171).

Cluster order number, n	Prevalence 1 .%*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group‡, %	Prevalence in cluster, %	Most prevalent diagnoses of the block	AU p-value*
MC1^ 160,216	83.8	50.4	E70-E90:Metabolic disorders	42.2	50.3	E78 Disorders of lipoprotein metabolism and other lipidaemias	0.69 (0.64-0.75)
			F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	40.1	F17 Mental and behavioural disorders due to use of tobacco	
			I10-I15:Hypertensive diseases	32.5	38.8	I10 Essential (primary) hypertension	
			M50-M54:Other dorsopathies	27.8	33.2	M54 Dorsalgia	
			E65-E68:Obesity and other hyperalimentation	14.6	17.4	E66 Obesity	
			E10-E14:Diabetes mellitus	14.2	16.9	E11 Non-insulin-dependent diabetes mellitus	
MC2 110,200	57.6	24.2	M70-M79:Other soft tissue disorders	16.9	29.3	M75 Shoulder lesions	0.87 (0.84-0.90)
			N40-N51:Diseases of male genital organs	12.1	21.0	N40 Hyperplasia of prostate	
			M20-M25:Other joint disorders	12.0	20.9	M25 Other joint disorders, not elsewhere classified	
			K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	20.0	K29 Gastritis and duodenitis	
			I80-I89:Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	10.0	17.4	I84 Haemorrhoids	
			K40-K46:Hernia	8.8	15.2	K40 Inguinal hernia	

			D10-D36:Benign neoplasms	8.6	14.9	D17 Benign lipomatous neoplasm	
			M15-M19:Arthrosis	7.7	13.4	M17 Gonarthrosis [arthrosis of knee]	
			K55-K63:Other diseases of intestines	6.4	11.1	K62 Other diseases of anus and rectum	
MC3 103,365	54.1	20.7	F40-F48:Neurotic, stress-related and somatoform disorders	13.5	24.9	F41 Nutritional marasmus	0.79 (0.74-0.84)
			K00-K14:Diseases of oral cavity, salivary glands and jaws	12.0	22.3	K02 Dental caries	
			J40-J47:Chronic lower respiratory diseases	9.3	17.2	J44 Other chronic obstructive pulmonary disease	
			J00-J06:Acute upper respiratory infections	8.9	16.4	J00 Acute nasopharyngitis [common cold]	
			J30-J39:Other diseases of upper respiratory tract	8.0	14.8	J30 Vasomotor and allergic rhinitis	
			L20-L30:Dermatitis and eczema	7.5	13.9	L21 Seborrhoeic dermatitis	
			G40-G47:Episodic and paroxysmal disorders	7.4	13.7	G47 Sleep disorders	
			F50-F59:Behavioural syndromes associated with physiological disturbances and physical factors	6.6	12.2	F52 Sexual dysfunction, not caused by organic disorder or disease	
			F30-F39:Mood [affective] disorders	6.3	11.6	F32 Depressive episode	
			B35-B49:Mycoses	4.1	7.6	B35 Dermatophytosis	
MC4 48,140	25.2	4.7	H90-H95:Other disorders of ear	7.7	30.6	H91 Disorders of sclera and cornea in diseases classified elsewhere	0.87 (0.83-0.91)
			H53-H54:Visual disturbances and blindness	3.9	15.5	H54 Blindness and low vision	
			B00-B09:Viral infections characterized by skin and	3.5	13.9	B07 Viral warts	

			mucous membrane lesions				
			L60-L75:Disorders of skin appendages	3.5	13.9	L72 Follicular cysts of skin and subcutaneous tissue	
			H10-H13:Disorders of conjunctiva	3.0	12.0	H10 Conjunctivitis	
			H49-H52:Disorders of ocular muscles, binocular movement, accommodation and refraction	2.8	11.2	H52 Disorders of refraction and accommodation	
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	10.0	L82 Seborrhoeic keratosis	
			L00-L08:Infections of the skin and subcutaneous tissue	2.1	8.3	L02 Cutaneous abscess, furuncle and carbuncle	
			H00-H06:Disorders of eyelid, lacrimal system and orbit	1.6	6.5	H00 Hordeolum and chalazion	
MC5 30,315	15.9	2.9	I30-I52:Other forms of heart disease	6.9	43.7	I45 Other conduction disorders	0.77 (0.73-0.82)
			I20-I25:Ischaemic heart diseases	5.0	31.0	I25 Chronic ischaemic heart disease	
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	15.2	I73 Other peripheral vascular diseases	
			I60-I69:Cerebrovascular diseases	1.8	11.4	I64 Stroke, not specified as haemorrhage or infarction	
			H30-H36:Disorders of choroid and retina	1.7	11.0	H36 Retinal disorders in diseases classified elsewhere	
			N17-N19:Renal failure	1.5	9.2	N18 Chronic renal failure	
MC6 29,907	15.6	1.5	M05-M14:Inflammatory polyarthropathies	5.4	34.5	M10 Gout	0.66 (0.60-0.72)
			S80-S89:Injuries to the knee and lower leg	3.2	20.7	S83 Dislocation, sprain and strain of joints and ligaments of knee	

			L40-L45:Papulosquamous disorders	3.2	20.5	L40 Psoriasis	
			G50-G59:Nerve, nerve root and plexus disorders	2.7	17.4	G56 Mononeuropathies of upper limb	
			M65-M68:Disorders of synovium and tendon	2.6	16.8	M65 Synovitis and tenosynovitis	
MC7 19,313	10.1	0.9	K70-K77:Diseases of liver	5.2	51.9	K76 Other diseases of liver	1.00
			B15-B19:Viral hepatitis	3.2	32.0	B18 Chronic viral hepatitis	
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	2.6	25.8	K80 Cholelithiasis	
MC8 19,160	10.0	0.7	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	41.0	T14 Injury of unspecified body region	0.80 (0.75-0.86)
			L50-L54:Urticaria and erythema	2.1	20.8	L50 Urticaria	
			S60-S69:Injuries to the wrist and hand	1.9	18.5	S61 Open wound of wrist and hand	
			S90-S99:Injuries to the ankle and foot	1.4	13.9	S93 Dislocation, sprain and strain of joints and ligaments at ankle and foot level	
			S00-S09:Injuries to the head	1.3	13.3	S01 Open wound of head	
MC9 13,752	7.2	0.3	E00-E07:Disorders of thyroid gland	2.4	32.8	E03 Other hypothyroidism	0.97 (0.96-0.98)
			K50-K52:Noninfective enteritis and colitis	1.7	24.1	K52 Other non-infective gastro-enteritis and colitis	
			D50-D53:Nutritional anaemias	1.2	16.3	D50 Iron deficiency anaemia	
			Q38-Q45:Other congenital malformations of the digestive system	1.2	16.2	Q40 Other congenital malformations of upper alimentary tract	
			D70-D77:Other diseases of blood and blood-forming organs	1.1	15.3	D72 Other disorders of white blood cells	
MC10	7.1	0.5	J20-J22:Other acute lower	2.6	37.2	J20 Acute bronchitis	0.86 (0.82-0.91)

13,490			respiratory infections				
			A00-A09:Intestinal infectious diseases	2.4	34.2	A09 Diarrhoea and gastro-enteritis of presumed infectious origin	
			J09-J18:Influenza and pneumonia	1.6	22.0	J11 Influenza, virus not identified	
			B25-B34:Other viral diseases	1.0	14.6	B34 Viral infection of unspecified site	
MC11 13,434	7.0	0.4	H60-H62:Diseases of external ear	3.9	55.3	H61 Other disorders of external ear	0.88 (0.85-0.92)
			H65-H75:Diseases of middle ear and mastoid	2.1	30.5	H65 Nonsuppurative otitis media	
			H80-H83:Diseases of inner ear	1.4	19.9	H81 Disorders of vestibular function	
MC12 10,952	5.7	0.1	N20-N23:Urolithiasis	4.3	75.5	N20 Calculus of kidney and ureter	0.99 (0.98-1.00)
			N30-N39:Other diseases of urinary system	1.6	27.1	N39 Other disorders of urinary system	
MC13 8,794	4.6	0.3	M45-M49:Spondylopathies	3.1	66.7	M47 Spondylosis	0.99 (0.95-1.00)
			M40-M43:Deforming dorsopathies	1.8	38.9	M41 Scoliosis	
MC14 7,444	3.9	0.2	H40-H42:Glaucoma	2.3	59.6	H40 Glaucoma	0.91 (0.88-0.95)
			H25-H28:Disorders of lens	1.8	45.4	H25 Senile cataract	
MC15 6,161	3.2	0.0	T51-T65:Toxic effects of substances chiefly nonmedicinal as to source	2.1	65.8	T65 Toxic effect of other and unspecified substances	0.90 (0.87-0.93)
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	35.3	F20 Schizophrenia	

**Approximately unbiased (AU) probability-value

Appendix 6. Factors in women aged 45-64 years (n= 217,823).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in group, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
WF1^ 130,072	59.7	25.4	M50-M54:Other dorsopathies	35.8	59.9	10.6	69.1
			M70-M79:Other soft tissue disorders	27.0	45.2		
			M15-M19:Arthrosis	15.7	26.2		
			G50-G59:Nerve, nerve root and plexus disorders	8.5	14.3		
			M45-M49:Spondylopathies	4.3	7.3		
			M40-M43:Deforming dorsopathies	3.8	6.4		
WF2 82,301	37.8	12.0	I10-I15:Hypertensive diseases	25.6	67.6	7.0	84.5
			E65-E68:Obesity and other hyperalimentation	19.0	50.2		
			E10-E14:Diabetes mellitus	7.7	20.3		
WF3 71,436	32.8	8.1	I10-I15:Hypertensive diseases	25.6	78.0	20.2	58.6
			E10-E14:Diabetes mellitus	7.7	23.4		
			I30-I52:Other forms of heart disease	4.4	13.3		
			H25-H28:Disorders of lens	1.7	5.3		
			H30-H36:Disorders of choroid and retina	1.2	3.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	1.1	3.2		
			I20-I25:Ischaemic heart diseases	1.0	3.1		
WF4 60,027	27.6	5.9	J00-J06:Acute upper respiratory infections	12.6	45.8	38.3	38.3
			N30-N39:Other diseases of urinary system	5.9	21.3		
			H60-H62:Diseases of external ear	3.6	13.1		
			J20-J22:Other acute lower respiratory infections	3.4	12.2		
			A00-A09:Intestinal infectious diseases	2.7	10.0		
			H65-H75:Diseases of middle ear and mastoid	2.5	9.2		

			J09-J18:Influenza and pneumonia	1.7	6.1		
			B25-B34:Other viral diseases	1.3	4.8		
			M60-M63:Disorders of muscles	1.2	4.4		
WF5 56,671	26.0	5.3	L20-L30:Dermatitis and eczema	9.3	35.7	8.4	77.5
			B35-B49:Mycoses	5.7	22.1		
			L60-L75:Disorders of skin appendages	4.8	18.3		
			H53-H54:Visual disturbances and blindness	4.4	16.9		
			H10-H13:Disorders of conjunctiva	3.8	14.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	3.3	12.8		
			H55-H59:Other disorders of eye and adnexa	1.2	4.7		
WF6 46,391	21.3	3.8	K20-K31:Diseases of oesophagus, stomach and duodenum	11.4	53.6	5.9	90.4
			K40-K46:Hernia	5.7	26.7		
			K80-K87:Disorders of gallbladder, biliary tract and pancreas	4.0	18.8		
			K70-K77:Diseases of liver	2.9	13.8		
			Q38-Q45:Other congenital malformations of the digestive system	1.5	7.1		
WF7 41,492	19.0	0.6	M80-M85:Disorders of bone density and structure	11.3	59.5	5.1	95.5
			D50-D53:Nutritional anaemias	8.3	43.5		
WF8 40,619	18.6	0.0	F10-F19:Mental and behavioural disorders due to psychoactive substance use	18.6	100.0	4.3	99.8
WF9 23,729	10.9	0.6	J40-J47:Chronic lower respiratory diseases	8.2	74.9	4.1	103.9
			J20-J22:Other acute lower respiratory infections	3.4	30.8		

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor
^ Abbreviation of sex, method and number (WF1: Women Factor 1)

Appendix 7. Factors in men aged 45-64 years (n= 191,171).

Factorial order number, n	Prevalence 1, %*	Prevalence 2, %†	Blocks of diagnoses	Prevalence in age, %	Prevalence in factor, %	Variance proportion, %	Cumulative Variance proportion, %
MF1^ 118,037	61.7	26.1	E70-E90:Metabolic disorders	42.2	68.3	5.1	94.8
			I10-I15:Hypertensive diseases	32.6	52.7		
			E65-E68:Obesity and other hyperalimentation	14.6	23.6		
			M05-M14:Inflammatory polyarthropathies	5.4	8.7		
MF2 75,315	39.4	8.7	I10-I15:Hypertensive diseases	32.5	82.6	28.5	28.5
			I30-I52:Other forms of heart disease	6.9	17.6		
			I20-I25:Ischaemic heart diseases	5.0	12.6		
			I70-I79:Diseases of arteries, arterioles and capillaries	2.4	6.1		
			I60-I69:Cerebrovascular diseases	1.8	4.6		
			N17-N19:Renal failure	1.5	3.7		
MF3 73,638	38.5	4.4	F10-F19:Mental and behavioural disorders due to psychoactive substance use	33.6	87.2	5.3	89.6
			K70-K77:Diseases of liver	5.2	13.6		
			B15-B19:Viral hepatitis	3.2	8.4		
			F20-F29:Schizophrenia, schizotypal and delusional disorders	1.1	2.9		
MF4 66,303	34.7	5.1	M50-M54:Other dorsopathies	27.8	80.2	7.3	77.8
			M15-M19:Arthrosis	7.7	22.2		
			M45-M49:Spondylopathies	3.1	8.8		
			M40-M43:Deforming dorsopathies	1.8	5.2		
MF5 34,903	18.3	2.5	L20-L30:Dermatitis and eczema	7.5	41.3	22.2	50.7
			B35-B49:Mycoses	4.1	22.5		
			H53-H54:Visual disturbances and	3.9	21.3		

			blindness				
			H10-H13:Disorders of conjunctiva	3.0	16.6		
			L80-L99:Other disorders of the skin and subcutaneous tissue	2.5	13.8		
MF6 27,697	14.5	1.8	J00-J06:Acute upper respiratory infections	8.9	61.3	10.4	61.0
			J20-J22:Other acute lower respiratory infections	2.6	18.1		
			A00-A09:Intestinal infectious diseases	2.4	16.7		
			J09-J18:Influenza and pneumonia	1.6	10.7		
			B25-B34:Other viral diseases	1.0	7.1		
MF7 33,568	17.6	2.2	E10-E14:Diabetes mellitus	14.2	80.8	9.5	70.5
			H40-H42:Glaucoma	2.3	13.2		
			H25-H28:Disorders of lens	1.8	10.1		
			H30-H36:Disorders of choroid and retina	1.7	10.0		
MF8 25,121	13.1	0.7	K20-K31:Diseases of oesophagus, stomach and duodenum	11.5	87.8	6.5	84.3
			D50-D53:Nutritional anaemias	1.2	8.9		
			Q38-Q45:Other congenital malformations of the digestive system	1.2	8.9		
MF9 15,974	8.4	0.4	T08-T14:Injuries to unspecified part of trunk, limb or body region	4.1	49.2	5.0	99.8
			S80-S89:Injuries to the knee and lower leg	3.3	38.7		
			S90-S99:Injuries to the ankle and foot	1.4	16.7		
MF10 8,271	4.3	0.0	N20-N23:Urolithiasis	4.3	100.0	3.9	103.7

*Individuals from ≥1 diagnosis in the factor/ † Individuals from with ≥2 diagnosis in the factor

^ Abbreviation of sex, method and number (MF1: Men Factor 1)

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cross-sectional studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*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	6-8
Bias	9	Describe any efforts to address potential sources of bias	-
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	
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	6-9
		(*d*) If applicable, describe analytical methods taking account of sampling strategy	6-9
		(*e*) Describe any sensitivity analyses	6-9
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	10
		(b) Give reasons for non-participation at each stage	10
		(c) Consider use of a flow diagram	Supplementary File Appendix 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	Tables
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	Tables 2 and 3.
		(*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	-
Discussion			
Key results	18	Summarise key results with reference to study objectives	23
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	20-21
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	19
Generalisability	21	Discuss the generalisability (external validity) of the study results	20
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	23

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

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.