American Journal of Epidemiology **Submitted Manuscript**

Special Collection: Click or tap here to enter text.

Title: Multimorbidity patterns, sociodemographic characteristics, and mortality: Data science insights from low-resource settings

Authors: Juan Carlos Bazo-Alvarez (1,2), Darwin Del Castillo (1,10), Luis Piza (1,3), Antonio Bernabé-Ortiz (1), Rodrigo M Carrillo-Larco (1,4), Liam Smeeth (5), Robert H. Gilman (6), William Checkley (7,8), J. Jaime Miranda (1,3,5,9)

ORCiD IDs: JCB-A: 0000-0002-6169-8049, DdC: 0000-0002-8609-0312, LP: 0000-0002-6384-6957, ABO: 0000-0002-6834-1376, RMCL: 0000-0002-2090-1856, LS: 0000-0002-9168-6022, RHG: 0000-0002-3704-9217, WC: 0000-0003-1106-8812, JJM: 0000-0002-4738-5468

Correspondence Address: J. Jaime Miranda, Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia <jaime.miranda@sydney.edu.au> Juan Carlos Bazo-Alvarez Research Department of Primary Care and Population Health, University College London (UCL), London, UK <juan.alvarez.16@ucl.ac.uk>

Joint Authorship: JCBA and JJM are co-joint corresponding authors

Affiliations: (1) CRONICAS Centre of Excellence in Chronic Diseases, Universidad Peruana Cayetano Heredia, Lima, Peru, (2) Research Department of Primary Care and Population Health, University College London, London, UK, (3) Facultad de Medicina "Alberto Hurtado", Universidad Peruana Cayetano Heredia, Lima, Peru, (4) Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA. (5) Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK, (6) Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, (7) Division of Pulmonary and Critical Care, Johns Hopkins University School of Medicine, Baltimore, MD, USA, (8) Center for Global Non-Communicable Disease Research and Training, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, (9) Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia, (10) Department of Global Health, University of Washington, Seattle, Washington, USA.

Key words: multimorbidity patterns, sociodemographic characteristics, mortality risk, cluster analysis, unsupervised machine learning, low-and-middle-income countries.

Acknowledgments[1]**:** The authors would like to thank everyone who took part in the study. JCBA is supported by the National Institute for Health and Care Research ARC North Thames. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. The establishment of the CRONICAS Cohort Study was supported by the National Heart, Lung, and Blood Institute Global Health Initiative under the contract Global Health Activities in Developing Countries to Combat Non-Communicable Chronic Diseases (project number 268200900033C-1-0-1).

[1] Study investigators, conference presentations, preprint publication information, thanks.

Funding: JCBA was funded by the National Institute for Health and Care Research, Three Schools Mental Health Programme, Grant Reference Number MHF012. JJM acknowledges having received support from the Alliance for Health Policy and Systems Research (2009/32034, 2012/253750), Bloomberg Philanthropies (grant 46129, via University of North Carolina at Chapel Hill School of Public Health), FONDECYT via CIENCIACTIVA/CONCYTEC, British Council, British Embassy and the Newton-Paulet Fund (223-2018, 224-2018), DFID/MRC/Wellcome Global Health Trials (MR/M007405/1), Fogarty International Center (R21TW009982, D71TW010877, R21TW011740, K01TW011478), Grand Challenges Canada (GMH-POC-0335-04), International Development Research Center Canada (IDRC 106887, 108167), Inter-American Institute for Global Change Research (IAI CRN3036), National Cancer Institute (NCI 1P20CA217231), National Council for Scientific and Technological Development (CNPq Brasil 408523/2023-9), National Health and Medical Research Council (NHMRC 2022566), National Heart, Lung and Blood Institute (NHLBI HHSN268200900033C, 5U01HL114180, 1UM1HL134590), National Institute for Health and Care Research (NIHR 150261, NIHR 150287), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK K23DK135798), National Institute of Mental Health (NIMH 1U19MH098780), NSW Health, Cardiovascular Elite Postdoctoral Researcher Grants (H23/37663), Swiss National Science Foundation (40P740-160366), UKRI BBSRC (BB/T009004/1), UKRI EPSRC (EP/V043102/1), UKRI MRC (MR/P008984/1, MR/P024408/1, MR/P02386X/1, MR/X004163/1, MR/X020851/1), Wellcome (074833/Z/04/Z, 093541/Z/10/Z, 103994/Z/14/Z, 107435/Z/15/Z, 205177/Z/16/Z, 214185/Z/18/Z, 218743/Z/19/Z), World Diabetes Foundation (WDF15-1224) and the World Health Organization (2021/1189041, 2022/1249357).

Conflict of Interest: The authors declare no conflict of interest.

Disclaimer: N/A

Data Availability Statement: All data used for this study are available from the corresponding Author upon reasonable request.

ABSTRACT

Multimorbidity data is typically analysed by tallying disease counts, which overlooks nuanced relationships among conditions. We identified clusters of multimorbidity and subpopulations with varying risks and examined their association with all-cause mortality using a data-driven approach. We analysed 8-year follow-up data of people ≥35 years who were part of the CRONICAS Cohort Study, a multisite cohort from Peru. First, we used Partitioning Around Medoids and multidimensional scaling to identify multimorbidity clusters. We then estimated the association between multimorbidity clusters and all-cause mortality. Second, we identified subpopulations using finite mixture modelling. Our analysis revealed three clusters of chronic conditions: respiratory (cluster 1: bronchitis, COPD and asthma), lifestyle, hypertension, depression and diabetes (cluster 2), and circulatory (cluster 3: heart disease, stroke and peripheral artery disease). While only the cluster comprising circulatory diseases showed a significant association with all-cause mortality in the overall population, we identified two latent subpopulations (named I and II) exhibiting differential mortality risks associated with specific multimorbidity clusters. These findings underscore the importance of considering multimorbidity clusters and sociodemographic characteristics in understanding mortality risks. They also highlight the need for tailored interventions to address the unique needs of different subpopulations living with multimorbidity to reduce mortality risks effectively.

INTRODUCTION

Multimorbidity poses a significant global challenge, particularly in low- and middle-income countries (LMICs).[1–4] The recent COVID-19 pandemic has shown that people with multimorbidity have a higher risk of severe disease and mortality due to SARS-CoV-2 infection.[5–7] Thus far, the bulk of literature on multimorbidity has originated from high-income settings, with limited cross-sectional evidence originating from low- and middle-income countries (LMICs),[2,8] and even fewer longitudinal studies.[9]

Multimorbidity in relation to increased mortality has become a significant public health concern. A meta-analysis by Nunes et al.[10] showed that individuals with three or more chronic conditions have a two- to three-fold higher risk of mortality compared to those without multimorbidity. The burden is especially critical in low- and middle-income countries (LMICs), where health systems are often under-resourced. Evidence from a systematic review highlighted that in LMICs, multimorbidity is associated with a 1.5- to 2-fold increase in mortality[11]. This poses significant challenges for healthcare systems, as they are often designed to treat single diseases rather than manage the complexity of multiple conditions.

Despite growing awareness of multimorbidity and its link to increased mortality, significant gaps remain in understanding how these conditions interact, particularly in low- and middle-income countries (LMICs). Recently, Whitty and Watt[12] emphasised that recognising patterns of diseases that occur together holds significant clinical relevance, particularly in individuals across all ethnicities and socio-economic levels. However, there is limited research on the specific patterns of multimorbidity in LMIC populations and how socioeconomic, cultural, and environmental factors influence these patterns and their impact on mortality [13,14].

Sociodemographic characteristics are crucial for studying multimorbidity patterns; however, to studying them properly is challenging.[15–17] While multimorbidity is often associated with elderly age groups, there is a growing appreciation that younger adults also experience multimorbidity.[18,19] In LMIC settings, diverse populations reside in rural and urban areas, further challenging the study of multimorbidity.[20] Moreover, multimorbidity tends to affect socioeconomically disadvantaged populations disproportionately.[1,21] All these sociodemographic factors do not operate isolated. Thus, they need to be studied from a multifactorial angle that facilitates the description and detection of those subpopulations at more risk.

Benefiting from an 8-year ascertainment of mortality in an adult cohort spanning various geographical settings of an LMIC, we aimed to study the association between multimorbidity and mortality and how this can vary across subpopulations defined by different sociodemographic features. To accomplish this, we set three objectives: a) to delineate clusters of physical and mental chronic conditions within the target population, b) to assess the relationship between these clusters (at baseline) and all-cause 8-year mortality across the entire population, c) pinpoint and characterise subpopulations of participants based on their sociodemographic profiles and explore whether the studied association varies across these groups.

METHODS

Study design and settings

This follow-up analysis builds upon the CRONICAS Cohort Study, conducted across four distinct Peruvian settings characterised by variations in urbanisation, air pollution levels, and altitude. Comprehensive details of this study have been previously published.[20,22] The study sites encompassed highly urbanised Lima, Peru's capital, alongside semi-urban communities in Tumbes, located in northern Peru at sea level. Additionally, high-altitude sites included urban and rural communities in Puno, positioned at an elevation of 3,825 meters above sea level in the southern Peruvian Andes.

Participants

The study enrolled full-time residents of the area aged 35 and older. It employed a stratified random sampling method, using sex and age categories (35-44, 45-54, 55-64, and 65+ years). Only one participant per household was included in the study.

Study procedures at baseline

Baseline assessments were conducted from September 2010 to January 2012, with detailed procedures outlined elsewhere.[22] Fieldwork personnel and site coordinators underwent comprehensive training covering participant selection, ethical considerations of human subjects participating in research, informed consent procedures, and clinical assessments.

Sociodemographic information, risk factors, and cardiopulmonary symptom history were collected through a questionnaire administered at enrollment.[22] In rural areas, fieldworkers proficient in local languages assisted participants with poor literacy in completing surveys. Physical measurements included weight, height, blood pressure, and spirometry before and

after bronchodilator administration. Measurement techniques are described in previous publications.[22] Blood samples were analysed at a single laboratory facility.

Follow-up procedures

Mortality status was ascertainment in 2018 using Peru's National Registry of Identification and Civil Status (RENIEC), the national authority responsible for vital records. Vital status and date of death or censoring (09/2018 in Lima, Tumbes and Puno) were the only data utilised in this study. For deceased participants, the date of death was used for analysis, while for those still alive, the date of the database search was considered the censoring date. Due to typical delays in death recording, there may be a 1-3 month lag between the actual date of death and its recording in the registry.

Study variables

The primary outcome of interest was all-cause mortality, defined as the incidence of any fatal event from the baseline enrollment until any point during the follow-up period. The duration between the baseline assessment and the date of death or censorship was measured in years.

At baseline, multimorbidity was defined as the presence of two or more chronic conditions out of a predefined list of 12 conditions: alcohol disorder, asthma, chronic bronchitis (CBR), chronic obstructive pulmonary disease (COPD), depression, heart disease, hypertension, peripheral artery disease (PAD), stroke, and type 2 diabetes mellitus, consistent with our previous study.[20] Additionally, for this analysis, gastroesophageal reflux (n=3) and lung cancer (n=2) were excluded due to the small number of cases. Instead, we incorporated obesity, defined as a body mass index BMI ≥30 kg/m^2, and total cholesterol, with high total cholesterol defined as ≥240 mg/dL[23]. While multimorbidity was defined as having two or more chronic conditions, we further quantified the number of diseases within each cluster to create a Multimorbidity Index, as described in the Statistical Methods section.

The analyses considered various sociodemographic characteristics, including sex, age (in years), education level (categorised as low, low-middle, middle, middle-high, and high), wealth index (also classified into five levels: low, low-middle, middle, middle-high, and high), and geographical site (Lima, Urban Puno, Rural Puno, Tumbes). These covariates were all assessed at baseline. Further information on geographical sites and wealth index definitions can be found elsewhere.[22,24]

Statistical methods

Our analysis strategy used a two-stage approach, summarised in **Figure 1**.

In the first stage, we utilised two algorithms, Partitioning Around Medoids (PAM) and multidimensional scaling, to identify clusters of multimorbidity. We then developed a Multimorbidity Index for each cluster to facilitate comparisons in subsequent survival analyses. We then estimated the association between multimorbidity clusters and all-cause mortality in the full population by fitting Weibull survival models, adjusting for specific comorbidities and covariates. In the second stage, we identified subpopulations within the full population using finite mixture modelling[25,26] based on socio-demographic characteristics while considering multimorbidity clusters and mortality risk. For survival models, we assessed the proportional hazard assumption graphically. Missing data on covariates were quantified and addressed using an imputation approach. We conducted formal analyses using Stata 18.0 and R 4.3.1. To ensure reproducibility, we include essential details of all models, along with analysis codes, in the Supplementary Information.

Stage One: Clustering of diseases, Multimorbidity Index, and associations with mortality

First, clusters of multimorbidity were identified using two algorithms: Partitioning Around Medoids (PAM),[27–29] and multidimensional scaling.[30] In this process, all health conditions were dichotomised into yes/no categories, and a distance matrix was computed based on the Jaccard distance, measuring the proximity of each condition. PAM operates by searching for k representative objects or medoids in this matrix. Once a set of k medoids is identified, the algorithm constructs k clusters by assigning each observation to the nearest medoid. The optimal k value is the one that minimises the sum of dissimilarities between each health condition and its closest representative object. The multidimensional scaling algorithm takes the distance matrix and returns a set of points such that the dissimilarities are approximately equal to the distances between points. The algorithm adjusts the positions of points in a low-dimensional space iteratively to minimise the selected stress function, which quantifies the difference between the original dissimilarities and the distances between points. Combining both approaches, PAM and multidimensional scaling, we drew a plot to visualise the clusters of multimorbidities. This way, we integrated information from both techniques to determine the final clusters. To verify the consistency of the clusters identified, we applied an alternative hierarchical clustering approach to the same data and contrasted its output against the PAM/multidimensional scaling's. Finally, two medical doctors verified the clinical

plausibility of the clusters. Ours contrasts with the traditional approach of merely tallying disease counts, which overlooks nuanced relationships among conditions. Thus, we were able to identify linked comorbidity groups to be analysed in terms of mortality risk, providing more actionable insights for clinical practice.

Second, a Multimorbidity Index was calculated for each multimorbidity cluster by first counting the number of conditions present per patient. Next, we subtracted the mean and then divided it by the standard deviation, resulting in a Z-score with a mean of zero and a standard deviation of one. This index was created to enable meaningful comparisons in subsequent survival analyses. Since clusters contain varying numbers of diseases, hazard ratios for each cluster cannot be directly compared when using simple counts. The standardisation provided by the index addresses this issue. It is important to note that Multimorbidity Index values are directly proportional to the number of diseases within clusters; if a person presents more diseases in the same cluster, the Multimorbidity Index of that cluster increases accordingly. Additionally, with k clusters, each person has k values of the Multimorbidity Index, representing one value per cluster.

Third, to explore the relationship between clusters of multimorbidity and mortality, we fitted Weibull survival models to estimate hazard ratios (HR) as a measure of relative risk. Specifically, HR indicates the increase or decrease in mortality risk per unit change in the multimorbidity index. For example, HR=1.3 means a 30% extra risk per each one-unit increase in the multimorbidity index. We fitted one model per cluster, adjusting for specific comorbidities within each model. This approach helped us avoid the inclusion of comorbidities as confounders when they are in the causal pathway. Consequently, we reported unadjusted HRs for each cluster and adjusted HRs after accounting for covariates.

Stage Two: Subpopulations based on sociodemographic characteristics and mortality risk associated with multimorbidity within each subpopulation

We delineated classes or sub-groups of individuals within the target population, here referred to as subpopulations, based on their socio-demographic characteristics (such as sex, age, wealth, and site), while considering their multimorbidity cluster profile (i.e., Multimorbidity Index values for all clusters) and mortality risk. Performing finite mixture modelling (FMM),[31] we identified these latent subpopulations. FMM integrates two models: a multinomial regression to estimate the probability of belonging to each subpopulation considering socio-demographics and a survival regression to predict time-to-death based on multimorbidity clusters. Following FMM analysis, we calculated posterior probabilities to

assign each patient to the most likely subpopulation. For each identified subpopulation, we report their socio-demographic attributes, morbidities, and HRs reflecting mortality risk (both per subpopulation and per cluster within each subpopulation). To provide a comprehensive overview of our findings, we generated a plot illustrating individuals categorised by age, subpopulation, and comorbidity cluster.

Research ethics

The CRONICAS Cohort Study received approval from the ethics committees of Universidad Peruana Cayetano Heredia (UPCH) and Asociación Benéfica PRISMA in Peru, as well as from the Johns Hopkins Bloomberg School of Public Health in the US. Additionally, the ethics committee at UPCH reviewed and approved the mortality follow-up assessments.

Data availability

Datasets used for the present analysis could be available from the corresponding author on reasonable request.

Code availability

Analysis code is presented in Supplementary Information.

RESULTS

Data from 3,326 participants, 51% female, mean age 55.6 (SD 12.6) years, was included in the analysis. Nearly half of the population fell into the low and low-middle education categories. Urban areas (Lima and urban Puno) were home to 55% of participants, while 45% resided in rural areas (Tumbes and rural Puno). Regarding site elevation, 63% lived at sea level (Lima and Tumbes), while 37% were in high-altitude settings (Puno). Only a quarter of the sample had a normal BMI. The most prevalent conditions at baseline were overweight/obesity (65%), hypertension (25%) and depressive symptoms (18%), followed by alcohol disorders (14%) and diabetes (8%). The prevalence of other conditions studied was ≤5%. Missing data were consistently below 18% in each study variable (**Table 1**).

Table 1. Sociodemographics and health conditions at baseline of the study cohort (N=3,326)

Sociodemographics and health conditions		n	%
sex			
	women	1699	51.1
	men	1624	48.8
	missing	3	0.1
age	mean (sd)	*55.6*	*(12.6)*
wealth index			
	Low	578	17.4
	Low-middle	646	19.4
	Middle	697	21.0
	Middle-high	692	20.8
	High	713	21.4
education			
	Low	773	23.2
	Low-middle	690	20.8
	Middle	1083	32.6
	Middle-high	123	3.7
	High	654	19.7
	missing	3	0.1
site			
	Lima	1079	32.4
	Urban Puno	756	22.7
	Rural Puno	473	14.2
	Tumbes	1018	30.6
alcohol disorder			
	normal	2848	85.6
	hazardous	477	14.4
	missing	1	0.0
asthma			
	no	3250	97.7
	yes	73	2.2
	missing	3	0.1
chronic bronchitis			
	no	2762	83.0
	yes	171	5.2
	missing	393	11.8
COPD			
	no	2605	78.3
	yes	154	4.6
	missing	567	17.1
heart disease diagnosed			
	no	3200	96.2
	yes	124	3.7
	missing	2	0.1
hypertension			
	no	2204	66.2
	yes	831	25.0
	missing	291	8.8

artery disease			
	no	2917	87.7
	yes	72	2.2
	missing	337	10.1
stroke diagnosed			
	no	3309	99.5
	yes	15	0.5
	missing	2	0.1
diabetes			
	no	2657	79.9
	yes	262	7.9
	missing	407	12.2
bmi			
	normal	832	25.0
	overweight/obesity	2163	65.0
	missing	331	10.0
total cholesterol			
	<274	2784	83.7
	>=274	121	3.6
	missing	421	12.7
depressive symptoms			
	no	2739	82.3
	yes	584	17.6
	missing	3	0.1

Clusters of multimorbidity and all-cause mortality in the overall population

PAM identified two clusters, named “cluster 1, respiratory” and “cluster 2, lifestyle, hypertension, depression and diabetes”. The remaining three health conditions, PAD, stroke, and heart disease, were non-clustered by PAM (**Figure 2**). However, our multidimensional scaling plot revealed that these conditions form another cluster designated “cluster 3, circulatory”. Medical professionals of the study team validated these clusters, concluding they were clinically plausible. Although the three clusters were partially consistent with those obtained from the alternative hierarchical clustering analysis, some variations were noted (see **Supplementary Information**). Given that cluster 2 encompasses six health conditions, while the other two clusters contain three each, employing a Multimorbidity Index seemed justified to standardise measures and facilitate comparisons of HRs between clusters 1, 2, and 3.

In the overall population (**Table 2**), unadjusted estimates revealed a longitudinal association between specific clusters and all-cause mortality. Specifically, there was a 37% higher mortality for cluster 1, respiratory, and a 32% higher all-cause mortality for cluster 3, circulatory, per each unit increase in the multimorbidity index. However, these estimates were attenuated after further adjustment for covariates, with only cluster 3, circulatory,

showing evidence of an association with all-cause mortality in the overall population (HR 1.16).

Table 2. All-cause mortality and multimorbidity clusters (standardised index) by people classes*, which are described in terms of sociodemographics and health conditions (N=3,326)

Multimorbidity index (MMI)**, sociodemographics, health conditions	Overall N=3,326 (100%)		Subpopulation I N=2,679 (80.5%)		Subpopulation II N=647 (19.5%)	
	n	%	n	%	n	%
All-cause mortality						
no	3127	94.0	2552	95.3	127	88.9
yes	199	6.0	575	4.7	72	11.1
Subpopulations HR(95%CI)***				reference		2.6 (2.0 - 3.5)
MMI (std) - HR(95%CI)+						
cluster 1 (respiratory)	1.37	(1.24 - 1.53)	1.33	(1.16-1.51)	1.49	(1.25-1.77)
cluster 2 (lifestyle, depression & diabetes)	1.02	(0.87 - 1.19)	1.18	(0.98-1.41)	0.86	(1.65-1.11)
cluster 3 (circulatory)	1.32	(1.21 - 1.44)	1.40	(1.27-1.54)	1.26	(1.02-1.55)
MMI (std) - adjusted HR(95%CI)++						
cluster 1 (respiratory)	1.11	(0.98 - 1.25)	1.06	(0.90 - 1.23)	1.31	(1.07 - 1.60)
cluster 2 (lifestyle, depression & diabetes)	1.05	(0.89 - 1.23)	1.18	(0.97 - 1.43)	0.91	(0.68 - 1.23)
cluster 3 (circulatory)	1.16	(1.04 - 1.29)	1.19	(1.07 - 1.33)	1.18	(0.90 - 1.55)
	%	(95%CI)	%	(95%CI)	%	(95%CI)
sex						
women	51.1	(49.4 - 52.8)	52.9	(51.0 - 54.8)	44.0	(40.2 - 47.8)
men	48.9	(47.2 - 50.6)	47.1	(45.2 - 49.0)	56.0	(52.2 - 59.8)
age mean(95%CI)	55.6	(55.2 - 56.0)	55.2	(54.8 - 55.7)	57.1	(56.2 - 58.1)
wealth index						
Low	17.4	(16.1 - 18.7)	12.4	(11.1 - 13.6)	38.0	(34.3 - 41.8)
Low-middle	19.4	(18.1 - 20.8)	18.7	(17.2 - 20.2)	22.4	(19.2 - 25.6)
Middle	21.0	(19.6 - 22.3)	24.0	(22.4 - 25.6)	8.3	(6.2 - 10.5)
Middle-high	20.8	(19.4 - 22.2)	24.8	(23.2 - 26.5)	4.2	(2.6 - 5.7)
High	21.4	(20.0 - 22.8)	20.1	(18.6 - 21.6)	27.0	(23.6 - 30.5)
education						
Low	23.2	(21.8 - 24.7)	22.4	(20.8 - 23.9)	26.7	(23.3 - 30.2)
Low-middle	20.8	(19.4 - 22.1)	20.3	(18.8 - 21.8)	22.7	(19.5 - 26.0)
Middle	32.6	(31.0 - 34.2)	34.3	(32.5 - 36.1)	25.5	(22.1 - 28.9)
Middle-high	3.7	(3.1 - 4.3)	4.1	(3.4 - 4.9)	2.0	(0.9 - 3.1)
High	19.7	(18.3 - 21.0)	18.9	(17.4 - 20.4)	23.0	(19.8 - 26.3)
site						
Lima	32.4	(30.8 - 34.0)	39.7	(37.8 - 41.5)	2.5	(1.3 - 3.7)
Urban Puno	22.7	(21.3 - 24.2)	21.9	(20.3 - 23.5)	26.1	(22.7 - 29.5)
Rural Puno	14.2	(13.0 - 15.4)	0.6	(0.3 - 0.8)	70.6	(67.1 - 74.1)
Tumbes	30.6	(29.0 - 32.2)	37.8	(35.9 - 39.6)	0.8	(0.1 - 1.4)
alcohol disorder						
normal	85.7	(84.5 - 86.8)	86.9	(85.6 - 88.2)	80.5	(77.5 - 83.6)
hazardous	14.3	(13.2 - 15.5)	13.1	(11.8 - 14.4)	19.5	(16.4 - 22.5)
asthma						

	no	97.8	(97.3 - 98.3)	97.3	(96.7 - 97.9)	99.7	(0.99 - 100.0)
	yes	2.2	(1.7 - 2.7)	2.7	(2.0 - 3.3)	0.3	(0.1 - 0.7)
chronic bronchitis							
	no	93.9	(93.0 - 94.8)	94.2	(93.3 - 95.2)	92.5	(90.1 - 94.9)
	yes	6.1	(5.2 - 7.0)	5.8	(4.8 - 6.7)	7.5	(5.1 - 9.9)
COPD							
	no	93.7	(92.7 - 94.7)	94.4	(93.4 - 95.4)	90.9	(87.9 - 93.8)
	yes	6.3	(5.3 - 7.3)	5.6	(4.6 - 6.6)	9.1	(6.2 - 12.1)
heart disease diagnosed							
	no	96.3	(95.6 - 96.9)	96.1	(95.3 - 96.8)	97.1	(95.8 - 98.4)
	yes	3.7	(3.1 - 4.4)	3.9	(3.2 - 4.7)	2.9	(1.6 - 4.2)
hypertension							
	no	73.0	(71.5 - 74.6)	71.5	(69.8 - 73.3)	79.1	(75.8 - 82.4)
	yes	27.0	(25.4 - 28.5)	28.5	(26.7 - 30.2)	20.9	(17.6 - 24.2)
artery disease							
	no	97.6	(97.0 - 98.2)	97.3	(96.7 - 97.9)	98.7	(97.7 - 99.7)
	yes	2.4	(1.8 - 3.0)	2.7	(2.0 - 3.3)	1.3	(0.3 - 2.3)
stroke diagnosed							
	no	99.5	(99.3 - 99.8)	99.5	(99.2 - 99.8)	99.8	(99.5 - 100.0)
	yes	0.5	(0.2 - 0.7)	0.5	(0.2 - 0.8)	0.2	(0 - 0.5)
diabetes							
	no	91.0	(89.9 - 92.0)	90.1	(88.9 - 91.3)	94.6	(92.5 - 96.6)
	yes	9.0	(8.0 - 10.1)	9.9	(8.7 - 11.1)	5.4	(3.4 - 7.5)
bmi							
	normal	28.2	(26.5 - 29.9)	24.1	(22.3 - 25.9)	45.5	(41.4 - 49.6)
	overweight/obese	71.8	(70.1 - 73.4)	75.9	(74.1 - 77.7)	54.5	(50.4 - 58.6)
total cholesterol							
	<274	95.7	(94.8 - 96.4)	95.4	(94.5 - 96.3)	96.7	(95.1 - 98.3)
	>=274	4.3	(3.6 - 5.1)	4.6	(3.7 - 5.5)	3.3	(1.7 - 4.9)
depressive symptoms							
	no	82.4	(81.1 - 83.7)	84.3	(82.9 - 85.7)	74.8	(71.4 - 78.1)
	yes	17.6	(16.3 - 18.9)	15.7	(14.3 - 17.1)	25.2	(21.9 - 28.6)

(*) Subpopulations were detected by finite mixture modelling (FMM). Modelling details can be seen in the methods section.

(**) For each multimorbidity cluster, we counted the number of conditions present and then calculated a Z-score (mean=0, sd=1). We called this score Multimorbidity Index or MMI.

(***) HR estimate from Weibull survival models (reference group is Subpopulation I, the outcome is time-to-death). Hazard Ratios (HR) for subpopulations are unadjusted.

(+) HR estimates from Weibull survival models (one model per cluster, outcome is time-to-death).

(++) HR adjusted for sex, age, education, wealth index, site and all conditions from other clusters but body mass index and cholesterol.

Subpopulations and all-cause mortality

FMM uncovered two latent groups within the overall population: subpopulation I (N=2,679; 80.5%) and subpopulation II (N=647; 19.5%) (**Table 2**). Individuals in subpopulation II faced a significantly higher risk of mortality, being 2.6 times more likely to die than those in subpopulation I.

When investigating the relationship between multimorbidity clusters and all-cause mortality in both subpopulations, cluster 2, lifestyle, hypertension, depression and diabetes, showed no association with all-cause mortality. Unadjusted HRs, per each one-unit increase in the

multimorbidity index, were consistent in magnitude across cluster 1 (respiratory) and cluster 3 (circulatory). However, after adjusting for covariates, cluster 1 remained associated with mortality in subpopulation II (HR 1.31), while cluster 3 remained associated with mortality in subpopulation I (HR 1.19).

In simpler terms, our analysis revealed that respiratory multimorbidity poses a higher risk of all-cause mortality among individuals from subpopulation II. In contrast, cardiovascular multimorbidity presents a greater risk among those from subpopulation I. Interestingly, this distinction was not apparent when analysing the entire population.

Characteristics of subpopulations

The profile of individuals in subpopulation II predominantly consisted of males, with a higher proportion falling into the lowest categories of wealth and residing primarily in highland areas, mainly rural Puno, followed by urban Puno. In contrast, individuals in subpopulation I were predominantly female, slightly younger (by approximately two years), and predominantly resided in sea-level areas such as Lima and Tumbes. The morbidity profile of both subpopulations is detailed in **Table 2**.

Figure 3 illustrates that individuals from subpopulation II exhibit fewer morbidities within cluster 3 (circulatory) than those in subpopulation I. Moreover, individuals from subpopulation II tend to have more simultaneous morbidities within cluster 2 (lifestyle, hypertension, depression, diabetes) before the age of 70, whereas those in subpopulation I may experience a similar number of comorbidities before and after that age. Additionally, any subpopulation may present with more than one morbidity from cluster 1 (respiratory) between the ages of 40 and 80.

DISCUSSION

Main findings

We used mortality data from a cohort study spanning various socioeconomic and environmental settings to identify clusters of multimorbidity and subpopulations with varying risks and examined their association with all-cause mortality. Our analysis revealed three distinct clusters of chronic conditions within the target population: respiratory (cluster 1), lifestyle, hypertension, depression, and diabetes (cluster 2), and circulatory (cluster 3). Interestingly, only the cluster comprising circulatory diseases showed a significant association with all-cause mortality in the overall population.

Additionally, our investigation identified two latent subpopulations within the target population (based on socioeconomic factors), each exhibiting different mortality risks and, more interestingly, differential risks associated with specific multimorbidity clusters. Individuals in subpopulation II, characterised by a majority of males, older age, lower wealth, and highland residence (mostly rural), faced i) more than twice the mortality risk of subpopulation I and ii) a higher mortality risk linked to respiratory multimorbidity (cluster 1). Conversely, individuals in subpopulation I, primarily females, younger, wealthier, and residing at sea level, faced a higher mortality risk associated with circulatory multimorbidity (cluster 3).

Clustering in the study of multimorbidity

Chronic conditions manifest in clusters rather than randomly arising, often sharing common risk factors.[31,32] However, there is currently no standardised method for identifying and analysing these cluster patterns, with factor analysis and hierarchical clustering being two commonly employed techniques.[32,33] Replicating multimorbidity clusters using different approaches may prove challenging, highlighting the importance of appropriate measures and analysis to understand multimorbidity patterns and inform adequate responses, including prevention strategies.[12,34–36]

However, it's essential to exercise caution when interpreting clustering outcomes, as data do not solely dictate these[37] but are also influenced by methodological decisions and researcher input. Our analysis, as shown in the case of the identification of cluster 3 (circulatory), demonstrates the significance of such considerations. We employed multiple clustering methods and ensured clinical relevance through team review, underscoring the meticulous approach adopted. The association between multimorbidity clusters, mortality, and, particularly, sociodemographic characteristics guided the subsequent detection of subpopulations.

While our analytical approach may diverge from others due to methodological nuances, it provides valuable insights and contributes to the ongoing discourse on establishing standardised methods for analysing multimorbidity patterns.[37] We recommend considering these methodological nuances when comparing our study with literature employing similar approaches, such as unsupervised machine learning. Awareness of these nuances will enhance the interpretation of our study findings within the broader context of multimorbidity research.

Comparison with the literature

The existing literature highlights a spectrum of methodologies used to evaluate multimorbidity,[38] ranging from simple disease counts[39–41] to more nuanced approaches

considering mortality risk[42] and affected body systems.[43] While disease count remains prevalent, its limitations in capturing the complexity of disease combinations are increasingly recognised. In contrast, clustering methods offer a more nuanced understanding by grouping conditions based on shared characteristics. Notably, our data-driven analysis revealed clusters aligned with affected body systems, such as respiratory and cardiovascular, enhancing clinical interpretation compared to generic disease counting.

Our study takes a longitudinal approach by using an 8-year follow-up of a robust outcome, such as mortality, to examine multimorbidity patterns, thus departing from the more common cross-sectional studies prevalent in the literature. This approach provides a deeper understanding of how chronic conditions interact within populations over time. Importantly, our focus on LMICs addresses a notable gap in the literature, as these settings are often underrepresented in longitudinal multimorbidity studies.[44] By including a range of LMIC-specific settings in our analysis, we aim to capture the unique challenges and complexities faced by populations in these regions, thereby enriching our understanding of multimorbidity on a global scale.

Studies across different populations and employing varied analytical techniques identify clusters predominantly comprising cardiovascular and respiratory conditions. For instance, studies in the UK,[45,46] China,[47] and Chile,[36,47] have consistently reported clusters featuring cardiovascular and respiratory diseases, with diabetes and hypertension often clustered together.[37,48] Despite differences in population demographics and analysis methods, the consistency of these clusters underscores their relevance across diverse settings. Notably, our approach differed from traditional methods by first identifying disease clusters and subsequently characterising subgroups of individuals based on sociodemographic factors. This innovative approach allowed for a more nuanced understanding of multimorbidity dynamics within the target population, revealing distinct subgroups with differential mortality risks.

Our findings for Subpopulation II align with evidence showing that rural and socioeconomically disadvantaged groups in LMICs face elevated health risks due to limited healthcare access and a higher prevalence of chronic diseases[49,50]. Respiratory multimorbidity (cluster 1) emerged as a major contributor to mortality in this group, emphasising the need for interventions targeting respiratory conditions in rural and high-altitude areas where poor indoor air quality and limited healthcare infrastructure may worsen these issues[51]. In contrast, the risk profile of Subpopulation I likely reflects the effects of urbanisation and lifestyle factors like dietary patterns and physical inactivity, both linked to cardiovascular disease in more affluent LMIC settings[52,53]. These findings highlight the need

for context-specific multimorbidity management that addresses sociodemographic and environmental influences on health outcomes in varied populations.

Relevance for public health

As our study revealed, multimorbidity significantly impacts mortality rates and healthcare utilisation.[1,54] Individuals with multimorbidity face higher risks of premature death and more frequent hospital admissions compared to those with single chronic conditions,[55,56] aligning with our findings of distinct clusters of chronic conditions associated with varying mortality risks. Moreover, managing multimorbidity often involves complex treatment regimens and increased healthcare costs,[57,58] which burden individuals and healthcare systems.[59–61] Our identification of specific clusters of chronic conditions, such as respiratory and circulatory diseases, underscores the importance of tailored priority settings to address the unique needs of different subpopulations, as highlighted in our study.

The economic burden of multimorbidity extends beyond healthcare costs, affecting households through increased caregiving responsibilities and potential loss of income.[34,62–64] These challenges are particularly pronounced in LMICs, where social determinants of health exacerbate the impact of multimorbidity. In light of these findings, accurately identifying multimorbidity clusters is crucial for optimising healthcare delivery and resource allocation. By understanding how clusters form within populations, healthcare providers can develop targeted interventions to improve outcomes such as mortality rates, hospitalisations, and overall healthcare costs, ultimately enhancing public health efforts in managing multimorbidity.

Strengths and limitations

Our study leveraged longitudinal data from diverse socioeconomic and environmental settings, highlighting the broad spectrum of populations represented and underscoring the robustness of our findings, which reflect the multifaceted nature of multimorbidity across different contexts. Additionally, our emphasis on mortality as a primary outcome provides a robust measure of disease burden and underscores the severity of multimorbidity's impact on health outcomes. Focusing on mortality ensured a comprehensive assessment of the risks associated with different multimorbidity clusters, contributing valuable insights for healthcare decision-making and resource allocation. Also, our two-stage analytical approach involving health conditions and individuals yielded valuable insights. Moreover, identifying subpopulations based on sociodemographic characteristics further enriches our understanding of how multimorbidity patterns manifest across different groups. Ultimately, our study provides valuable insights into these patterns and their association with mortality,

and it also contributes to advancing methodological approaches in this field, thus facilitating more comprehensive analyses in future research endeavours.

However, it is essential to acknowledge several limitations in our study. Firstly, the inclusion of self-reported conditions, such as depression assessed through validated questionnaires, introduces a degree of uncertainty regarding the accuracy of prevalence estimates. Secondly, our analysis was constrained by the availability of only all-cause mortality data, precluding a detailed examination of specific causes of death associated with different multimorbidity clusters. While our findings provide valuable insights into overall mortality risks, the lack of granularity limits our ability to discern the relative contributions of various chronic conditions to mortality outcomes. Additionally, our multimorbidity definition encompassed a limited number of diseases comprising 12 prevalent chronic conditions. However, this approach may only partially capture the evolving landscape of chronic diseases, particularly those experiencing increasing prevalence rates, such as non-alcoholic fatty liver disease and certain neoplasms, warranting consideration in future investigations seeking to provide a more holistic assessment of multimorbidity patterns. Our population was relatively young at baseline, which posed challenges, such as a low incidence of mortality. However, as we initially noted, there is growing recognition that multimorbidity also affects younger adults, making it essential to generate this new evidence for that population. Lastly, while our study was conducted within a single country, namely Peru, it is essential to acknowledge that Peru presents a unique testing ground for examining multimorbidity dynamics. The country's diverse populations, languages, settings, and socioeconomic differentials contribute to a rich tapestry of health determinants that offset the limitation of a single-country focus. Indeed, the participants in this study reflect populations residing across variations in urbanisation, air pollution levels, and altitude.

In conclusion, our study elucidated three distinct clusters of chronic physical and mental conditions, with only the cluster of circulatory diseases demonstrating a significant association with all-cause mortality in the entire population. We used a two-stage data-driven analytical approach to identify two latent subpopulations with different sociodemographic profiles, each exhibiting differential risks associated with specific multimorbidity clusters. Specifically, individuals in subpopulation II, characterised by older age, lower wealth, and residence in high-altitude settings, faced a higher mortality risk linked to respiratory multimorbidity. Conversely, individuals in subpopulation I, primarily younger, wealthier, and residing at sea level, faced a higher mortality risk associated with circulatory multimorbidity.

References

1. Skou, S. T. *et al.* Multimorbidity. *Nat Rev Dis Primers* **8**, 48 (2022).
2. Basto-Abreu, A. *et al.* Multimorbidity matters in low and middle-income countries. *J Multimorb Comorb* **12**, 26335565221106070 (2022).
3. Banerjee, A., Hurst, J., Fottrell, E. & Miranda, J. J. Multimorbidity: Not Just for the West. *Glob. Heart* **15**, 45 (2020).
4. Academy of Medical Sciences. *Multimorbidity: A Priority for Global Health Research*. https://acmedsci.ac.uk/policy/policy-projects/multimorbidity (2018).
5. McQueenie, R. *et al.* Multimorbidity, polypharmacy, and COVID-19 infection within the UK Biobank cohort. *PLoS One* **15**, e0238091 (2020).
6. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
7. Abate, S. M., Checkol, Y. A., Mantedafro, B. & Basu, B. Prevalence and risk factors of mortality among hospitalized patients with COVID-19: A systematic review and Meta-analysis. (2020) doi:10.2471/blt.20.260737.
8. Abebe, F., Schneider, M., Asrat, B. & Ambaw, F. Multimorbidity of chronic non-communicable diseases in low- and middle-income countries: A scoping review. *J Comorb* **10**, 2235042X20961919 (2020).
9. Nazar, G. *et al.* Latent class analyses of multimorbidity and all-cause mortality: A prospective study in Chilean adults. *PLoS One* **18**, e0295958 (2023).
10. Nunes, B. P., Flores, T. R., Mielke, G. I., Thumé, E. & Facchini, L. A. Multimorbidity and mortality in older adults: A systematic review and meta-analysis. *Arch. Gerontol. Geriatr.* **67**, 130–138 (2016).
11. Arokiasamy, P. *et al.* The impact of multimorbidity on adult physical and mental health in low- and middle-income countries: what does the study on global ageing and adult health (SAGE) reveal? *BMC Med.* **13**, 178 (2015).

12. Whitty, C. J. M. & Watt, F. M. Map clusters of diseases to tackle multimorbidity. *Nature* **579**, 494–496 (2020).
13. Asogwa, O. A. *et al.* Multimorbidity of non-communicable diseases in low-income and middle-income countries: a systematic review and meta-analysis. *BMJ Open* **12**, e049133 (2022).
14. Oni, T. *et al.* Patterns of HIV, TB, and non-communicable disease multi-morbidity in peri-urban South Africa- a cross sectional study. *BMC Infect. Dis.* **15**, 20 (2015).
15. Caballero, F. F. *et al.* Multimorbidity Patterns in Older Adults: the Role of Social Variables and Lifestyle Behaviors. *Gerontology* **69**, 716–727 (2023).
16. Álvarez-Gálvez, J. *et al.* Social determinants of multimorbidity patterns: A systematic review. *Front Public Health* **11**, 1081518 (2023).
17. Beridze, G. *et al.* Patterns of multimorbidity in primary care electronic health records: A systematic review. *J Multimorb Comorb* **14**, 26335565231223350 (2024).
18. Barnett, K. *et al.* Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* **380**, 37–43 (2012).
19. Vos, T. *et al.* Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222 (2020).
20. Miranda, J. J. *et al.* Multimorbidity at sea level and high-altitude urban and rural settings: The CRONICAS Cohort Study. *J Comorb* **9**, 2235042X19875297 (2019).
21. Pathirana, T. I. & Jackson, C. A. Socioeconomic status and multimorbidity: a systematic review and meta-analysis. *Aust. N. Z. J. Public Health* **42**, 186–194 (2018).
22. Miranda, J. J. *et al.* Addressing geographical variation in the progression of non-communicable diseases in Peru: the CRONICAS cohort study protocol. *BMJ Open* **2**, e000610 (2012).
23. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Executive summary of the third report of the national cholesterol education program

(NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *JAMA* **285**, 2486–2497 (2001).

24. Quispe, R. *et al.* The Relationship Between Socioeconomic Status and CV Risk Factors: The CRONICAS Cohort Study of Peruvian Adults. *Glob. Heart* **11**, 121-130.e2 (2016).
25. Deb, P. & Trivedi, P. K. Demand for Medical Care by the Elderly: A Finite Mixture Approach. *J. Appl. Econometrics* **12**, 313–336 (1997).
26. Deb, P. FMM: Stata module to estimate finite mixture models. *Statistical Software Components* (2012).
27. Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J. Math. Model. Algorithms* **5**, 475–504 (2006).
28. Schubert, E. & Rousseeuw, P. J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. in *Similarity Search and Applications* 171–187 (Springer International Publishing, 2019).
29. Schubert, E. & Rousseeuw, P. J. Fast and Eager k-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. *arXiv [cs.LG]* (2020).
30. Cox, T. & Cox, M. *Multidimensional Scaling*. (Chapman and Hall/CRC, 2000).
31. McLachlan, G. J., Lee, S. X. & Rathnayake, S. I. Finite Mixture Models. *Annu. Rev. Stat. Appl.* **6**, 355–378 (2019).
32. Bower, P. *et al.* Multimorbidity, service organization and clinical decision making in primary care: a qualitative study. *Fam. Pract.* **28**, 579–587 (2011).
33. Smith, S. M., O'Kelly, S. & O'Dowd, T. GPs' and pharmacists' experiences of managing multimorbidity: a "Pandora's box." *Br. J. Gen. Pract.* **60**, 285–294 (2010).
34. Jan, S. *et al.* Action to address the household economic burden of non-communicable diseases. *Lancet* **391**, 2047–2058 (2018).
35. Perone, S. A. *et al.* Report of the WHO independent high-level commission on NCDs: where is the focus on addressing inequalities? *BMJ Global Health* **5**, e002820 (2020).

36. Macinko, J., Andrade, F. C. D., Nunes, B. P. & Guanais, F. C. Primary care and multimorbidity in six Latin American and Caribbean countries. *Rev. Panam. Salud Publica* **43**, e8 (2019).

37. Ng, S. K., Tawiah, R., Sawyer, M. & Scuffham, P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. *Int. J. Epidemiol.* **47**, 1687–1704 (2018).

38. Ho, I. S.-S. *et al.* Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies. *Lancet Public Health* **6**, e587–e597 (2021).

39. Stirland, L. E. *et al.* Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. *BMJ* **368**, (2020).

40. Nguyen, H. *et al.* Prevalence of multimorbidity in community settings: A systematic review and meta-analysis of observational studies. *J Comorb* **9**, 2235042X19870934 (2019).

41. Johnston, M. C., Crilly, M., Black, C., Prescott, G. J. & Mercer, S. W. Defining and measuring multimorbidity: a systematic review of systematic reviews. *Eur. J. Public Health* **29**, 182–189 (2019).

42. Brilleman, S. L. & Salisbury, C. Comparing measures of multimorbidity to predict outcomes in primary care: a cross sectional study. *Fam. Pract.* **30**, 172–178 (2013).

43. Kato, D., Kawachi, I., Saito, J. & Kondo, N. Complex multimorbidity and mortality in Japan: a prospective propensity-matched cohort study. *BMJ Open* **11**, e046749 (2021).

44. Cezard, G., McHale, C. T., Sullivan, F., Bowles, J. K. F. & Keenan, K. Studying trajectories of multimorbidity: a systematic scoping review of longitudinal approaches and evidence. *BMJ Open* **11**, e048485 (2021).

45. Siah, K. W., Wong, C. H., Gupta, J. & Lo, A. W. Multimorbidity and mortality: A data science perspective. *J Multimorb Comorb* **12**, 26335565221105430 (2022).

46. Zhu, Y., Edwards, D., Mant, J., Payne, R. A. & Kiddle, S. Characteristics, service use and mortality of clusters of multimorbid patients in England: a population-based study. *BMC Med.* **18**, 78 (2020).
47. Fan, J. *et al.* Multimorbidity patterns and association with mortality in 0.5 million Chinese adults. *Chin. Med. J.* **135**, 648–657 (2022).
48. Rajoo, S. S., Wee, Z. J., Lee, P. S. S., Wong, F. Y. & Lee, E. S. A Systematic Review of the Patterns of Associative Multimorbidity in Asia. *Biomed Res. Int.* **2021**, 6621785 (2021).
49. Abegunde, D. O., Mathers, C. D., Adam, T., Ortegon, M. & Strong, K. The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet* **370**, 1929–1938 (2007).
50. Atun, R. *et al.* Improving responsiveness of health systems to non-communicable diseases. *Lancet* **381**, 690–697 (2013).
51. Burney, P. *et al.* Chronic obstructive pulmonary disease mortality and prevalence: the associations with smoking and poverty--a BOLD analysis. *Thorax* **69**, 465–473 (2014).
52. Yusuf, S., Reddy, S., Ôunpuu, S. & Anand, S. Global burden of cardiovascular diseases. *Circulation* **104**, 2746–2753 (2001).
53. Miranda, J. J., Kinra, S., Casas, J. P., Davey Smith, G. & Ebrahim, S. Non-communicable diseases in low- and middle-income countries: context, determinants and health policy. *Trop. Med. Int. Health* **13**, 1225–1234 (2008).
54. Salisbury, C., Johnson, L., Purdy, S., Valderas, J. M. & Montgomery, A. A. Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. *Br. J. Gen. Pract.* **61**, e12-21 (2011).
55. Menotti, A. *et al.* Prevalence of morbidity and multimorbidity in elderly male populations and their impact on 10-year all-cause mortality: The FINE study (Finland, Italy, Netherlands, Elderly). *J. Clin. Epidemiol.* **54**, 680–686 (2001).

56. Vogeli, C. *et al.* Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. *J. Gen. Intern. Med.* **22 Suppl 3**, 391–395 (2007).

57. Lehnert, T. *et al.* Review: health care utilization and costs of elderly persons with multiple chronic conditions. *Med. Care Res. Rev.* **68**, 387–420 (2011).

58. Townsend, A., Hunt, K. & Wyke, S. Managing multiple morbidity in mid-life: a qualitative study of attitudes to drug use. *BMJ* **327**, 837 (2003).

59. Shippee, N. D., Shah, N. D., May, C. R., Mair, F. S. & Montori, V. M. Cumulative complexity: a functional, patient-centered model of patient complexity can improve research and practice. *J. Clin. Epidemiol.* **65**, 1041–1051 (2012).

60. May, C. R. *et al.* Rethinking the patient: using Burden of Treatment Theory to understand the changing dynamics of illness. *BMC Health Serv. Res.* **14**, 281 (2014).

61. Mair, F. S. & May, C. R. Thinking about the burden of treatment. *BMJ* **349**, g6680 (2014).

62. Jaspers, L. *et al.* The global impact of non-communicable diseases on households and impoverishment: a systematic review. *Eur. J. Epidemiol.* **30**, 163–188 (2015).

63. Thrush, A. & Hyder, A. A. The neglected burden of caregiving in low- and middle-income countries. *Disabil. Health J.* **7**, 262–272 (2014).

64. Pesantes, M. A., Brandt, L. R., Ipince, A., Miranda, J. J. & Diez-Canseco, F. An exploration into caring for a stroke-survivor in Lima, Peru: Emotional impact, stress factors, coping mechanisms and unmet needs of informal caregivers. *eNeurologicalSci* **6**, 33–50 (2017).

Figure 1. Two-stage approach for clustering diseases, detecting subpopulations based on their sociodemographic characteristics, and analysing mortality risk related to multimorbidity

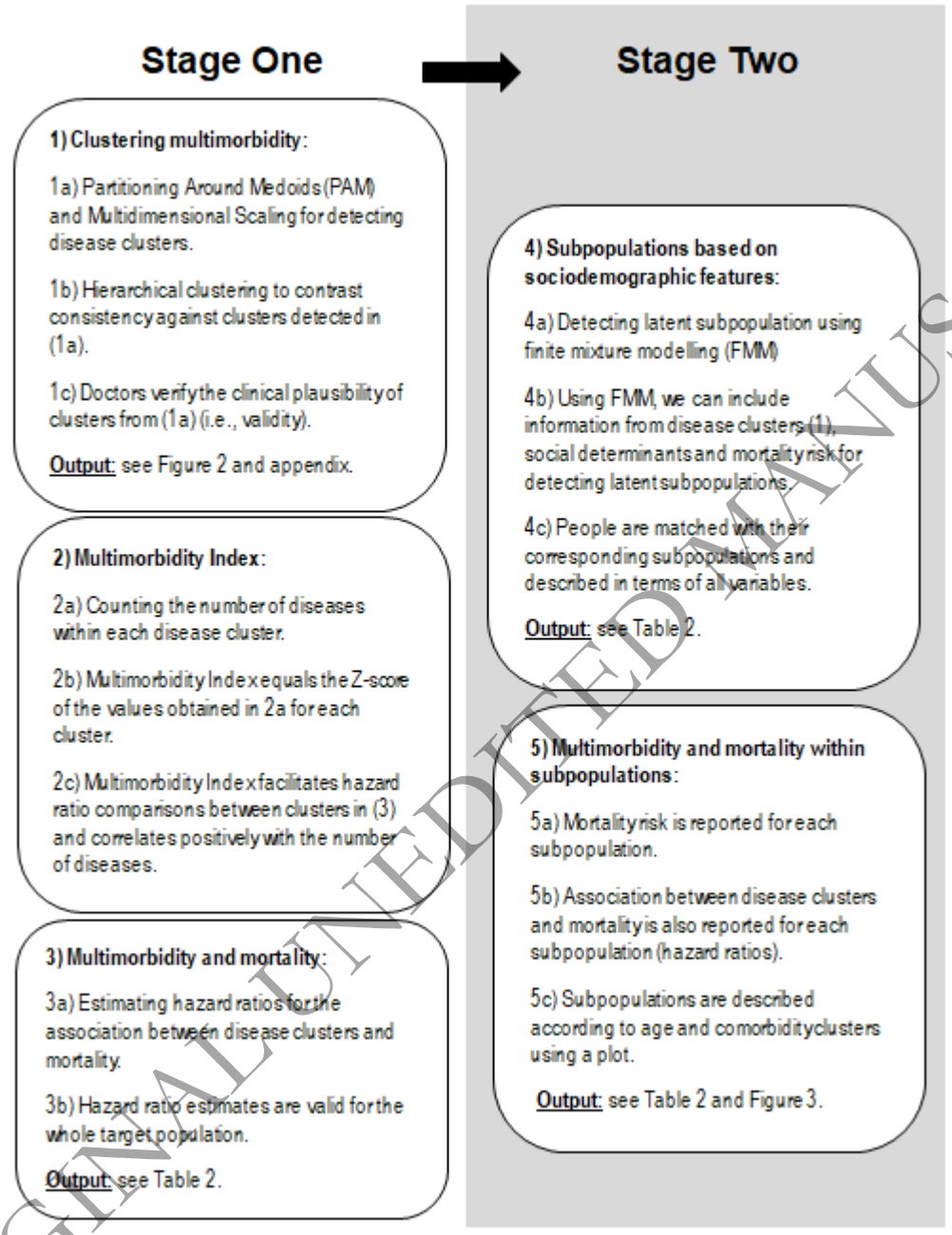

Figure 2. Multidimensional scaling plot with PAM clustering

PAM = Partition Around Medoids. The dot colour represents PAM clustering (i.e., red for cluster one, black for cluster two, and three different colours for cluster three). Dimensions (y-axis and x-axis) come from the Multidimensional scaling method. In this plot, the closer the dots are to each other, the more likely they are in the same cluster. BMI = obesity based on body mass index, PAD = Peripheral Artery Disease, CBR = Chronic Bronchitis, COPD = Chronic obstructive pulmonary disease.

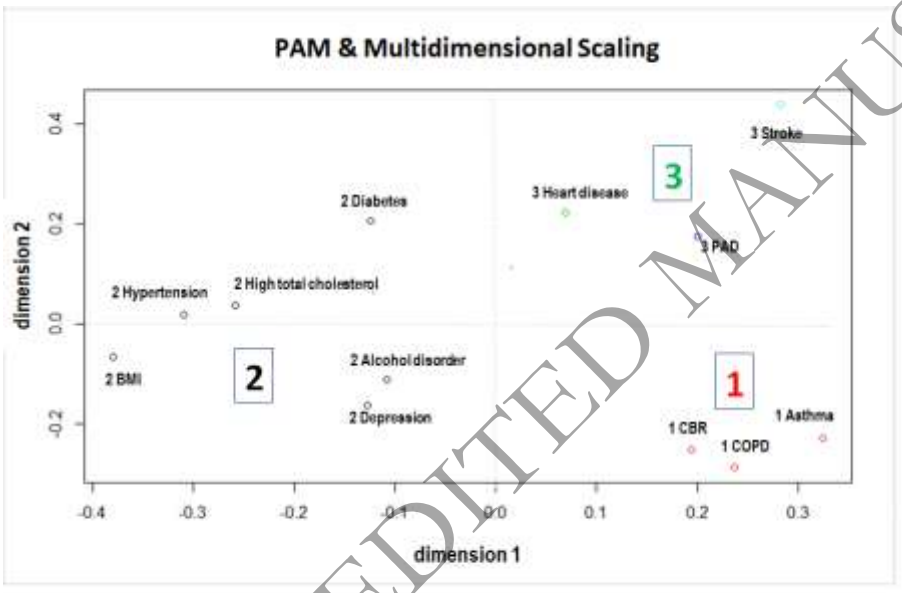

Figure 3. Multimorbidity clusters (1, 2, and 3) in people from subpopulation I (orange) and subpopulation II (blue) (N=3326)

Note: The larger the shape, the higher the number of comorbidities of the specific cluster. Circles for cluster 1, V symbols for cluster 2, and triangles for cluster 3. Orange is the subpopulation-I, and blue is the subpopulation-II. Plots show 20% of the sample for Cluster 2 only (randomly selected to facilitate visualisation).