Classification of patients with osteoarthritis through clusters of comorbidities using 633,330 individuals from Spain

Marta Pineda-Moncusí* 1, Francesco Dernie* 1, Andrea Dell'Isola 2, Anne Kamps 3, Jos Runhaar 3, Subhashisa Swain 4, Weiya Zhang 5, Martin Englund 2, Irene Pitsillidou 6, Victoria Y Strauss 1, Danielle E Robinson 1, Daniel Prieto-Alhambra 1 ✉ and Sara Khalid 1

1 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, U.K.
2 Clinical Epidemiology Unit, Department of Clinical Sciences Lund, Orthopedics, Lund University, Sweden
3 Erasmus MC University Medical Center Rotterdam, Department of General Practice, Rotterdam, the Netherlands
4 Nuffield Department of Primary Care Health Sciences, University of Oxford, United Kingdom
5 Academic Rheumatology, Pain Centre Versus Arthritis, National Institution for Health and Research Biological Centre, School of Medicine, University of Nottingham, United Kingdom
6 EULAR Patient Research Partner (PRP)

*Joint first authors, ✉ Corresponding author

Corresponding Author
Daniel Prieto-Alhambra
Botnar Research Centre, Old Road, Headington, Oxford OX3 7LD
daniel.prietoalhambra@ndorms.ox.ac.uk

Abstract

Objectives
To explore clustering of comorbidities among patients with a new diagnosis of osteoarthritis (OA) and estimate the 10-year mortality risk for each identified cluster.

Methods
This is a population-based cohort study of individuals with first incident diagnosis of OA of the hip, knee, ankle/foot, wrist/hand, or 'unspecified' site between 2006 and 2020, using SIDIAP (a primary care database representative from Catalonia, Spain). At the time of OA diagnosis, conditions associated with OA in the literature that were found in ≥1% of the individuals (n=35) were fitted into two cluster algorithms, K-means and latent class analysis (LCA). Models were assessed using a range of internal and external evaluation procedures. Mortality risk of the obtained clusters was assessed by survival analysis using Cox proportional hazards.

Results
We identified 633,330 patients with a diagnosis of OA. Our proposed best solution used LCA to identify four clusters: 'Low-morbidity (relatively low number of comorbidities), 'Back/neck pain plus mental health', 'Metabolic syndrome' and 'Multimorbidity' (higher prevalence of all study comorbidities). Compared to the 'Low-morbidity, the 'Multimorbidity' cluster had the highest risk of 10-year mortality (adjusted HR: 2.19 [95%CI: 2.15-2.23]), followed by 'Metabolic syndrome' (adjusted HR: 1.24 [95%CI: 1.22-1.27]]) and 'Back/neck pain plus mental health' (adjusted HR: 1.12 [95%CI: 1.09-1.15]).

Conclusion
Patients with a new diagnosis of OA can be clustered into groups based on their comorbidity profile, with significant differences in 10-year mortality risk. Further research is required to understand the interplay between OA and particular comorbidity groups, and the clinical significance of such results.

Keywords: Epidemiology, osteoarthritis, comorbidities, clustering

Key Messages

- Patients with newly diagnosed osteoarthritis can by classified into different clusters by their comorbidity patterns.
- Such classification can help identify ‘high-risk’ patients who require more intense attention from healthcare providers.
- The main patient sub-groups were ‘Low-morbidity’, ‘Back/neck pain plus mental health’, ‘Metabolic syndrome’ and ‘Multimorbidity’.

Introduction

Osteoarthritis (OA) is a common chronic condition affecting about 250 million people worldwide[1]. The progressive degenerative nature of the disease causes functional impairment, often severe pain, and loss of quality of life[2].

Given its chronic nature, OA often coexists alongside other chronic conditions (a.k.a. comorbidities). A systematic review has shown that patients with OA are more likely to have multiple conditions compared to patients without OA[3], and further studies have shown that this increased likelihood exists both in the years preceding a diagnosis of OA, as well as in the years after[4].

The co-existence of two or more chronic conditions is termed multimorbidity[5], and is estimated to affect between 19-27% of the UK general population[6-8]. Studies have shown that increasing multimorbidity is associated with lower socioeconomic status[7,8] and increasing age[7], and drives higher healthcare utilisation including primary care usage, prescription costs and hospitalisation[8,9]. There is a growing realisation that understanding multimorbidity is important, both in clinical practice and in the development of clinical guidelines[10,11].

Within the context of multimorbidity, there is increasing recognition of the concept of comorbidities existing in groups or 'clusters'[12]. Examining the exact conditions which co-exist within an individual, rather than simply the number of comorbidities, would allow us to understand whether a patient's chronic comorbid conditions are 'concordant' (may be treated with a unified approach), or 'discordant' (may worsen or compete with treatments for individual conditions)[13], with important repercussions for the treatment of that individual, including polypharmacy[14].

Clustering of comorbidities among individuals with OA has only recently started to be explored. Studies examining general multimorbidity have shown that musculoskeletal problems including OA are very common among people with multimorbidity[15], and often cluster with cardiovascular disease[16,17]. OA is a particularly common contributor to multimorbidity among the elderly[17]. Such multimorbidity involving OA not only leads to further negative effects on quality of life, but also complicate treatment and increase requirements for analgesia[18]. With respect to the clustering of comorbidities specifically in individuals with OA, one large scale study in the UK has recently demonstrated five distinct clusters of comorbidities which predicted general practice (GP) consultation rates and mortality[19].

In this study we used machine learning techniques to examine large-scale data from patients with OA to further explore clustering of comorbidities in primary care patients with OA in the Spanish population.

Methods

Study design, setting and data sources

We conducted a population-based cohort study using the Information System for Research in Primary Care (SIDIAP) healthcare database, which collects de-identified patient records from 279 primary care providers in Catalonia, Spain, covering around 80% of the Catalan population, or 5.8 million people[20]. This study forms part of the Comorbidities in Osteoarthritis (ComOA) project, the protocol for which has been published previously[21].

Participants and study size

We included all participants aged ≥18 years with a least one physician-recorded diagnosis of OA of the hip, knee, ankle/foot, wrist/hand, or 'unspecified' site between 1st of January 2006 to 31st of June 2020, using ICD-10 codes (International Classification of Diseases 10th revision). The index date (date of their first incidence diagnosis of OA) was identified for each participant, and participants were followed from this date. Participants were excluded if they did not have at least one year of data recorded prior

to their index date, or if they had a specific non-OA diagnosis (soft-tissue disorders, other bone/cartilage diseases) at the same joint in the 12 months prior to or after the index OA/joint pain date.

Outcomes

The outcomes of interest were 1) clusters of comorbidities in people with OA and 2) risk of mortality in 10 years.

Mortality follow-up: individuals were followed from the date of OA diagnosis until the earliest: 1) date of death, or 2) date of transfer out of catchment area or end-date of data availability in SIDIAP.

Variables

Comorbidities

A comprehensive initial list of 58 comorbidities was informed by a literature review and by expert opinion (Table 1). The extraction of comorbidities from individuals was performed at the time of OA diagnosis.

Other variables

A set of external characteristics (i.e., not included in the cluster algorithms) from individuals at index-date was used to describe the obtained clusters: sex, age, body mass index (BMI), socioeconomic status, smoking and alcohol risk. BMI was classified into four categories: 1 (underweight, BMI<18.5); 2 (healthy weight, $18.5 \leq$ BMI < 25); 3 (overweight, $25 \leq$ BMI <30); 4 (obese, BMI ≥ 30). Socioeconomic status of the individuals was measured using of the MEDEA deprivation index[22]: urban areas are represented as quintiles (i.e., from U1 to U5), were U1 are the less deprived areas and U5 are the most deprived, and rural areas (R) are diferentiated[23].

Statistical methods

The external characteristics of participants and the prevalence of each comorbidity were described at the index date. Comorbidities found in less than 1% of the study population were excluded: their inclusion in the cluster algorithms increases the running times and the sample noise rather than drive to specific cluster solutions. Individuals were then classified into different clusters using K-means and latent class analysis (LCA) algorithms.

K-means is a type of 'hard' clustering approach, where individuals can only belong to one group in a binary fashion[24,25]. In order to identify the optimal number of clusters (*k*), we evaluated the clusters both internally, using within-cluster sum of squares (WCSS) and externally, by validating the clusters based on the external characteristics of the participants within each cluster: we selected the three cluster solutions from the WCSS before their change became lower than ±1 standard deviation (compared to the prior value); and then we explored them by assessing the prevalence of the comorbidities in each of the clusters and the external variables.

In contrast to 'hard' clustering approaches, 'soft' approaches such as LCA[26,27] return the probability of an individual belonging to a particular group/cluster. To identify the potential optimal *k*, we compared the performance of the models from k=1 to k=10, using a number of metrics: entropy R[28,29]; goodness of fit tests [30,31,32] ; and log-likelihood ratio. Participants were assigned to the cluster with the higher posterior probability and then internally and externally validated using the same strategy as K-means, except for the initial selection of *k* clusters, that in this case depended on the lack of change (>±1 standard deviation) of entropy and goodness of fit tests and likelihood values.

For an easier understanding of the results, both K-means and LCA resulting clusters were assigned to a tag/identifier that clinically represents the grouped patients.

To calculate the 10-year mortality risk for each cluster, survival analysis[33] was performed using Kaplan-Meier to plot the unadjusted curve of mortality in each cluster, and the Cox proportional hazards to calculate hazard ratios (HRs). We report the HRs with 95% confidence intervals (CI), both unadjusted and adjusted for age and sex.

All statistical analyses were conducted using R 4.1.1 for Windows.

Results

A total of 633,330 patients were identified with a diagnosis of OA between 1st January 2006 and 31st June 2020. Our cohort was predominantly female (67.2%), with a mean age of 67.3 years. A large proportion of participants were either overweight (40.2%) or obese (39.9%). The baseline characteristics of the cohort is given in Table 1.

After exclusion of comorbidities with a prevalence of less than 1% (Table 2), a total of 35 comorbidities were included in the cluster analysis. The most common comorbidities were back/neck pain (33.6%) and hypertension (23.5%).

Clustering by K-means

Internal validation using WCSS showed us that the biggest reduction of the within clusters distance occur up to k=4, and solutions initially selected as potentially optimal were k=4, k=5 and k=6 (representative of the number of groups which participants could be clustered into, I.e., 4-cluster, 5-cluster and 6-cluster solutions, respectively) (Supplementary Figure S1A). However, no significant improvement was observed in 5- and 6- cluster solutions after assessing the distribution of comorbidity patterns within each cluster solution and the external variables. Thus, the 4-cluster solution was selected as the best K-means solution (Table 3).

For k=4, distribution of comorbidity patterns led us to identify the following clusters (ordered from the largest to the lowest size): 'low-morbidity' (n= 302,733, 47.8%), 'metabolic syndrome' (n= 125,590,19.8%), 'back and neck pain' (n= 124,496, 19.7%), and 'mental health' (n= 80,511, 12.7%). (Figure 1A)

Those labelled as 'low-morbidity' were defined as individuals with a lower prevalence of other comorbidities compared to the general OA population. In contrast, the term 'multimorbidity' refers to clusters of individuals with a higher prevalence of all the listed comorbidities compared to the general OA population. Cluster of 'metabolic syndrome' was characterized by the presence of hypertension in all individuals, plus above average prevalence of obesity and diabetes. This group presented a higher ratio of males (37.80%) and obese individuals (44.9% had BMI ≥30) (Figure 1B). 'Back and neck pain' cluster was defined by the 100% prevalence of this condition in all the cluster members. Whilst 'mental health' label was assigned by the significant proportion of anxiety and depression: notably, all participants with anxiety were classified into this cluster. In addition, the 'mental health' group had the highest ratio of females (78.60%).

Supplementary Figures S2 and S3 displays the 5- and 6- cluster solutions, respectively.

Clustering by LCA

After clustering by LCA, internal validation (Supplementary Figure S1B) using ABIC, BIC, CAIC and the likelihood ratio did not show a statistically optimal model. However, the decline ratio of the different parameters allowed us to exclude the clusters solutions equal or higher than k=6, since those did not improve model fit substantively. Evaluation of the mean posterior probability values show better discrimination for 4-cluster than 5-cluster models (Supplementary Tables S1). Hence, we selected the 4-cluster solution as tour preferred model.

When k=4, we identified the following clusters: 'back and neck pain plus mental health', 'multimorbidity', 'low-morbidity' and 'metabolic syndrome'. Again, 'low-morbidity' refers to individuals with a lower prevalence of other comorbidities and 'multimorbidity' refers to individuals with a higher prevalence of all the listed comorbidities, compared to the general OA population.

The cluster with the highest proportion of participants was the 'heathier' (n=394,940, 62.36%), followed by 'back and neck pain plus mental health' (n=114,718, 18.11%), 'metabolic syndrome' (n=72,532, 11.45%), and 'multimorbidity' (n=51,140, 8.07%). Whilst our overall cohort was predominantly female (67.20%), females only made up 39.00% of the 'metabolic syndrome' cluster, which had the highest proportion of men. Conversely, 'back and neck pain plus mental health' cluster had a remarkable proportion of women (83.30%) and the youngest population (mean age [years, SD] 64.2, 12.5). In contrast, 'multimorbidity' cluster had the oldest population (mean age [years, SD] 79.20, 9.47). (Figure 2)

Supplementary Figures S4 and S5 reports the 5- and 6- cluster solutions, respectively.

Survival analyses

Survival analyses for 10-year mortality (HR, 95%CI adjusted for sex and age) revealed differences between the 4-clusters identified using K-means (Table 3a) and LCA (Table 3b). The 'low-morbidity' cluster was used as the reference group in both analyses:

For K-means, the 'back and neck pain' cluster had a reduced risk of 10-year mortality (0.93, 0.91-0.95), while the 'mental health' (1.21, 1.18-1.24) and 'metabolic syndrome' (1.18, 1.16-1.20) clusters had an increased risk.

In our LCA results, all clusters, including 'back and neck pain plus mental health' (1.12, 95% CI 1.09-1.15), 'metabolic syndrome' (1.24, 95% CI 1.22-1.27) and 'multimorbidity' (2.19, 95% CI 2.15-2.23), had increased risk of mortality.

Supplementary Tables S2 and S3 reports the survival analysis for 5- and 6- cluster solutions in K-means and LCA, respectively.

Discussion

Our study of 633,330 individuals with OA from the SIDIAP database is, to our knowledge, the largest to date exploring the clustering of comorbidities among individuals with a diagnosis of OA. We found that individuals with OA can be clustered based on their comorbidity patterns into groups with significantly different risks of 10-year mortality.

While we explored clustering using two separate methods, and three different cluster solutions in each of them, a number of patterns emerged: in all solutions the larger group was the 'low-morbidity' cluster, where patients with a new diagnosis of OA had the lowest prevalence of comorbid conditions; the 'back and neck pain plus mental health' groups tended to have the highest proportion of females; those designated as 'metabolic syndrome' groups had the highest proportion of males and the highest BMI; and the 'multimorbidity' groups had high mean age. Whilst age and sex varied between groups, socioeconomic status remained relatively stable. Nonetheless, the preferred solution for both clustering methods was the 4-cluster.

When K-means and LCA 4-cluster results are compared, soft classification of LCA allows higher flexibility to detect more complex patterns, such as the interaction between back and neck pain along with mental health comorbidities, or the 'Multimorbidity' cluster. Thus, clusters obtained by LCA better represented the behaviour and interaction within the different comorbidities (i.e., the comorbidity patterns). In addition, differences in 10-year mortality were most marked in the outgoing clusters from the LCA analyses, which may therefore be of more use when risk-stratifying patients in clinical practice.

With the caveat that more studies using different populations may shed further light on an optimal clustering solution in the future, we propose the 4-clusters identified by the LCA algorithm: 'low-morbidity', 'back/neck pain plus mental health', 'metabolic syndrome' and 'multimorbidity'.

Comparison with other literature and interpretation

A number of general patterns of multimorbidity have previously been established. Systematic reviews have identified 'mental health', 'cardiovascular / metabolic' and 'musculoskeletal' as common clusters of comorbid conditions[34,35], and have found that OA with cardiovascular and/or metabolic disease is a common multimorbidity profile presenting in primary care[36]. Despite our study focussing specifically on patients with OA diagnoses, rather than the wider population, we nevertheless observed these established clusters of comorbidities in most of our analyses.

The association between cardiovascular disease and OA is established[37,38], but whether they simply co-exist or share a common aetiology, perhaps due to age-related, inflammatory, hormonal or drug-related mechanisms, remains unclear[39]. Metabolic syndrome, classically characterised by both obesity and diabetes, is a risk factor for the development of OA through metabolic changes which affect joint function[40]. The level of obesity is also associated with the clinical severity of the disease[41], and management guidelines therefore frequently recommend physical activity and weight loss as first-line treatment strategies in an effort to halt or slow the progression of the disease[42]. The association between musculoskeletal (especially back and neck) pain and mental health is also established[43] and studies have shown that this link can commence early in life[44], which may contribute to our observation that our 'back and neck pain with mental health' have low mean ages.

A previous study used LCA to cluster 221,807 OA patients from the UK into five groups[19]. The five groups identified were 'low-morbidity', 'cardiovascular', 'musculoskeletal and mental health', 'cardiovascular and mental health', and 'metabolic', which, despite differences in the specific comorbidities used for analysis, reflect our own LCA k=5 results.

Several systematic reviews have explored links between OA and mortality with varied results, likely due to underlying methodological differences between them[45-47]. In order to address some of the issues intrinsic to meta-analyses and shed further light on mortality risk in OA, a recent study used large-scale individual patient-level data from six geographically diverse cohorts and found that patients with OA-related pain, or pain and radiographic OA, had between a 35-37% increased association with reduced time to death when compared to people without OA[48]. Our data revealed that among patients with OA, their 10-year mortality risk may vary widely depending on their particular comorbidities. The largest difference seen, when compared to patients with OA who were otherwise 'low-morbidity', was among our 'multimorbidity' groups, who in some cases had almost three times the risk of 10-year mortality.

Strengths and limitations of the study

Our study has several strengths. Firstly, we used a large established database which gathers information from >80% of its target population, allowing us to extract baseline characteristics as well as information surrounding diagnoses from a large number of participants. Secondly, our exploration of different clustering methods has allowed us to assess a variety of potential clustering results for translational potential and clinical utility. Our approach to internal and external validation, as well as assessment of mortality risk, helps to improve both the reliability and the usefulness of our findings.

Our study also has limitations. Despite the inclusion of a large number of participants, we cannot be sure that our findings are generalisable to populations in other geographic regions. Secondly, the diagnosis of OA in primary care is predominantly clinical (i.e., there is no requirement for radiographic confirmation)[42], so there is a lack of validation of individual OA diagnoses. However, we attempted to partially mitigate this by excluding participants who had other soft tissue or bone related pathology. Furthermore, the recording of knee and hip OA within SIDIAP has previously been validated, both through comparison to self-reported physician diagnosed OA[49], and through the analysis of free text records[50]. On the other hand, this analysis focuses on the time of OA diagnosis, so we cannot ignore the possibility that we may be observing different stages of OA, were the low-morbidity would represent an earlier stage of the diseases and the multimorbidity the other end of the spectrum. To unravel this, further work analysing patients' trajectories is necessary.

Conclusions

The comorbidity clusters we establish in our study for patients with a new diagnosis of OA reflect established multimorbidity patterns and are similar to those reported in a previous study using a different patient population. Such classification of patients may in the future be useful to help guide specific treatment strategies for particular groups of patients, to address both their OA as well as their other comorbidities, and may help identify 'high-risk' patients who require more intense input from healthcare providers. Furthermore, clustering may provide insight into shared underlying pathophysiological mechanisms between different comorbid conditions. There is a need to further validate our results in other patient cohorts, as well as research to investigate the underlying pathological mechanisms which may link the comorbidities we see in our clusters, and trials to determine the optimal treatment strategies for different groups of patients.

Funding statement

This research was founded by the Foundation for Research in Rheumatology (FOREUM) grant (2019-2022).

Conflicts of interest statement

DPA receives funding from the UK National Institute for Health and Care Research (NIHR) in the form of a senior research fellowship and from the Oxford NIHR Biomedical Research Centre. His research group has received funding form the European Medicines Agency and Innovative Medicines Initiative. His research group has received research grant/s from Amgen, Chiesi-Taylor, GSK, Novartis, and UCB Biopharma. His department has also received advisory or consultancy fees from Amgen, Astellas, Astra Zeneca, Johnson and Johnson, and UCB Biopharma; and speaker fees from Amgen and UCB Biopharma. Janssen and Synapse Management Partners have supported training programmes organised by DPA's department and open for external participants organized by his department outside the submitted work. AK's Institute received/receives FOREUM grant for the contributions of the (co)authors of this institution to the entire project. JR and SS research group receives FOREUM research grant. WZ received European Foundation of Research for Rheumatology (FOREUM grant to support the project, NIHR-BRC Centre for infrastructure support and Pain Centre Versus Arthritis centre grant for infrastructure support and also received consulting fees from Eli Lily and Regeneration in the form of advisory board, speakers fees from Harbin Rheumatology, and Shenzhen Rheumatology and Infection Summit and also received payment/ honoraria for lectures, presentation, manuscript writing/ educational events. VS is a Full-time employee in Boehringer-Ingelheim since Feb 2022 and receives payment from Pfizer for lectures. SK receives grant from Health Data Research UK, the Alan Turing Institute and Amgen BioPharma. AD, IP, DR, MPM, FD and ME has nothing to declare.

Ethics statement

In this study, all participants records were previously collected and anonymised by SIDIAP. Thus, no direct participant recruitment was done.

Data availability statement

Data that supports the findings of this study was provided by SIDIAP database. Data access is limited to researchers from public institutions, and collaboration with private organizations is only allowed for studies required by a regulatory agency or for non-commercial studies within a European project financed by the European Commission. Moreover, availability of data is subject to protocol approval by SIDIAP's Scientific Committee and Clinical Research Ethics Committee of IDIAPJGol.

Acknowledgments

We thank the Patient Research Participants (PRP) members Jenny Cockshull, Stevie Vanhegan, and Irene Pitsillidou for their involvement since the beginning of the project. We would like to thank the FOREUM for financially supporting the research.

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *The Lancet.* 2019;393(10182):1745-1759.
2. Martel-Pelletier J, Barr AJ, Cicuttini FM, et al. Osteoarthritis. *Nature Reviews Disease Primers.* 2016;2(1):16072.
3. Swain S, Sarmanova A, Coupland C, Doherty M, Zhang W. Comorbidities in Osteoarthritis: A Systematic Review and Meta-Analysis of Observational Studies. *Arthritis Care Res (Hoboken).* 2020;72(7):991-1000.
4. Swain S, Coupland C, Mallen C, et al. Temporal relationship between osteoarthritis and comorbidities: a combined case control and cohort study in the UK primary care setting. *Rheumatology.* 2021;60(9):4327-4339.
5. Boyd CM, Fortin M. Future of Multimorbidity Research: How Should Understanding of Multimorbidity Inform Health System Design? *Public Health Reviews.* 2010;32(2):451-474.
6. Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN. Patterns of Multimorbidity in Middle-Aged and Older Adults: An Analysis of the UK Biobank Data. *Mayo Clinic Proceedings.* 2018;93(7):857-866.
7. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet.* 2012;380(9836):37-43.
8. Cassell A, Edwards D, Harshfield A, et al. The epidemiology of multimorbidity in primary care: a retrospective cohort study. *British Journal of General Practice.* 2018;68(669):e245-e251.
9. Soley-Bori M, Ashworth M, Bisquera A, et al. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. *British Journal of General Practice.* 2021;71(702):e39-e46.
10. Whitty CJM, MacEwen C, Goddard A, et al. Rising to the challenge of multimorbidity. *BMJ.* 2020;368:l6964.
11. Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity. *BMJ : British Medical Journal.* 2012;345:e6341.
12. Chudasama YV, Khunti K, Davies MJ. Clustering of comorbidities. *Future Healthcare Journal.* 2021;8(2):e224-e229.
13. !!! INVALID CITATION !!! 13.
14. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. *Expert Opin Drug Saf.* 2014;13(1):57-65.
15. Duffield SJ, Ellis BM, Goodson N, et al. The contribution of musculoskeletal disorders in multimorbidity: Implications for practice and policy. *Best Pract Res Clin Rheumatol.* 2017;31(2):129-144.
16. Simões D, Araújo FA, Monjardino T, et al. The population impact of rheumatic and musculoskeletal diseases in relation to other non-communicable disorders: comparing two estimation approaches. *Rheumatology International.* 2018;38(5):905-915.
17. Collerton J, Jagger C, Yadegarfar ME, et al. Deconstructing Complex Multimorbidity in the Very Old: Findings from the Newcastle 85+ Study. *BioMed Research International.* 2016;2016:8745670.

18. Muckelt PE, Roos EM, Stokes M, et al. Comorbidities and their link with individual health status: A cross-sectional analysis of 23,892 people with knee and hip osteoarthritis from primary care. *Journal of Comorbidity.* 2020;10:2235042X20920456.
19. Swain S, Coupland C, Strauss V, et al. Clustering of comorbidities and associated outcomes in people with osteoarthritis - A UK Clinical Practice Research Datalink study. *Osteoarthritis Cartilage.* 2022.
20. García-Gil Mdel 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.* 2011;19(3):135-145.
21. Swain S, Kamps A, Runhaar J, et al. Comorbidities in osteoarthritis (ComOA): a combined cross-sectional, case–control and cohort study using large electronic health records in four European countries. *BMJ Open.* 2022;12(4):e052816.
22. Dominguez-Berjon MF, Borrell C, Cano-Serral G, et al. [Constructing a deprivation index based on census data in large Spanish cities(the MEDEA project)]. *Gac Sanit.* 2008;22(3):179-187.
23. Nolasco A, Moncho J, Quesada JA, et al. Trends in socioeconomic inequalities in preventable mortality in urban areas of 33 Spanish cities, 1996–2007 (MEDEA project). *International Journal for Equity in Health.* 2015;14(1):33.
24. Pinedo-Villanueva R, Khalid S, Wylde V, Gooberman-Hill R, Soni A, Judge A. Identifying individuals with chronic pain after knee replacement: a population-cohort, cluster-analysis of Oxford knee scores in 128,145 patients from the English National Health Service. *BMC Musculoskelet Disord.* 2018;19(1):354.
25. Khalid S, Prieto-Alhambra D. Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research. *Current Epidemiology Reports.* 2019;6(3):364-372.
26. Naldi L, Cazzaniga S. Research Techniques Made Simple: Latent Class Analysis. *Journal of Investigative Dermatology.* 2020;140(9):1676-1680.e1671.
27. Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prev Sci.* 2013;14(2):157-168.
28. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification.* 1996;13(2):195-212.
29. Boeschoten L, Oberski D, Waal Td. Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC). *Journal of Official Statistics.* 2017;33(4):921-962.
30. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika.* 1987;52(3):345-370.
31. Gideon S. Estimating the Dimension of a Model. *The Annals of Statistics.* 1978;6(2):461-464.
32. Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. In. Vol 52. Germany: Springer; 1987:333-343.
33. George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol.* 2014;21(4):686-694.
34. Busija L, Lim K, Szoeke C, Sanders KM, McCabe MP. Do replicable profiles of multimorbidity exist? Systematic review and synthesis. *Eur J Epidemiol.* 2019;34(11):1025-1053.

35. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. *J Clin Epidemiol.* 2014;67(3):254-266.
36. 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(7):e102149.
37. Wang H, Bai J, He B, Hu X, Liu D. Osteoarthritis and the risk of cardiovascular disease: a meta-analysis of observational studies. *Sci Rep.* 2016;6:39672-39672.
38. Hall AJ, Stubbs B, Mamas MA, Myint PK, Smith TO. Association between osteoarthritis and cardiovascular disease: Systematic review and meta-analysis. *European Journal of Preventive Cardiology.* 2016;23(9):938-946.
39. Fernandes GS, Valdes AM. Cardiovascular disease and osteoarthritis: common pathways and patient outcomes. *Eur J Clin Invest.* 2015;45(4):405-414.
40. Courties A, Sellam J, Berenbaum F. Metabolic syndrome-associated osteoarthritis. *Current Opinion in Rheumatology.* 2017;29(2).
41. Raud B, Gay C, Guiguet-Auclair C, et al. Level of obesity is directly associated with the clinical and functional consequences of knee osteoarthritis. *Sci Rep.* 2020;10(1):3601.
42. NICE. Osteoarthritis: care and management. Clinical guideline [CG177}. https://www.nice.org.uk/guidance/cg177.
43. Demyttenaere K, Bruffaerts R, Lee S, et al. Mental disorders among persons with chronic back or neck pain: results from the World Mental Health Surveys. *Pain.* 2007;129(3):332-342.
44. Rees CS, Smith AJ, O'Sullivan PB, Kendall GE, Straker LM. Back and neck pain are related to mental health problems in adolescence. *BMC Public Health.* 2011;11(1):382.
45. Xing D, Xu Y, Liu Q, et al. Osteoarthritis and all-cause mortality in worldwide populations: grading the evidence from a meta-analysis. *Sci Rep.* 2016;6:24393.
46. Veronese N, Cereda E, Maggi S, et al. Osteoarthritis and mortality: A prospective cohort study and systematic review with meta-analysis. *Semin Arthritis Rheum.* 2016;46(2):160-167.
47. Han X, Liu Z, Kong L, Wang L, Shen Y. Association between osteoarthritis and mortality: a meta-analysis. *Int J Clin Exp Med.* 2017;10(1):1094-1110.
48. Leyland KM, Gates LS, Sanchez-Santos MT, et al. Knee osteoarthritis and time-to all-cause mortality in six community-based cohorts: an international meta-analysis of individual participant-level data. *Aging Clinical and Experimental Research.* 2021;33(3):529-545.
49. Prieto-Alhambra D, Nogues X, Javaid MK, et al. An increased rate of falling leads to a rise in fracture risk in postmenopausal women with self-reported osteoarthritis: a prospective multinational cohort study (GLOW). *Ann Rheum Dis.* 2013;72(6):911-917.
50. Prieto-Alhambra D, Judge A, Javaid MK, Cooper C, Diez-Perez A, Arden NK. 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(9):1659-1664.

Tables

Table 1. Baseline characteristics.

External Variables	
Patients	**N=633,330**
Sex (n(%)):	
Female	425,826 (67.2%)
Male	207,504 (32.8%)
Age (mean (sd))	67.3 (13.0)
Body mass index (mean(sd))	29.3 (5.3)
NA	*541,318*
Body mass index by categories (n(%)):	
<18.5	524 (0.57%)
18.5 to 24.9	17791 (19.3%)
25 to 29.9	36998 (40.2%)
30+	36699 (39.9%)
NA	*541,318*
QMEDEA deprivation index (n(%)):	
Urban area 1 (less deprived area)	85,843 (13.6%)
Urban area 2	87,071 (13.8%)
Urban area 3	90,159 (14.2%)
Urban area 4	89,832 (14.2%)
Urban area 5 (more deprived area)	82,812 (13.1%)
Unknown Urban area	72,498 (11.5%)
Rural area	124,629 (19.7%)
NA	*486*
Smoke status (n(%)):	
Never smokers	340834 (64.8%)
Current smokers	79004 (15.0%)
Ex-smokers	106546 (20.2%)
NA	*106,946*
Risk of alcoholism (n(%)):	
None/low	60794 (61.7%)
Moderate	36523 (37.1%)
High/alcoholic	1198 (1.2%)
NA	*534,815*

Table 2 Prevalence of individual comorbidities at baseline. Comorbidities with a prevalence of <1%, excluded from final cluster analyses, are highlighted in bold.

Comorbidities (total = 58)	n (%)
Allergy	80,449 (12.70%)
Anaemia	48,281 (7.62%)
Ankylosing spondylitis	**550 (0.09%)**
Anxiety	80,554 (12.70%)
Arrythmia	32,605 (5.15%)
Asthma	15,960 (2.52%)
Autism	**24 (0.00%)**
Back and neck pain	212,986 (33.60%)
Benign prostate hypertrophy	33,560 (5.30%)
Cataracts	**0 (0%)**
Coronary heart disease	34,300 (5.42%)
Chronic heart failure	15,850 (2.50%)
Chronic Kidney disease	36,098 (5.70%)
Chronic obstructive pulmonary disease	23,961 (3.78%)
Dementia	12,467 (1.97%)
Depression	48,757 (7.70%)
Diabetes	57,498 (9.08%)
Eczema	21,924 (3.46%)
Epilepsy	**2671 (0.42%)**
Fatigue	16,852 (2.66%)
Fibromyalgia	10,008 (1.58%)
Gall bladder stone	21,346 (3.37%)
Gastro-esophagle reflux disease	6474 (1.02%)
Gout	12,388 (1.96%)
Hearing impairment	41,563 (6.56%)
Hepatitis	**455 (0.07%)**
HIV/AIDs	**252 (0.04%)**
Hyperlipidemia	11,602 (1.83%)
Hypertension	14,9092 (23.5%)
Hyperthyroidism	**4789 (0.76%)**
Hypothyroidism	22,153 (3.50%)
Inflammatory bowel disease	14,810 (2.34%)
Insomnia	44,278 (6.99%)
Irritable bowel syndrome	**4520 (0.71%)**
Leukaemia	**915 (0.14%)**
Liver	**2336 (0.37%)**
Lymphoma	**948 (0.15%)**
Migraine	10,401 (1.64%)
Multiple sclerosis	**248 (0.04%)**
Obesity	80,387 (12.70%)
Osteoporosis	45,261 (7.15%)
Other vessel diseases	9621 (1.52%)
Parkinson	**3872 (0.61%)**
Peripheral vascular disease	**2773 (0.44%)**
Polymyalgia rheumatica	**3408 (0.54%)**

Psoriasis	8179 (1.29%)
Psoriatic arthritis	**580 (0.09%)**
Rheumatoid arthritis	**3250 (0.51%)**
Schizophrenia	**985 (0.16%)**
Sinusitis	**2675 (0.42%)**
Sjogren's syndrome	**2070 (0.33%)**
Solid malignancy	23,946 (3.78%)
Stroke	20,986 (3.31%)
Substance abuse	40,423 (6.38%)
Systemic lupus erythematosus	**504 (0.08%)**
Thrombotic diseases	**823 (0.13%)**
Tuberculosis	**1321 (0.21%)**
Vitamin D deficiency	7569 (1.20%)

Table 3. Survival analysis for 10-year mortality in 4-cluster solutions of a) K-means and B) Latent Class Analysis:

a) K-means

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
3	Low-morbidity	Ref.	Ref.
1	Back and neck pain	0.72 [0.70 - 0.73]	0.93 [0.91 - 0.95]
4	Mental health	0.87 [0.84 - 0.89]	1.21 [1.18 - 1.24]
2	Metabolic Syndrome	1.62 [1.60 - 1.65]	1.18 [1.16 - 1.20]
-	Age	-	1.14 [1.14 - 1.14]
-	Sex (male)	-	1.73 [1.71 - 1.76]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			

b) Latent Class Analysis

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
3	Low-morbidity	Ref.	Ref.
2	'Back and neck pain' plus 'mental health'	0.85 [0.83 - 0.87]	1.12 [1.09 - 1.15]
4	Metabolic Syndrome	1.70 [1.67 - 1.74]	1.24 [1.22 - 1.27]
2	Multimorbidity	5.71 [5.61 - 5.81]	2.19 [2.15 - 2.23]
-	Age	-	1.13 [1.13 - 1.13]
-	Sex (male)	-	1.68 [1.66 - 1.70]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			

Figures

Figure 1. K-means cluster solution 4. A) Distribution of comorbidity patterns and B) External validation. Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other

vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.

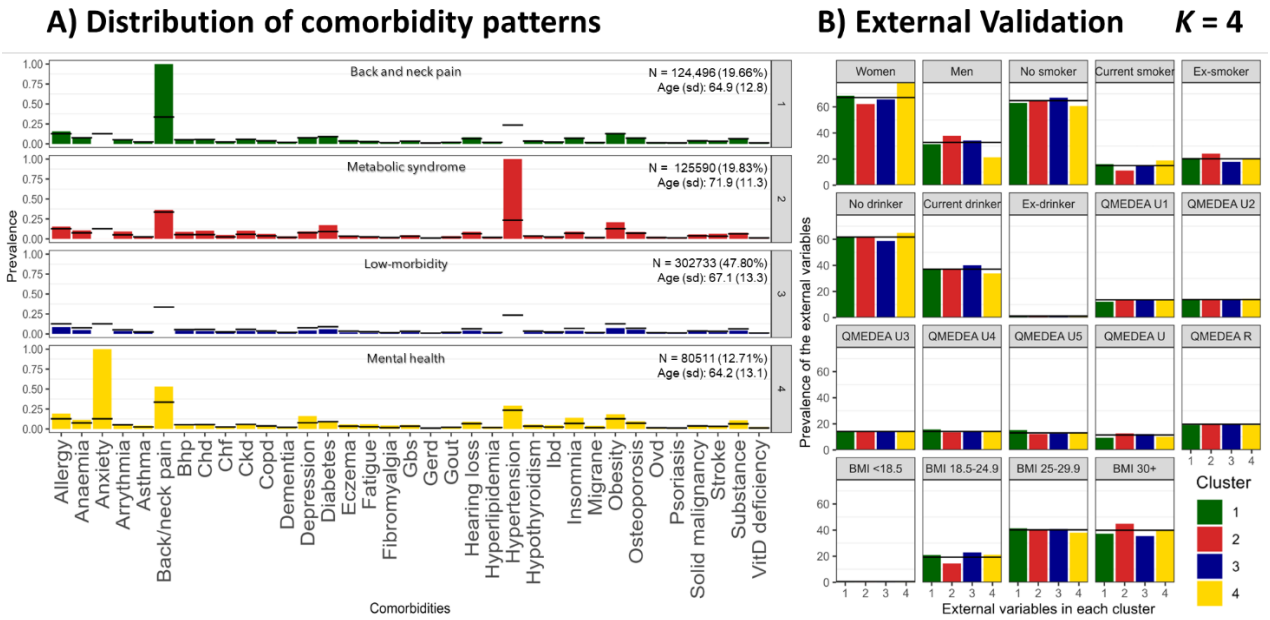
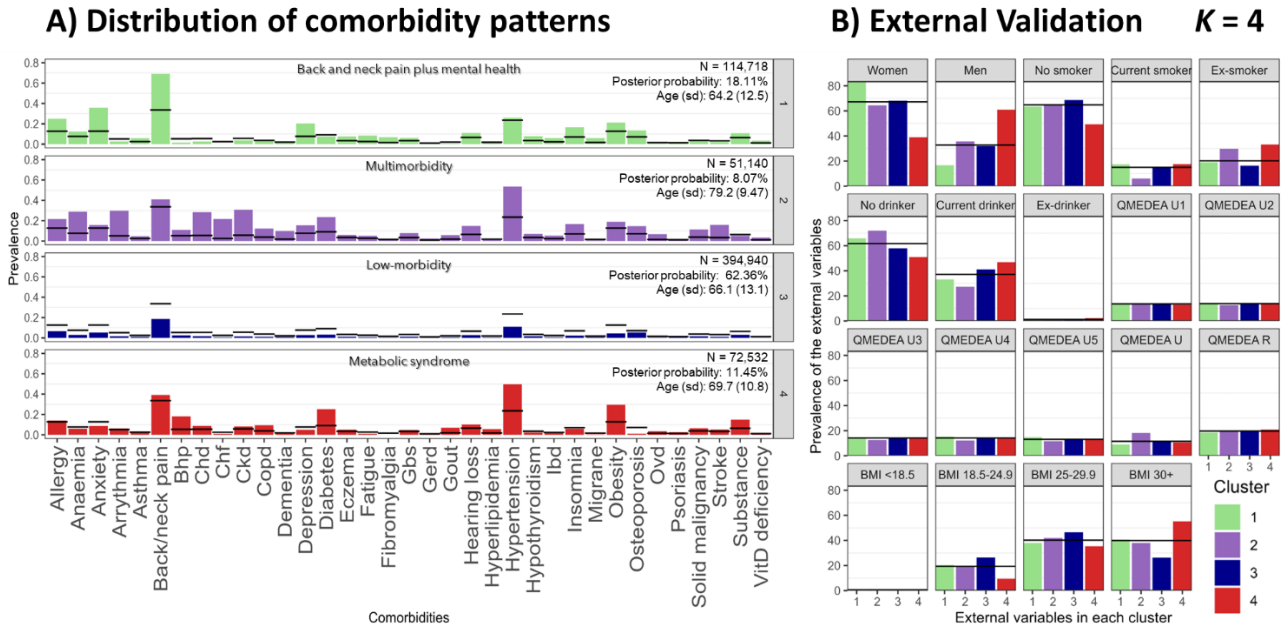
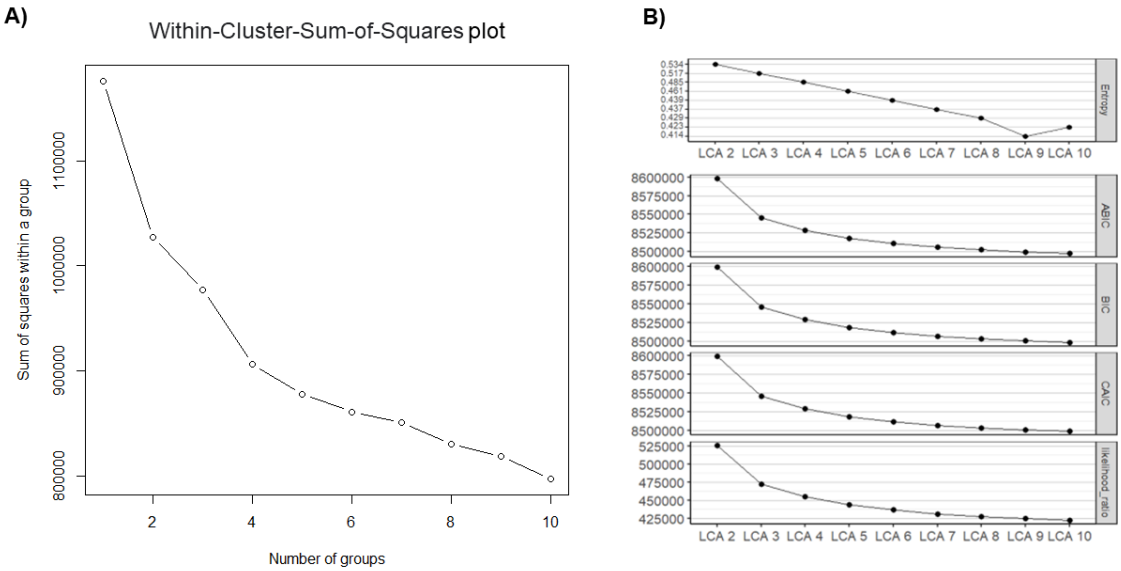


Figure 2. Latent Class Analysis cluster solution 4. A) Distribution of comorbidity patterns and B) External validation. Cluster colours are consistent in both sub-plots. Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.

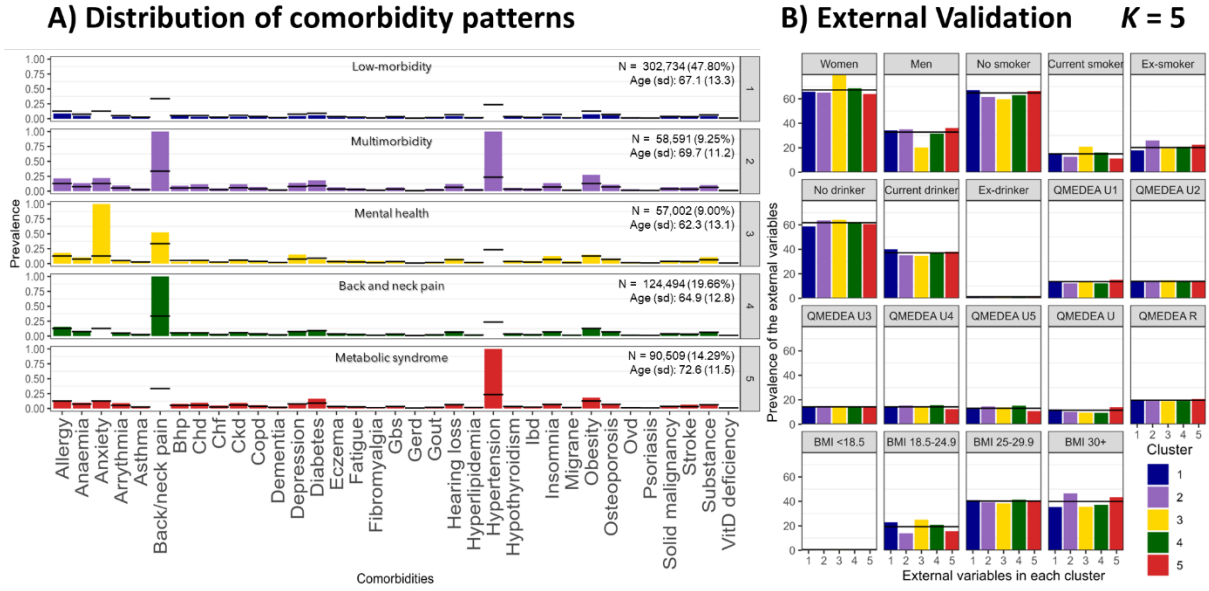


Supplementary material

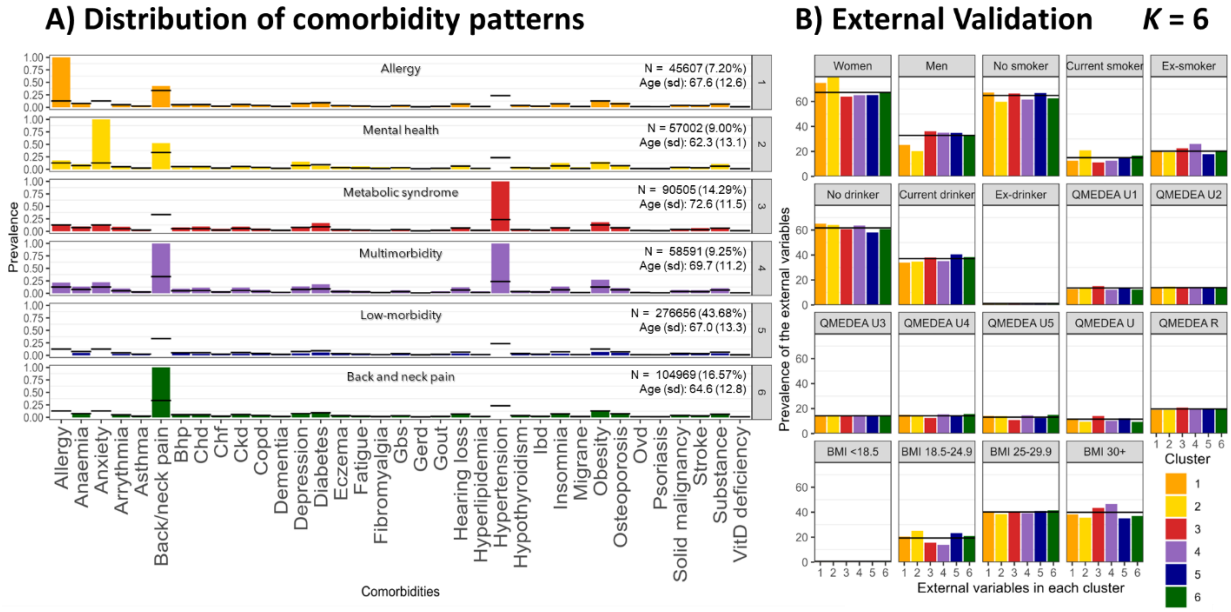
Supplementary Figure S1. Internal validations for A) K-means and B) Latent Class Analysis solutions. The internal validation of K-means is represented by the Within-Cluster-Sum-of-Squares (WCSS) for each number of k. The internal validation of LCA is represented by the entropy R2 values, goodness of fit tests (ABIC, BIC and CAIC) and likelihood ratio for each LCA when k ranges from 2 to 10.



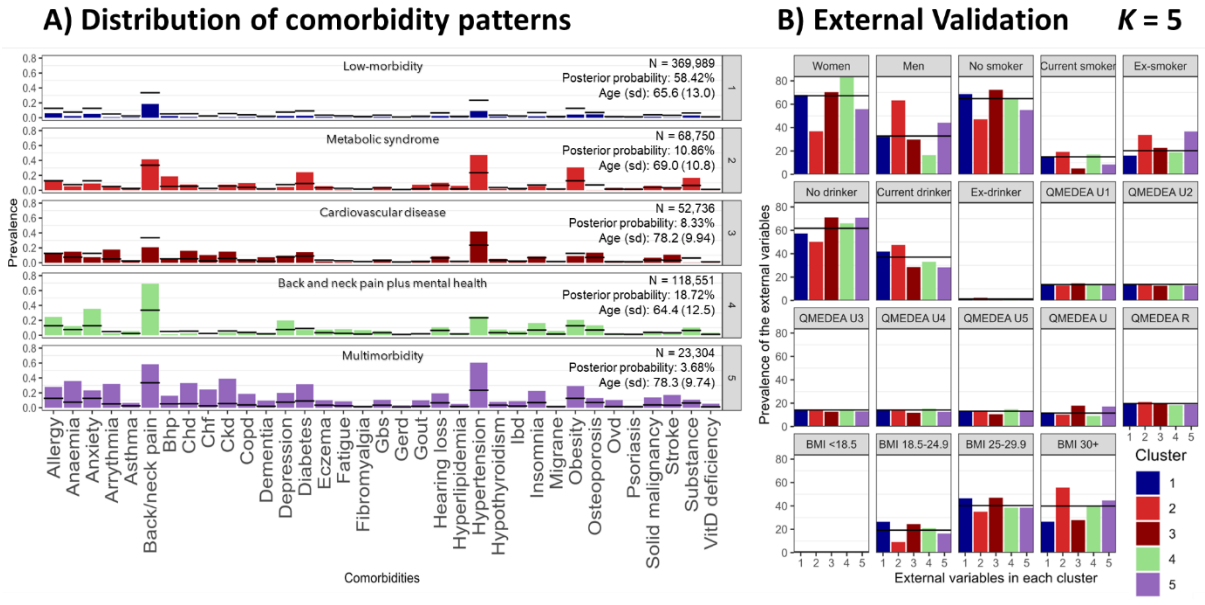
Supplementary Figure S2. K-means cluster solution 5. A) Distribution of comorbidity patterns and B) External validation. Cluster colours are consistent in both sub-plots. Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.



Supplementary Figure S3. K-means cluster solution 6. A) Distribution of comorbidity patterns and B) External validation. Cluster colours are consistent in both sub-plots. Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.

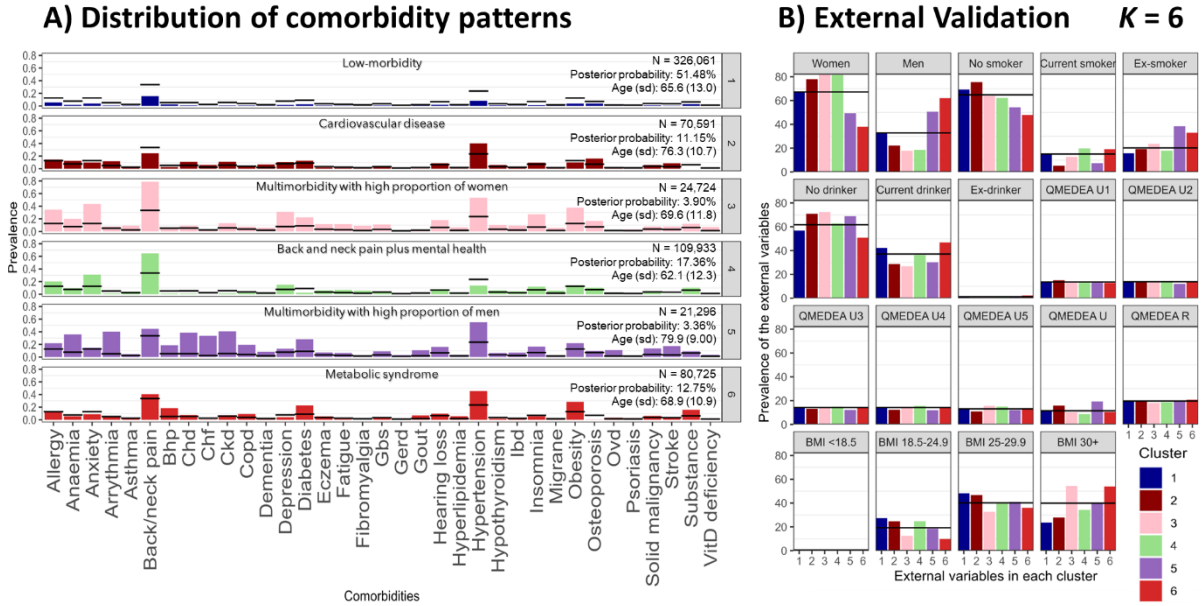


Supplementary Figure S4. Latent Class Analysis cluster solution 5. A) Distribution of comorbidity patterns and B) External validation. Cluster colours are consistent in both sub-plots. Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.



Supplementary Figure S5. Latent Class Analysis cluster solution 6. A) Distribution of comorbidity patterns and B) External validation. Cluster colours are consistent in both sub-plots.

Abbreviations: BMI, body mass index; Bhp, benign prostate hypertrophy; Chd, coronary heart disease; Ckd, chronic kidney disease; Copd, chronic obstructive pulmonary disease; Gbs, gall bladder stone; Gerd, gastroesophageal reflux disease; Ibd, inflammatory bowel disease; Ovd, other vessel diseases; Substance, substance abuse; QMEDEA, deprivation quintile index MEDEA where U is urban area (U1 is the less deprived and U5 the most), and R is rural area.



Supplementary Table S1. Mean posterior probability for each cluster solution in Laten Class Analysis (each person was assigned to the cluster with highest posterior probability).

Cluster Solution	Mean posterior probability					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
K=2	0.89	0.85	-	-	-	-
K=3	0.73	0.79	0.87	-	-	-
K=4	0.71	0.75	0.84	0.63	-	-
K=5	0.80	0.71	0.62	0.62	0.73	-
K=6	0.76	0.56	0.64	0.62	0.73	0.62

Supplementary Table S2. Survival analysis for 10-year mortality in a) 5- and b) 6-cluster solutions of K-means (each person was assigned to the cluster with highest posterior probability):

a) When k=5

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
1	Low-morbidity	Ref.	Ref.
4	Back and neck pain	0.72 [0.70 - 0.73]	0.93 [0.91 - 0.95]
3	Mental health	0.69 [0.66 - 0.71]	1.12 [1.09 - 1.15]
5	Metabolic Syndrome	1.77 [1.74 - 1.80]	1.22 [1.20 - 1.24]
2	Multimorbidity	1.22 [1.19 - 1.26]	1.14 [1.11 - 1.17]
-	Age	-	1.14 [1.14 - 1.14]
-	Sex (male)	-	1.73 [1.71 - 1.75]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			

b) When k=6

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
5	Low-morbidity	Ref.	Ref.
1	Allergy	1.00 [0.97 - 1.03]	1.03 [1.00 - 1.06]
6	Back and neck pain	0.71 [0.69 - 0.73]	0.93 [0.91 - 0.95]
2	Mental health	0.69 [0.67 - 0.71]	1.13 [1.09 - 1.16]
3	Metabolic Syndrome	1.78 [1.75 - 1.81]	1.22 [1.20 - 1.24]
4	Multimorbidity	1.24 [1.20 - 1.27]	1.15 [1.12 - 1.18]
-	Age	-	1.14 [1.14 - 1.14]
-	Sex (male)	-	1.73 [1.71 - 1.75]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			

Supplementary Table S3. Survival analysis for 10-year mortality in a) 5- and b) 6-cluster solutions of Latent Class Analysis (each person was assigned to the cluster with highest posterior probability):

a) When k=5

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
1	Low-morbidity	Ref.	Ref.
4	‘Back and neck pain’ plus ‘mental health’	0.91 [0.89 - 0.94]	1.15 [1.12 - 1.17]
3	Cardiovascular disease	4.56 [4.49 - 4.64]	1.89 [1.85 - 1.92]
2	Metabolic Syndrome	1.69 [1.66 - 1.73]	1.26 [1.23 - 1.29]
5	Multimorbidity	7.03 [6.85 - 7.21]	2.72 [2.65 - 2.79]
-	Age	-	1.13 [1.13 - 1.13]
-	Sex (male)	-	1.68 [1.65 - 1.70]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			

b) When k=6

Cluster number	Cluster name	Crude OR [95CI%]	Adjusted OR [95CI%]
1	Low-morbidity	Ref.	Ref.

4	'Back and neck pain' plus 'mental health'	0.68 [0.66 - 0.70]	1.04 [1.01 - 1.06]
2	Cardiovascular disease	3.46 [3.40 - 3.52]	1.68 [1.65 - 1.71]
5	Multimorbidity with high proportion of men	8.45 [8.25 - 8.66]	2.84 [2.77 - 2.91]
6	Metabolic syndrome	1.72 [1.68 - 1.75]	1.27 [1.25 - 1.30]
3	Multimorbidity high proportion of women	1.95 [1.86 - 2.03]	1.62 [1.55 - 1.69]
-	Age	-	1.13 [1.13 - 1.13]
-	Sex (male)	-	1.68 [1.65 - 1.70]
Abbreviations: CI, confidence intervals; Ref., Reference group; OR, Odds Ratio.			