

1 Investigating associations between physical
2 multimorbidity clusters and subsequent depression:
3 cluster and survival analysis of UK Biobank data

4

5 Lauren Nicole DeLong^{1,*}, Kelly Fleetwood², Regina Prigge², Paola Galdi¹, Bruce Guthrie³, and Jacques
6 D. Fleuriot^{1,3}

7

8 1 Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh,
9 Edinburgh, UK

10 2 Usher Institute, University of Edinburgh, Edinburgh, UK

11 3 Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, UK

12

13 * L.N.DELONG@sms.ed.ac.uk

14

15

16 Abstract

17 Background

18 Multimorbidity, the co-occurrence of two or more conditions within an individual, is a growing challenge
19 for health and care delivery as well as for research. Combinations of physical and mental health
20 conditions are highlighted as particularly important. The aim of this study was to investigate associations
21 between physical multimorbidity and subsequent depression.

22 Methods and Findings

23 We performed a clustering analysis upon physical morbidity data for UK Biobank participants aged 37-73
24 years at baseline data collection between 2006-2010. Of 502,353 participants, 142,005 had linked general
25 practice data with at least one physical condition at baseline. Following stratification by sex (77,785
26 women; 64,220 men), we used four clustering methods (agglomerative hierarchical clustering, latent class
27 analysis, *k*-medoids and *k*-modes) and selected the best-performing method based on clustering metrics.
28 We used Fisher's Exact test to determine significant over-/under-representation of conditions within each
29 cluster. Amongst people with no prior depression, we used survival analysis to estimate associations
30 between cluster-membership and time to subsequent depression diagnosis.

31 The *k*-modes models consistently performed best, and the over-/under-represented conditions in the
32 resultant clusters reflected known associations. For example, clusters containing an overrepresentation of
33 cardiometabolic conditions were amongst the largest clusters in the whole cohort (15.5% of participants,
34 19.7% of women, 24.2% of men). Cluster associations with depression varied from hazard ratio (HR)
35 1.29 (95% confidence interval (CI) 0.85-1.98) to HR 2.67 (95% CI 2.24-3.17), but almost all clusters
36 showed a higher association with depression than those without physical conditions.

37 Conclusions

38 We found that certain groups of physical multimorbidity may be associated with a higher risk of
39 subsequent depression. However, our findings invite further investigation into other factors, like social
40 ones, which may link physical multimorbidity with depression.

41

42 Introduction

43 Multimorbidity, the simultaneous occurrence of two or more long-term conditions in an individual is
44 increasingly common as populations age, and it challenges existing health systems^{1,2}. Multimorbidity is
45 more common with increasing age, in women and in the less affluent^{3,4}. Studying the co-occurrence of
46 multiple long-term conditions in the same individual has the potential to inform understanding of disease
47 causation and support planning of current and future health and care services⁵⁻⁷.

48 Depression affects millions of people worldwide^{8,9} and is ranked by the World Health
49 Organization as one of the most burdensome diseases^{9,10}. There is strong evidence that depression co-
50 occurs with other mental health disorders^{11,12}, and several ongoing studies aim to identify potential shared
51 mechanisms^{11,13}. However, previous studies have also found depression to be more common in people
52 with particular chronic physical illnesses, such as cardiovascular disease¹⁴, multiple sclerosis¹⁵, and
53 inflammatory bowel disease¹⁶. Physical ill-health might cause depression because it creates psychological
54 disturbance through ‘biographical disruption’ that threatens a sense of identity, or because of impact on
55 physical or social function. Alternatively, physical conditions may cause depression through intermediate
56 biological processes, like inflammation^{15,17}, in which case we might expect that different combinations or
57 patterns of physical conditions would be more strongly associated with depression than others.

58 Several studies have used cluster analyses to identify common patterns of physical conditions¹⁸⁻²¹,
59 typically using one method, such as agglomerative hierarchical clustering^{18,19,22}, *k*-medoids^{20,23}, Latent
60 Class Analysis^{21,24-26}, or *k*-means approaches²⁷⁻³⁰. Additionally, since morbidity data is binary (a person
61 has a condition or does not), some common clustering methods are inappropriate since they use similarity
62 measures incompatible with categorical data^{18,31}. Therefore, the aim of this study was to explore and
63 compare the use of four independent clustering methods appropriate for binary data and to examine
64 whether certain groups of physical conditions are associated with the subsequent diagnosis of depression.

65

66 Materials and methods

67 Data selection and pre-processing

68 We used data from UK Biobank³². Participants aged 37-73 years attended a baseline assessment during
69 2006-2010 which collected data on demography, lifestyle habits, health conditions, and a range of
70 physical and laboratory measurements. Participants provided written informed consent for linkage to
71 national datasets including general practice (GP) (primary care), hospital, cancer registry and death
72 records³². The UK Biobank has ethical approval from the NHS North West Research Ethics Committee
73 (reference: 21/NW/0157).

74 To robustly ascertain a broad range of long-term conditions, our study population included
75 participants with a continuous GP record from at least a year before to at least one day beyond their
76 baseline assessment. We excluded records from the UKB extract of the Vision practice management
77 system in England because the extraction process excluded participants who died prior to data extraction.
78 We also excluded participants who withdrew from the study (Supplementary Fig. 1).

79 We ascertained the presence of depression and 69 long-term physical health conditions at
80 baseline using data from the baseline visit and from linked GP, hospital, and cancer registry records based
81 on previously published lists^{33,34} (Supplementary Table 1). The UK National Health Service limits
82 registration to one practice at a time, and GP records transfer between practices so should capture an
83 individual's entire medical history. However, available hospital and cancer registry records began at
84 different times for England, Wales and Scotland. Therefore, we used all GP records up to baseline
85 assessment date, and to be consistent across countries, we used hospital and cancer registry records within
86 eight years before baseline assessment date. We used published codelists to identify diagnoses from GP
87 records using Read V2 and CTV3 diagnosis codes, hospital records using ICD-10 diagnosis codes and
88 OPCS-4 procedure codes, and cancer registry records using ICD-10 codes³⁴. We similarly ascertained

89 depression during follow-up using information from GP, hospital and death records. Eligible participants
90 with no history of depression prior to baseline were followed up to the earliest of depression diagnosis,
91 death or the end of their available GP or hospital records.

92

93 Models and metrics

94 We explored the suitability of four methods (k -modes^{35,36}, k -medoids²³, Latent Class Analysis
95 (LCA)²⁴, and agglomerative hierarchical clustering (AHC)²² (Supplementary Appendix 1)) to cluster all
96 participants based on binary features denoting the absence/presence of the 69 baseline physical
97 conditions. Participants with no physical conditions at baseline were excluded from the clustering
98 analysis. We additionally clustered separately for men (all 69 conditions) and women (67 conditions since
99 erectile dysfunction and hyperplasia of the prostate are only found in men), as well as in the whole
100 population, because of known sex differences in patterns of individual morbidities and multimorbidity
101^{37,38}.

102 To select the number of clusters for each method, we used various heuristics, including the elbow
103 method on a scree plot³⁹ for both k -modes and k -medoids, the minimal Bayesian Information Criterion⁴⁰
104 for LCA, and Hamming distance⁴¹ for AHC. To assess suitability and performance of these clustering
105 methods, we used three performance metrics (Calinski and Harabasz score⁴², Davies-Bouldin score⁴³,
106 and Silhouette score⁴⁴ (Supplementary Appendix 1)), which are appropriate for unsupervised clustering.
107 Since k -modes and LCA are sensitive to differences in initialization^{24,45}, we repeated them five times and
108 compared with other models using the mean and standard deviation across the five experiments.

109 Thereafter, we used two metrics to analyze over- and under-representation of conditions in each
110 cluster. We designed one metric, the *adjusted relative frequency (ARF)* to measure the magnitude of over-
111 or under-represented conditions within a cluster, relative to prevalence in the whole cohort. For each
112 condition, ARF is calculated as:

113
$$\text{adjusted relative frequency (ARF)} = \frac{\% \text{ with condition in cluster}}{\% \text{ with condition in the cohort}}$$

114 An ARF of exactly one indicates that the condition occurs at the same relative frequency as it does in the
115 entire cohort, and values greater or less than one indicate over- and under-representation, respectively.
116 We used Fisher's Exact Test (two-sided, $\alpha=0.05$) to evaluate whether over- or under-representation of a
117 condition in each cluster was statistically significant, using a Bonferroni correction to account for
118 multiple testing⁴⁶⁻⁴⁸. Finally, we visualised and compared statistically significant results on a *bubble*
119 *heatmap* (Supplementary Appendix 2). To allow others to conduct similar cluster analyses, we made the
120 code available as a software package (<https://github.com/laurendelong21/clusterMed>).

121

122 Survival analysis to predict depression diagnosis

123 Using participants without a record of depression at baseline, we applied Cox regression models⁴⁹ to
124 evaluate time to depression diagnosis by condition cluster, accounting for death as a competing risk⁵⁰.
125 Participants with no physical conditions at baseline were included as the reference group. We ran separate
126 models for the whole cohort and for men and women separately, examining associations between cluster
127 membership and subsequent depression. All models were adjusted for baseline age, ethnicity, country of
128 residence and deprivation. The model for the whole cohort was additionally adjusted for sex. Baseline age
129 was included in the models as a continuous variable; all other variables were categorical. Ethnicity was
130 self-reported at baseline, and we categorized it into five groups (Black, Mixed, South Asian, White, and
131 any other ethnic group⁵¹). Country of residence (England, Wales or Scotland) and area-based deprivation,
132 measured by the Townsend Deprivation Index⁵², were derived from participants' home addresses at
133 baseline. We divided the Townsend Deprivation Index into deciles within the entire UK Biobank cohort.
134 A small number of participants (368 women and 435 men) who were missing data on ethnicity, country
135 or deprivation were excluded from the survival analyses.

136 Results

137 Performance metrics across various clustering methods

138 There were 140,956 participants (73,036 women and 67,920 men) with at least one physical condition at
139 baseline who were included in the clustering analysis (Supplementary Fig. 1). Performance metrics for
140 each of the four methods explored are reported in Table 1.

141 Models based on AHC consistently achieved the poorest Calinski and Harabasz, and Davies-Bouldin
142 scores in all three cohorts. Models based on LCA had better metrics than AHC-based models, with
143 particularly high Calinski and Harabasz scores, but the Davies-Bouldin scores were consistently worse in
144 comparison to k -modes or k -medoids based models. In contrast, the best Davies-Bouldin scores were
145 achieved by the k -modes and k -medoids based models. However, since the Davies-Bouldin score assesses
146 similarity between the most similar clusters, scores may be optimistic when several clusters only contain a
147 single (or very few) participant(s). This was the case for the k -medoids models for the whole and men-
148 only cohorts. The presence of singleton clusters is also concurrent with a larger number of total clusters.
149 Specifically, all three models based on k -modes discovered eight clusters, all models using AHC
150 discovered ten clusters, and all models using LCA discovered five or six clusters. In contrast, the k -
151 medoids models discovered 25 clusters for the whole population (17 only had one participant, six for
152 women-only (no singletons), and 13 for men-only (seven singletons) (Table 1). Therefore, while k -
153 medoids models had comparable Davies-Bouldin scores to k -modes models, the results were less
154 informative and consistent. For each cohort, we therefore selected the best performing k -modes model
155 with the highest Calinski and Harabasz score among the five independent runs.

156 **Table 1. Performance metrics, number of clusters, and cluster sizes across four clustering methods.**

Cohort	Clustering method	Calinski and Harabasz score (higher is better)	Davies-Bouldin score (lower is better)	Silhouette score (higher is better)	No. of clusters	No. of singleton clusters	Median participants / cluster	(Min-max.) participants per cluster
Whole	LCA	5567.79 ± 175.06	3.80 ± 0.34	0.08 ± 0.03	6	0	17203.5	(6816 - 57564)
	AHC	628.39	4.38	0.09	10	0	2495.5	(11 - 110366)
	<i>k</i> -medoids	1270.00	2.48	0.08	25	17	1	(1 - 55553)
	<i>k</i> -modes	6079.53 ± 1228.07	2.43 ± 0.16	0.17 ± 0.02	8	0	14954.5	(974 - 39601)
Women-only	LCA	4098.56 ± 5.23	3.53 ± 0.00	0.15 ± 0.00	5	0	15794	(2373 - 30529)
	AHC	524.62	4.82	0.04	10	0	1374.5	(2 - 47728)
	<i>k</i> -medoids	3672.43	2.60	0.14	6	0	9472	(3942 - 28757)
	<i>k</i> -modes	3456.82 ± 555.59	2.36 ± 0.10	0.16 ± 0.04	8	0	9233.5	(562 - 22787)
Men-only	LCA	3134.68 ± 91.36	3.52 ± 0.02	0.10 ± 0.00	5	0	13868	(6827 - 16448)
	AHC	274.95	4.18	0.20	10	0	1709	(7 - 27353)
	<i>k</i> -medoids	1225.80	2.17	0.11	13	7	1	(1 - 25595)
	<i>k</i> -modes	2779.42 ± 463.66	2.35 ± 0.07	0.15 ± 0.03	8	0	8545.5	(649 - 15532)

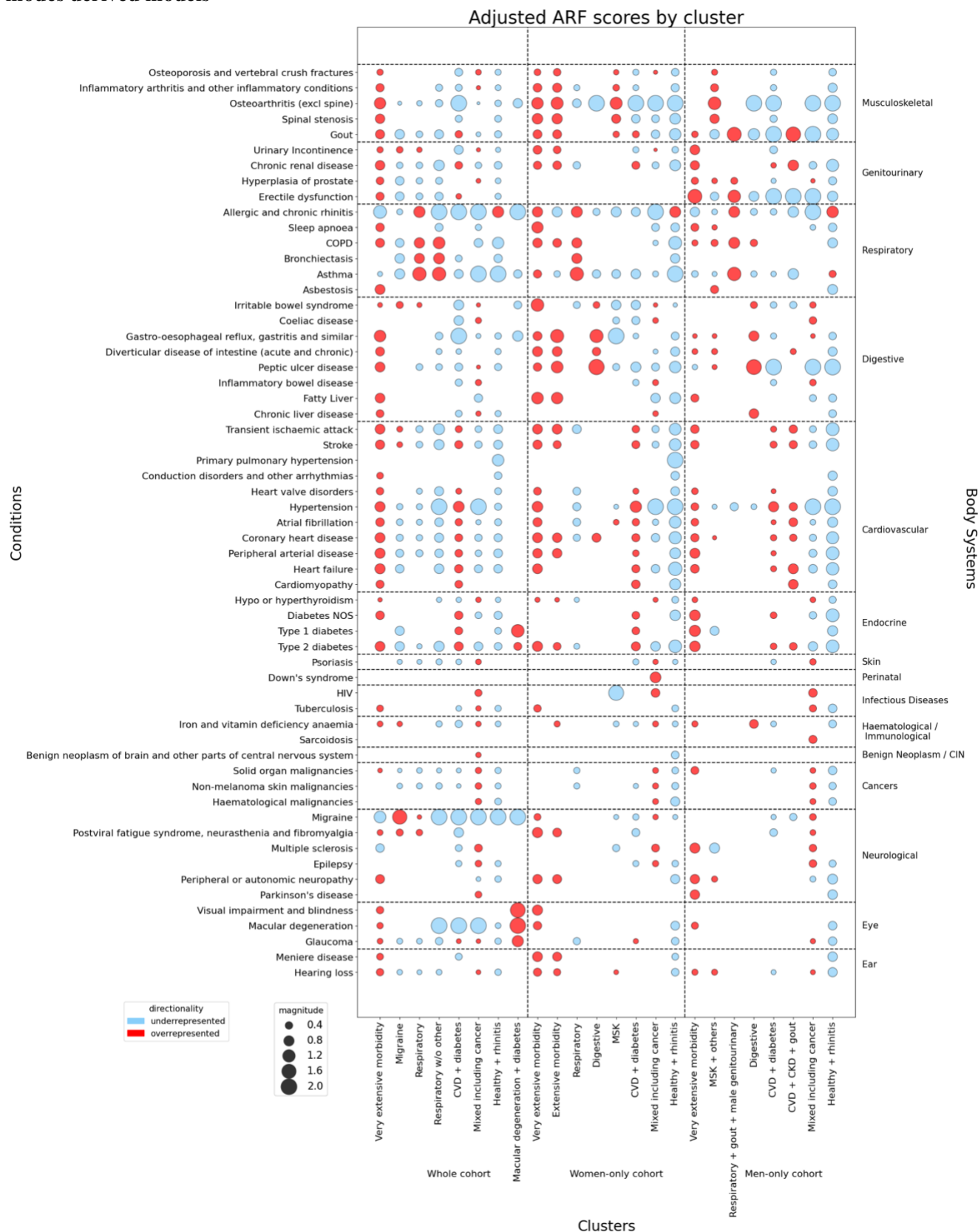
157 Results for LCA and *k*-modes are reported as average ± standard deviation across five independent models.

158 Differential representation of physical conditions within *k*-modes 159 clusters

160 Many of the significantly over-represented conditions within several clusters aligned with body
161 systems (Fig. 1) and we therefore used clinical judgement to name the clusters according to the systems or
162 conditions which were most prominent (Supplementary Tables 2-5). The four largest clusters in whole
163 cohort are *Mixed including cancer* (27.9% of participants in the cohort), *Healthy + Rhinitis* (22.2%),
164 *Cardiovascular disease (CVD) + diabetes* (15.5%), and *Very extensive morbidity* (12.5%). For women,
165 the four largest clusters are *Mixed including cancer* (29.3%), *CVD + diabetes* (19.7%), *Musculoskeletal*
166 (*MSK*) (16.4%), and *Healthy + Rhinitis* (15.9%). Finally, for men, the four largest clusters are *CVD +*
167 *diabetes* (24.2%), *Mixed including cancer* (20.8%), *MSK + others* (19.1%), and *Healthy + Rhinitis*
168 (17.2%) (Fig. 2, Table 2).

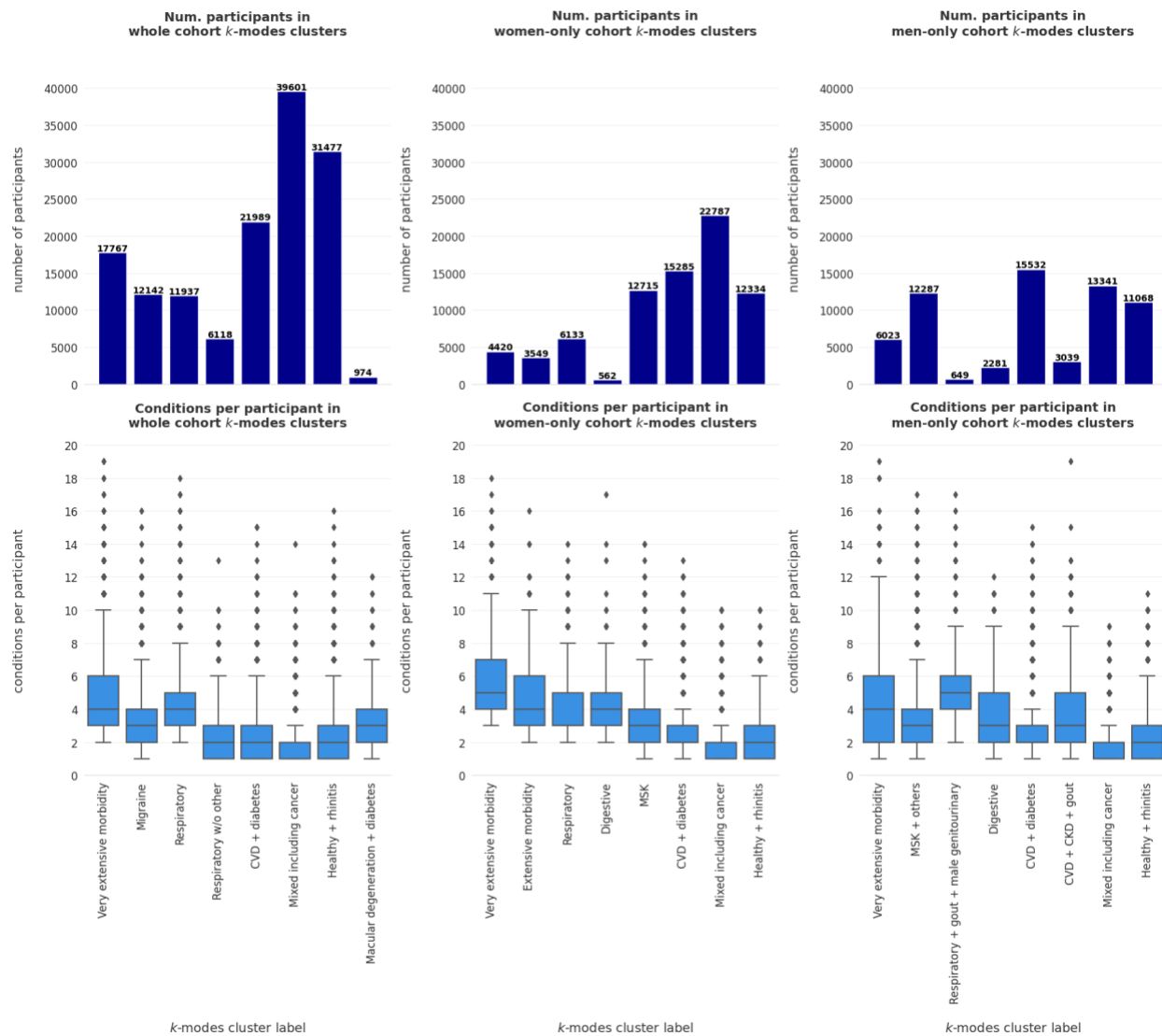
169 Of the remaining clusters, there were some similarities across all three cohorts (*e.g. Respiratory*
170 clusters). Generally, clusters with more participants also tended to have fewer conditions per participant
171 (Fig. 2). For example, the *Mixed including cancer* clusters had the lowest mean conditions per participant
172 (whole: 1.77; women: 1.75; men: 1.62). Such clusters may serve as “miscellaneous” categories for
173 participants with condition profiles that are not easily grouped and/or people with one dominant
174 condition. However, there were also differences. For example, there were clusters which only appeared in
175 the whole population (*Migraine*) and clusters which only appeared in the sex-stratified cohorts (*Digestive*
176 and *MSK* clusters).

177 **Fig. 1. Bubble heatmap shows under- and over-represented conditions in each cluster from the k -**
 178 **modes derived models**



179 Magnitude (adjusted ARF values (*Suppl. Appendix 2*)) is an ordinal representation of the level of over- or under-representation
 180 (e.g., a red bubble of magnitude 2.0 indicates that a condition is more overrepresented than a red bubble of magnitude 1.5).
 181

182 **Fig. 2. Cluster sizes and condition counts.**



183

184

185

186 Subsequent incident depression per identified cluster

187 Analysis of time to incident depression diagnosis included 141,001 participants (73,036 women
188 and 67,920 men), excluding 30,770 participants with a history of depression at baseline (20,592 women
189 and 10,178 men). In addition to participants included in the clustering analysis, this analysis also included
190 30,551 participants with no physical conditions at baseline (16,238 women and 14,313 men)
191 (Supplementary Fig. 1). During an average follow-up of 6.8 years, 5,904 (4.2%) participants, including
192 3,574 (4.9%) women and 2,330 (3.4%) men, had a new depression diagnosis. Generally, participants with
193 physical conditions at baseline had a higher rate of subsequent depression than participants with no
194 physical conditions at baseline (Table 2, Supplementary Fig. 3).

195 There were several consistencies across cohorts. The *Very extensive morbidity* clusters were the
196 most strongly associated with depression in all three cohorts (whole: HR 2.42, 95% CI 2.17-2.69; women:
197 HR 2.67, 95% CI 2.24-3.17; men: HR 2.65, 95% CI 2.22-3.18). Additionally, the *Healthy + rhinitis*
198 (whole: HR 1.59, 95% CI 1.46-1.75; women: HR 1.48, 95% CI 1.30-1.67; men: HR 1.50, 95% CI 1.29-
199 1.75) and *Mixed including cancer* (whole: HR 1.62, 95% CI 1.48-1.77; women: HR 1.63, 95% CI 1.46-
200 1.82; men: HR 1.60, 95% CI 1.38-1.86) clusters were generally the most weakly associated. Finally,
201 association with depression also appeared to increase with the number of conditions per participant (Fig.
202 3). However, there were some exceptions; for example, the whole cohort's *Macular degeneration +*
203 *diabetes* cluster had the fourth highest mean number of conditions (3.09), but was only weakly associated
204 with depression (HR 1.29, 95% CI 0.85-1.98).

205

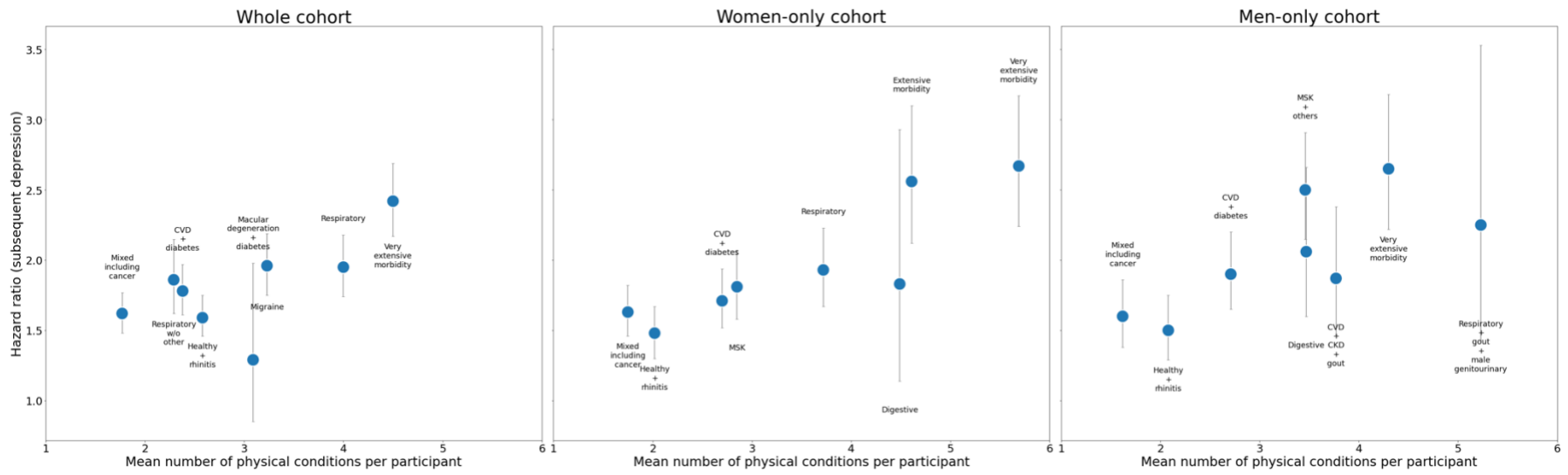
206

207 **Table 2. Hazard ratios per cluster for the development of subsequent depression.**

Cohort	Cluster Label	No. (%) of participants in cluster	Mean no. of conditions per participant in cluster	Hazard ratio (95% CI)
Whole cohort	Very extensive morbidity	17767 (12.5)	4.50	2.42 (2.17, 2.69)
	Migraine	12142 (8.6)	3.23	1.96 (1.75, 2.19)
	Respiratory	11937 (8.4)	4.00	1.95 (1.74, 2.18)
	Respiratory w/o other	6118 (4.3)	2.29	1.86 (1.62, 2.15)
	CVD + diabetes	21989 (15.5)	2.38	1.78 (1.61, 1.97)
	Mixed including cancer	39601 (27.9)	1.77	1.62 (1.48, 1.77)
	Healthy + rhinitis	31477 (22.2)	2.58	1.59 (1.46, 1.75)
	Macular degeneration + diabetes	974 (0.7)	3.09	1.29 (0.85, 1.98)
Women-only	Very extensive morbidity	4420 (5.7)	5.69	2.67 (2.24, 3.17)
	Extensive morbidity	3549 (4.6)	4.61	2.56 (2.12, 3.10)
	Respiratory	6133 (7.9)	3.72	1.93 (1.67, 2.23)
	Digestive	562 (0.7)	4.49	1.83 (1.14, 2.93)
	MSK	12715 (16.4)	2.85	1.81 (1.58, 2.07)
	CVD + diabetes	15285 (19.7)	2.70	1.71 (1.52, 1.94)
	Mixed including cancer	22787 (29.3)	1.75	1.63 (1.46, 1.82)
	Healthy + rhinitis	12334 (15.9)	2.02	1.48 (1.30, 1.67)
Men-only	Very extensive morbidity	6023 (9.4)	4.30	2.65 (2.22, 3.18)
	MSK + others	12287 (19.1)	3.46	2.50 (2.15, 2.91)
	Respiratory + gout + male genitourinary	649 (1.0)	5.23	2.25 (1.43, 3.53)
	Digestive	2281 (3.6)	3.47	2.06 (1.60, 2.66)
	CVD + diabetes	15532 (24.2)	2.71	1.90 (1.65, 2.20)
	CVD + CKD + gout	3039 (4.7)	3.77	1.87 (1.47, 2.38)
	Mixed including cancer	13341 (20.8)	1.62	1.60 (1.38, 1.86)
	Healthy + rhinitis	11068 (17.2)	2.08	1.50 (1.29, 1.75)

208 All models were adjusted for baseline age, ethnicity, country of residence, and deprivation.

209 **Fig. 3. Risk of subsequent depression by mean number of physical conditions.**



210

211 Each point represents a cluster value. Error bars represent a 95% confidence interval.

212 Discussion

213 Summary of findings

214 This study systematically explored clustering of physical health conditions using four methods
215 appropriate for binary data (*k*-modes^{35,36}, *k*-medoids²³, Latent Class Analysis²⁴, and agglomerative
216 hierarchical clustering (AHC)²²). *K*-modes performed best, and the clusters identified were reasonably
217 interpretable and often aligned with known associations between conditions. People with any physical
218 condition at baseline were generally more likely to develop depression than people without any physical
219 condition. There was some variation in this association by cluster which may be at least partly driven by
220 differences in the mean number of physical conditions in each cluster.

221 Comparison with other studies

222 Existing studies of morbidity clustering typically apply a single method. One study compared LCA to a
223 Bayesian, network-based approach, but used age and admission type, rather than conditions alone, to
224 drive cluster formation⁵³. Two other studies explored AHC and *k*-means in the same dataset, but chose *k*-
225 means on the basis of AHC being too computationally intensive rather than based on performance^{27,28}.
226 Additionally, despite the use of *k*-means⁵⁴ by several multimorbidity studies²⁷⁻³⁰, it typically relies upon
227 Euclidean distance as its similarity measure³¹, which is unsuitable for binary data¹⁸. Other
228 multimorbidity studies have used *k*-means *after* a Multiple Correspondence Analysis^{55,56}, which
229 represents categorical features as a low-dimensional Euclidean space^{29,30}. While this transforms the data
230 features into an appropriate format for *k*-means, it also manipulates the data based on their pairwise co-
231 occurrences, which may not be appropriate for every dataset.

232 This study finds that almost all physical morbidity clusters are associated with higher risk of subsequent
233 depression than the group with no physical conditions at baseline. Although the strength of association
234 varied by cluster, this seemed to be partly explained by the mean number of conditions in the cluster. This
235 is consistent with a similar study²⁹ which found associations between severe mental illness and a higher

236 number of physical conditions. Another similar study, which aimed to identify groups of physical
237 conditions associated with incident depression within a Taiwanese cohort, also found that social factors
238 played a role on the risk of subsequent depression diagnosis ²¹. Specifically, they found that amongst four
239 *Cardiometabolic*, *Arthritis-cataract*, *Multimorbidity*, and *Relatively healthy* clusters, those within the
240 *Arthritis-cataract* and *Multimorbidity* clusters had significantly higher risk of depression than healthy
241 individuals. However, this association was attenuated for participants who engaged in social activities,
242 including a job, volunteer experience, or community activities ²¹.

243 **Strengths and limitations**

244 Strengths of this study include the analysis of a large dataset which records morbidities in both baseline
245 research data and linked routine data, as well as the inclusion of a wide set of morbidities recommended
246 by a recent consensus study ⁵⁷. Notably, this study is unique for its implementation and comparison
247 between four clustering methods appropriate for binary data. A limitation is that the data are collected
248 from volunteers who are generally more affluent than the UK average, and people from ethnic minorities
249 are somewhat under-represented ⁵⁸. Additionally, there is no standard way to evaluate the validity of
250 identified clusters, although the observed clusters do include several known clinical associations.
251 Consequently, this warrants further validation studies in other datasets to explore reproducibility of
252 cluster solutions.

253 **Implications for research**

254 Many previous studies of morbidity clustering do not provide much information about which conditions
255 are over or under-represented in clusters, which leaves readers relying solely on author-chosen cluster
256 labels for interpretation ^{19,21}. For example, it is common for other studies to identify a ‘cardiometabolic’
257 cluster ²¹ and the *CVD + diabetes* cluster in our study was amongst the three largest clusters in all three
258 cohorts. However, it is not straightforward to compare clusters across studies because of considerable
259 variation in the conditions included in analysis, and because many clustering studies do not provide

260 detailed information about the nature of identified clusters. Key implications are that clustering studies
261 should be more consistent in the choice of conditions to include (and, at a minimum, follow consensus
262 recommendations⁵⁷). Additionally, they should report the nature of clusters to help understand them
263 beyond their high-level labels (by, for example, visualizing the prevalence of individual conditions in
264 each cluster alongside over/under representation). We believe that our Adjusted Relative Frequency
265 (ARF) measure with visualization in a bubble heatmap demonstrates one way to do this. However, there
266 is a need for multimorbidity researchers to develop improved and consistent cluster visualizations and
267 explanations to facilitate interpretation and to enhance clinical utility.

268 Morbidity clustering studies also typically use one clustering method, but there is no single clustering
269 method which is likely to be optimal for every dataset. We therefore believe that clustering studies should
270 more systematically explore different methods and make explicit how they choose the best method for
271 their datasets and purposes. To encourage similar systematic comparison of different cluster methods, we
272 have provided access to our code (<https://github.com/laurendelong21/clusterMed>). Many studies also
273 cluster the entire population which is likely not sensible given the very different incidence and prevalence
274 of disease with age and, to a lesser extent, sex and ethnicity. In this analysis, study participants were
275 mostly middle-aged (so we did not further stratify by age) and overwhelmingly white, but we found some
276 differences in the clusters identified in the whole population versus women or men separately. Although
277 whole population clustering may be appropriate in some circumstances, reporting of clusters stratified by
278 age and sex (and ethnicity if the data permits) would be valuable to explore how clustering varies by
279 demographic characteristics.

280 Finally, further research to better understand why physical multimorbidity is associated with subsequent
281 depression is needed. The general trend between increased risk of subsequent depression and mean
282 number of conditions suggests a social explanation: suffering more conditions may more strongly
283 interrupt one's life or sense of self. However, the relationship between depression and physical conditions

284 is very likely bidirectional and longitudinal research which better examines how the two interact over a
285 lifetime would be valuable.

286 Conclusions

287 Using the best performing of four different clustering methods, this study identified several
288 multimorbidity clusters which align with known clinical associations. Association with depression varied
289 between clusters, but this may be partly driven by differences in the number of conditions. More research
290 is needed to better understand the mechanisms underlying such associations.

291

292 Data Availability

293 The UK Biobank data is not openly available to protect the rights of participants. Researchers can register
294 for access here: <https://www.ukbiobank.ac.uk/enable-your-research/register> .

295 Code Availability

296 Corresponding code is available at <https://github.com/laurendelong21/clusterMed> .

297 Competing Interests

298 The authors declare no competing interests.

299 Author Contributions

300 All authors contributed to writing the manuscript. L.N.D. and K.F. conducted the analyses and made the
301 figures. L.N.D. and P.G. wrote the software. K.F. and R.P. processed and prepared the data. J.D.F. and
302 B.G. designed and supervised the study.

303 Acknowledgments

304 This work was co-funded by the Medical Research Council and the National Institute for Health Research
305 (grant number MC/S028013), and the NIHR AIM-CISC programme (grant number NIHR202639). The
306 views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of
307 Health and Social Care. The study was conducted using the UK Biobank Resource under application
308 number 57213.

309 LND is individually funded by a Global Informatics Scholarship from the School of Informatics at the
310 University of Edinburgh. The School of Informatics had no role in study design, data collection and
311 analysis, decision to publish, or preparation of the manuscript.

312 The authors would like to thank the UK Biobank participants and the UK Biobank staff for their
313 contributions to this study. The authors would like to thank the public members of our advisory board, Dr
314 Paul Kelly and Pat Watson, for providing thoughtful feedback throughout our project. This work has
315 made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF)
316 (<http://www.ecdf.ed.ac.uk/>).

317

318 References

- 319 1. Skou, S. T. *et al.* Multimorbidity. *Nat Rev Dis Primers* **8**, 1–22 (2022).
- 320 2. Harrison, C. *et al.* Comorbidity versus multimorbidity: Why it matters. *Journal of Multimorbidity*
321 *and Comorbidity* vol. 11 2633556521993993 Preprint at (2021).
- 322 3. Fortin, M., Stewart, M., Poitras, M.-E., Almirall, J. & Maddocks, H. A Systematic Review of
323 Prevalence Studies on Multimorbidity: Toward a More Uniform Methodology. *The Annals of*
324 *Family Medicine* **10**, 142–151 (2012).
- 325 4. Barnett, K. *et al.* Epidemiology of multimorbidity and implications for health care, research, and
326 medical education: a cross-sectional study. *The Lancet* **380**, 37–43 (2012).
- 327 5. Swain, S., Sarmanova, A., Coupland, C., Doherty, M. & Zhang, W. Comorbidities in
328 Osteoarthritis: A systematic review and meta-analysis of observational studies. *Arthritis Care Res*
329 *(Hoboken)* **72**, 991–1000 (2020).
- 330 6. Alexander, K. P. *et al.* Outcomes of apixaban versus warfarin in patients with atrial fibrillation and
331 multi-morbidity: Insights from the ARISTOTLE trial. *Am Heart J* **208**, 123–131 (2019).
- 332 7. Marrie, R. A. Comorbidity in multiple sclerosis: Past, present and future. *Clinical and*
333 *Investigative Medicine* **42**, E5–E12 (2019).
- 334 8. Pitsillou, E. *et al.* The cellular and molecular basis of major depressive disorder: towards a unified
335 model for understanding clinical depression. *Mol Biol Rep* **47**, 753–770 (2020).
- 336 9. Kraus, C., Kadriu, B., Lanzenberger, R., Zarate Jr, C. A. & Kasper, S. Prognosis and improved
337 outcomes in major depression: a review. *Transl Psychiatry* **9**, 1–17 (2019).
- 338 10. Malhi, G. S. & Mann, J. J. Depression. *The Lancet* **392**, 2299–2312 (2018).
- 339 11. Goodwin, G. M. The overlap between anxiety, depression, and obsessive-compulsive disorder.
340 *Dialogues Clin Neurosci* (2022).
- 341 12. Rao, S. & Broadbear, J. Borderline personality disorder and depressive disorder. *Australasian*
342 *Psychiatry* **27**, 573–577 (2019).
- 343 13. Gold, S. M. *et al.* Comorbid depression in medical diseases. *Nat Rev Dis Primers* **6**, 1–22 (2020).
- 344 14. Shao, M. *et al.* Depression and cardiovascular disease: Shared molecular mechanisms and clinical
345 implications. *Psychiatry Res* **285**, 112802 (2020).
- 346 15. Riemer, F. *et al.* Microstructural changes precede depression in patients with relapsing-remitting
347 Multiple Sclerosis. *Communications Medicine* **3**, 90 (2023).
- 348 16. Marrie, R. A., Graff, L. A., Fisk, J. D., Patten, S. B. & Bernstein, C. N. The relationship between
349 symptoms of depression and anxiety and disease activity in IBD over time. *Inflamm Bowel Dis* **27**,
350 1285–1293 (2021).

- 351 17. Davyson, E. *et al.* Metabolomic Investigation of Major Depressive Disorder Identifies a
352 Potentially Causal Association With Polyunsaturated Fatty Acids. *Biol Psychiatry* **94**, 630–639
353 (2023).
- 354 18. Cornell, J. E. *et al.* Multimorbidity Clusters: Clustering Binary Data From Multimorbidity
355 Clusters: Clustering Binary Data From a Large Administrative Medical Database. *Applied*
356 *Multivariate Research* **12**, 163 (2009).
- 357 19. Bisquera, A. *et al.* Identifying longitudinal clusters of multimorbidity in an urban setting: A
358 population-based cross-sectional study. *The Lancet Regional Health - Europe* **3**, 100047 (2021).
- 359 20. Robertson, L. *et al.* Identifying multimorbidity clusters in an unselected population of hospitalised
360 patients. *Sci Rep* **12**, 5134 (2022).
- 361 21. Ho, H.-E., Yeh, C.-J., Cheng-Chung Wei, J., Chu, W.-M. & Lee, M.-C. Association between
362 multimorbidity patterns and incident depression among older adults in Taiwan: the role of social
363 participation. *BMC Geriatr* **23**, 177 (2023).
- 364 22. Sasirekha, K. & Baby, P. Agglomerative hierarchical clustering algorithm-a. *International Journal*
365 *of Scientific and Research Publications* **83**, 83 (2013).
- 366 23. Jin Xin and Han, J. K-Medoids Clustering. *Encyclopedia of Machine Learning* 564–565 (2010)
367 doi:10.1007/978-0-387-30164-8_426.
- 368 24. Weller, B. E., Bowen, N. K. & Faubert, S. J. Latent class analysis: a guide to best practice. *Journal*
369 *of Black Psychology* **46**, 287–311 (2020).
- 370 25. Hall, M. *et al.* Multimorbidity and survival for patients with acute myocardial infarction in
371 England and Wales: Latent class analysis of a nationwide population-based cohort. *PLoS Med* **15**,
372 e1002501 (2018).
- 373 26. Eto, F. *et al.* Ethnic differences in early onset multimorbidity and associations with health service
374 use, long-term prescribing, years of life lost, and mortality: A cross-sectional study using
375 clustering in the UK Clinical Practice Research Datalink. *PLoS Med* **20**, e1004300 (2023).
- 376 27. Ioakeim-Skoufa, I. *et al.* Multimorbidity Clusters in the Oldest Old: Results from the EpiChron
377 Cohort. *Int J Environ Res Public Health* **19**, 10180 (2022).
- 378 28. Carmona-Pérez, J. *et al.* Multimorbidity clusters in patients with chronic obstructive airway
379 diseases in the EpiChron Cohort. *Sci Rep* **11**, 4784 (2021).
- 380 29. Lauenders, N., Hayes, J. F., Price, G. & Osborn, D. P. Clustering of physical health multimorbidity
381 in people with severe mental illness: An accumulated prevalence analysis of United Kingdom
382 primary care data. *PLoS Med* **19**, e1003976 (2022).
- 383 30. Guisado-Clavero, M. *et al.* Multimorbidity patterns in the elderly: a prospective cohort study with
384 cluster analysis. *BMC Geriatr* **18**, 16 (2018).
- 385 31. Sinaga, K. P. & Yang, M.-S. Unsupervised K-means clustering algorithm. *IEEE access* **8**, 80716–
386 80727 (2020).
- 387 32. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range
388 of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).

- 389 33. Ho, I. S. S. *et al.* Measuring multimorbidity in research: Delphi consensus study. *BMJ Medicine* **1**,
390 e000247 (2022).
- 391 34. Prigge, R. *et al.* Robustly Measuring Multiple Long-Term Health Conditions Using Disparate
392 Linked Datasets in UK Biobank. *Preprints with The Lancet* (2024).
- 393 35. Huang, Z. Clustering large data sets with mixed numeric and categorical values. in *Proceedings of*
394 *the 1st pacific-asia conference on knowledge discovery and data mining, (PAKDD)* 21–34 (1997).
- 395 36. Cao, F., Liang, J. & Bai, L. A new initialization method for categorical data clustering. *Expert Syst*
396 *Appl* **36**, 10223–10228 (2009).
- 397 37. Abad-D\`viez, J. M. *et al.* Age and gender differences in the prevalence and patterns of
398 multimorbidity in the older population. *BMC Geriatr* **14**, 1–8 (2014).
- 399 38. Agur, K., McLean, G., Hunt, K., Guthrie, B. & Mercer, S. W. How does sex influence
400 multimorbidity? Secondary analysis of a large nationally representative dataset. *Int J Environ Res*
401 *Public Health* **13**, 391 (2016).
- 402 39. Robert, L. Thorndike. “Who Belongs in the Family?”. *Psychometrika* **18**, 267–276 (1953).
- 403 40. Schwarz, G. Estimating the dimension of a model. *The annals of statistics* 461–464 (1978).
- 404 41. Hamming, R. W. Entropy and Shannon’s First Theorem. *Coding and information*
405 *theory.*(Prentice-Hall Inc. Englewood Cliffs, New Jersey) **107**, (1980).
- 406 42. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-*
407 *theory and Methods* **3**, 1–27 (1974).
- 408 43. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans Pattern Anal Mach*
409 *Intell* **2**, 224–227 (1979).
- 410 44. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster
411 analysis. *J Comput Appl Math* **20**, 53–65 (1987).
- 412 45. Huang, Z. A fast clustering algorithm to cluster very large categorical data sets in data mining.
413 *Data Min Knowl Discov* **3**, 34–39 (1997).
- 414 46. Dunn, O. J. Multiple comparisons among means. *J Am Stat Assoc* **56**, 52–64 (1961).
- 415 47. Bland, J. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *Bmj* **310**, 170
416 (1995).
- 417 48. Guide, P. Fisher’s Exact Test. Preprint at
418 https://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/.
- 419 49. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*
420 *(Methodological)* **34**, 187–202 (1972).
- 421 50. Satagopan, J. M. *et al.* A note on competing risks in survival data analysis. *Br J Cancer* **91**, 1229–
422 1235 (2004).
- 423 51. Khunti, K., Routen, A., Banerjee, A. & Pareek, M. The need for improved collection and coding
424 of ethnicity in health research. *J Public Health (Bangkok)* **43**, e270–e272 (2021).

- 425 52. Townsend, P. Deprivation. *J Soc Policy* **16**, 125–146 (1987).
- 426 53. Restocchi, V., Villegas, J. G. & Fleuriot, J. D. Multimorbidity profiles and stochastic block
427 modeling improve ICU patient clustering. in *2022 22nd IEEE International Symposium on*
428 *Cluster, Cloud and Internet Computing (CCGrid)* 925–932 (IEEE, 2022).
429 doi:10.1109/CCGrid54584.2022.00112.
- 430 54. MacQueen, J. Classification and analysis of multivariate observations. in *5th Berkeley Symp.*
431 *Math. Statist. Probability* 281–297 (1967).
- 432 55. Abdi, H. & Dominique Valentin. Multiple correspondence analysis. *Encyclopedia of measurement*
433 *and statistics* **2**, 651–657 (2007).
- 434 56. Beaney, T. *et al.* Identifying multi-resolution clusters of diseases in ten million patients with
435 multimorbidity in primary care in England. *Communications Medicine* **4**, 102 (2024).
- 436 57. Ho, I. S. S. *et al.* Measuring multimorbidity in research: Delphi consensus study. *BMJ Medicine* **1**,
437 e000247 (2022).
- 438 58. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK
439 Biobank Participants With Those of the General Population. *Am J Epidemiol* **186**, 1026–1034
440 (2017).
- 441

1 Supporting information

2 Supplementary Table 1. 69 physical conditions with corresponding bodily systems.

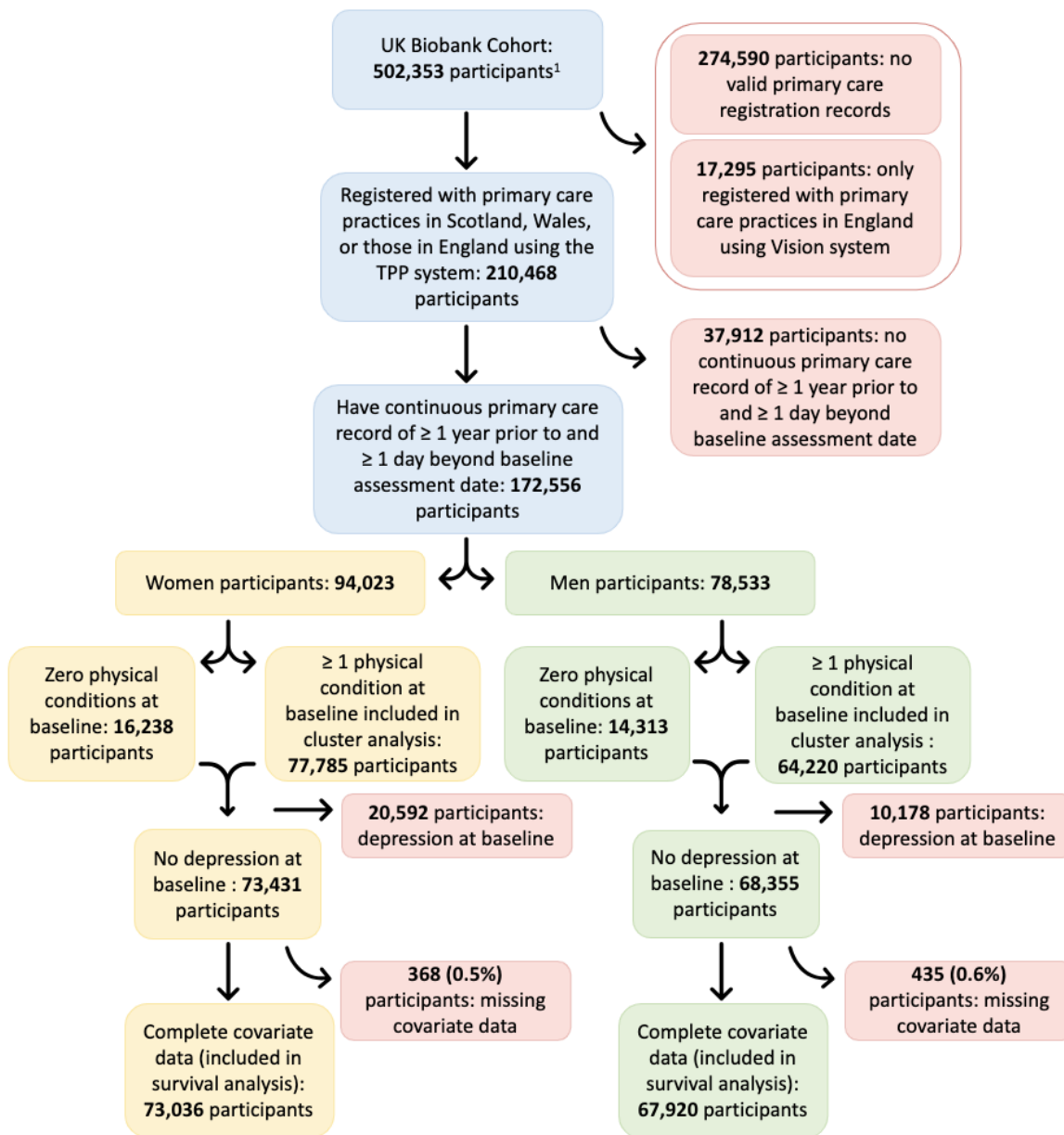
System	Condition
Skin conditions	Psoriasis
Perinatal conditions	Down's syndrome
Neurological conditions	Migraine Peripheral or autonomic neuropathy Epilepsy Cerebral Palsy Motor neuron disease Multiple sclerosis Myasthenia gravis Parkinson's disease Postviral fatigue syndrome, neurasthenia and fibromyalgia
Musculoskeletal conditions	Inflammatory arthritis and other inflammatory conditions Gout Osteoporosis and vertebral crush fractures Osteoarthritis (excl spine) Spinal stenosis
Mental Health Disorders	Dementia
Infectious Diseases	Tuberculosis
Infectious Diseases	HIV
Haematological/Immunological conditions	Sarcoidosis Iron and vitamin deficiency anaemia Immunodeficiencies Sickle-cell anaemia Thalassaemia
Diseases of the Respiratory System	Sleep apnoea COPD Bronchiectasis Asthma Asbestosis Allergic and chronic rhinitis
Diseases of the Genitourinary system	Chronic renal disease Urinary Incontinence Erectile dysfunction Non-acute cystitis Hyperplasia of prostate
Diseases of the Eye	Visual impairment and blindness Macular degeneration Glaucoma
Diseases of the Endocrine System	Hypo or hyperthyroidism Addisons disease Cystic Fibrosis Diabetes NOS

	Type 1 diabetes Type 2 diabetes
Diseases of the Ear	Meniere disease Hearing loss
Diseases of the Digestive System	Fatty Liver Chronic liver disease Peptic ulcer disease Irritable bowel syndrome Diverticular disease of intestine (acute and chronic) Gastro-oesophageal reflux, gastritis and similar Coeliac disease Inflammatory bowel disease
Diseases of the Circulatory System	Conduction disorders and other arrhythmias Coronary heart disease Cardiomyopathy Heart valve disorders Atrial fibrillation Transient ischaemic attack Peripheral arterial disease Hypertension Heart failure Primary pulmonary hypertension Stroke
Cancers	Solid organ malignancies Haematological malignancies Non-melanoma skin malignancies
Benign Neoplasm/CIN	Benign neoplasm of brain and other parts of central nervous system

3
4
5
6
7
8
9
10
11

12 **Supplementary Figure 1. Flow diagram explaining data filtration steps.**

13



1. Excluding participants who withdrew permission for their data to be included in research before 13 October 2023

14

15

16

17 Supplementary Appendix I. Methodological Descriptions.

18 To identify a clustering method best suited for our data, we explored several combinations of
19 metrics and methods which were, in theory, capable of handling the binary nature of the morbidity
20 data. Specifically, we used the following metrics within this study:

- 21 • **Hamming distance**¹: This is a dissimilarity metric denoting the number of mismatching
22 categories between two objects. It is formally defined as²:

$$23 \quad d_{\text{hamming}}(X, Y) = \sum_{j=1}^m \delta(x_j, y_j),$$

24 where:

$$25 \quad \delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_j = y_j \\ 1 & \text{if } x_j \neq y_j \end{cases}$$

- 26 • **cosine similarity**³: This is a measure which was originally used to indicate how similar two
27 vector angles are to one another. For binary data, it can be understood as the number of true,
28 or positive features are shared between two objects, divided by the product of the number of
29 true objects from each object. Formally, it is defined as⁴:

$$30 \quad s_{\text{cosine}}(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|}$$

31

32 With these metrics, we explored the following four clustering methods:

- 33 • **k-modes**^{5,6}: This uses the same methodology as *k*-means clustering⁷, but the centroid of each
34 cluster is defined on the number of matching categories between data points, computed via
35 the Hamming distance¹. In other words, the centroid represents the mode of the cluster, rather
36 than the mean^{5,6}. Centroid initialization was performed via the frequency-based *Huang*
37 metric⁵.
- 38 • **k-medoids**⁸: As above, this uses the same methodology as *k*-means clustering⁷, but the
39 centroid is an actual data point acting as the “median”, computed via cosine similarity.
- 40 • **Latent Class Analysis (LCA)**⁹: This aims to find groups or subtypes of cases (latent classes)
41 in multivariate categorical data. It gives probabilities of class membership, rather than
42 concrete class assignments, which are unique, so the user can see the likelihood that a data
43 point truly belongs to its assigned class⁹.
- 44 • **agglomerative hierarchical clustering (AHC)**¹⁰: This is best understood as a “bottom-up”
45 approach in which samples start out alone, then merge to form larger and larger clusters¹¹.
46 We used a *complete* linkage (the maximum distance between points in two clusters¹²),
47 computed via Hamming distance¹.

48

49 Finally, we assessed cluster performance, including separation and overlap, with the following three
50 performance metrics:

- 51 • **Calinski and Harabasz score**¹³: This is the ratio of between-cluster dispersion to within-
52 cluster dispersion. A higher Calinski and Harabasz score indicates better performance.
- 53 • **Davies Bouldin score**¹⁴: This is a measure of cluster similarity to each cluster’s most similar
54 cluster. A Davies Bouldin score closer to zero indicates better performance.
- 55 • **Silhouette score**¹⁵: This is a measure of cluster fit which accounts for the mean distance
56 between points in each individual cluster as well as the mean distance to points in the closest
57 neighboring cluster. A silhouette score closer to one indicates better performance. Hamming
58 distance was utilized as the distance metric.

59

60 The best similarity or dissimilarity metrics were selected for *k*-modes, *k*-medoids, and AHC by testing
61 each of them upon a random selection of participants (1,417). The *k*-modes method used with an
62 alternative initialization technique, called the *Cao* metric ⁶, resulted in some clusters containing less
63 than ten participants, while others contained hundreds. Such imbalance is uninformative for our
64 purposes. Similarly, *k*-medoids with Hamming distance and *Jaccard similarity* ¹⁶, another similarity
65 metric, resulted in several empty clusters. Finally, AHC with other metrics and linkage types resulted
66 in poor separation between clusters, with high overlap between branches. These issues were not
67 present with the specified metrics.

68

69

70 **Supplementary Appendix 2. Bubble Heatmap.**

71 The *bubble heatmap* places ARF values on a grid in which the y -axis contains conditions, the x -axis
72 contains clusters, and data points are colored blue (under-representation) or red (over-representation)
73 at each intersection. The magnitude of under- or over-representation is indicated by the size of the
74 data point, or *bubble*. Points are not statistically significant, as determined by the Fisher's Exact test
75 (REF), were omitted. Therefore, conditions with no significant values are omitted entirely from the y -
76 axis.

77 Notably, for visualisation purposes, the ARF values are adjusted so that values denoting under-
78 representation (between zero and one) were mapped to a similar scale as those denoting over-
79 representation (values greater than one). Specifically, we used the following function, in which x
80 denotes the original ARF value:

81
$$f(x) = \frac{2(x - 1)}{(x + 1)}$$

82

83

84 **Supplementary Table 2. ARF values and adjusted p-values per condition and cluster.**

85

86 Statistically significant values ($p < 0.05$) are denoted in bold.

87 See corresponding excel file (S2Table.xlsx).

88

89

90 **Supplementary Table 3.** Cluster labels for the *whole* cohort.

Cluster Label	Num. conditions significantly over-represented	Num. conditions significantly under-represented	Explanation
Very extensive morbidity	41	4	41 conditions are significantly over-represented with no stand-out conditions or groups.
Migraine	7	21	Migraine has large over-representation with relatively small over-representation of six other conditions.
Respiratory	8	19	Four respiratory conditions have large over-representation, with four other conditions significantly over-represented.
Respiratory w/o other	3	29	Three respiratory conditions are significantly over-represented, with widespread significant under-representation of other conditions.
CVD + diabetes	16	27	Nine cardiovascular disease (CVD) conditions and three diabetes types are significantly over-represented but, unlike ‘very extensive morbidity’, many other conditions are under-represented.
Mixed including cancer	23	21	Twenty-three conditions, including cancers, are significantly over-represented with no stand-out conditions or groups, but CVD and respiratory conditions are significantly under-represented. However, over-represented conditions are relatively rare so this group is relatively healthy.
Healthy + rhinitis	1	41	Allergic and chronic rhinitis is significantly over-represented, with significant under-representation of 41 other conditions
Macular degeneration + diabetes	5	6	Diabetes and eye conditions significantly over-represented, of which macular degeneration has the highest prevalence.

91

92

93 **Supplementary Table 4.** Cluster labels for the *women-only* cohort.

Cluster Label	No. of conditions significantly over-represented	No. of conditions significantly under-represented	Explanation
Very extensive morbidity	34	0	34 conditions are significantly over-represented with no stand-out conditions or groups.
Extensive morbidity	23	2	23 conditions are significantly over-represented with no stand-out conditions or groups and two relatively small under-representation of two conditions.
Respiratory	4	14	Four respiratory conditions have large over-representation, with no other significantly over-represented conditions.
Digestive	5	3	Four digestive conditions are significantly over-represented, two of which are largely over-represented. Additionally, one CVD condition is over-represented, but the other 11 CVD conditions are not.
MSK	7	11	Five musculoskeletal (MSK) conditions are significantly over-represented, with relatively small over-representation of two other conditions and under-representation of 11 other conditions.
CVD + diabetes	15	18	Nine CVD conditions and three diabetes types are significantly over-represented, but, unlike 'very extensive morbidity', many other conditions are under-represented.
Mixed including cancer	17	18	Seventeen conditions, including cancers, are significantly over-represented with no stand-out conditions or groups, but CVD, respiratory and MSK are significantly under-represented. However, over-represented conditions are relatively rare, so this group is relatively healthy.
Healthy + rhinitis	1	44	Allergic and chronic rhinitis is significantly over-represented, with significant under-representation of 44 other conditions.

94
95

96

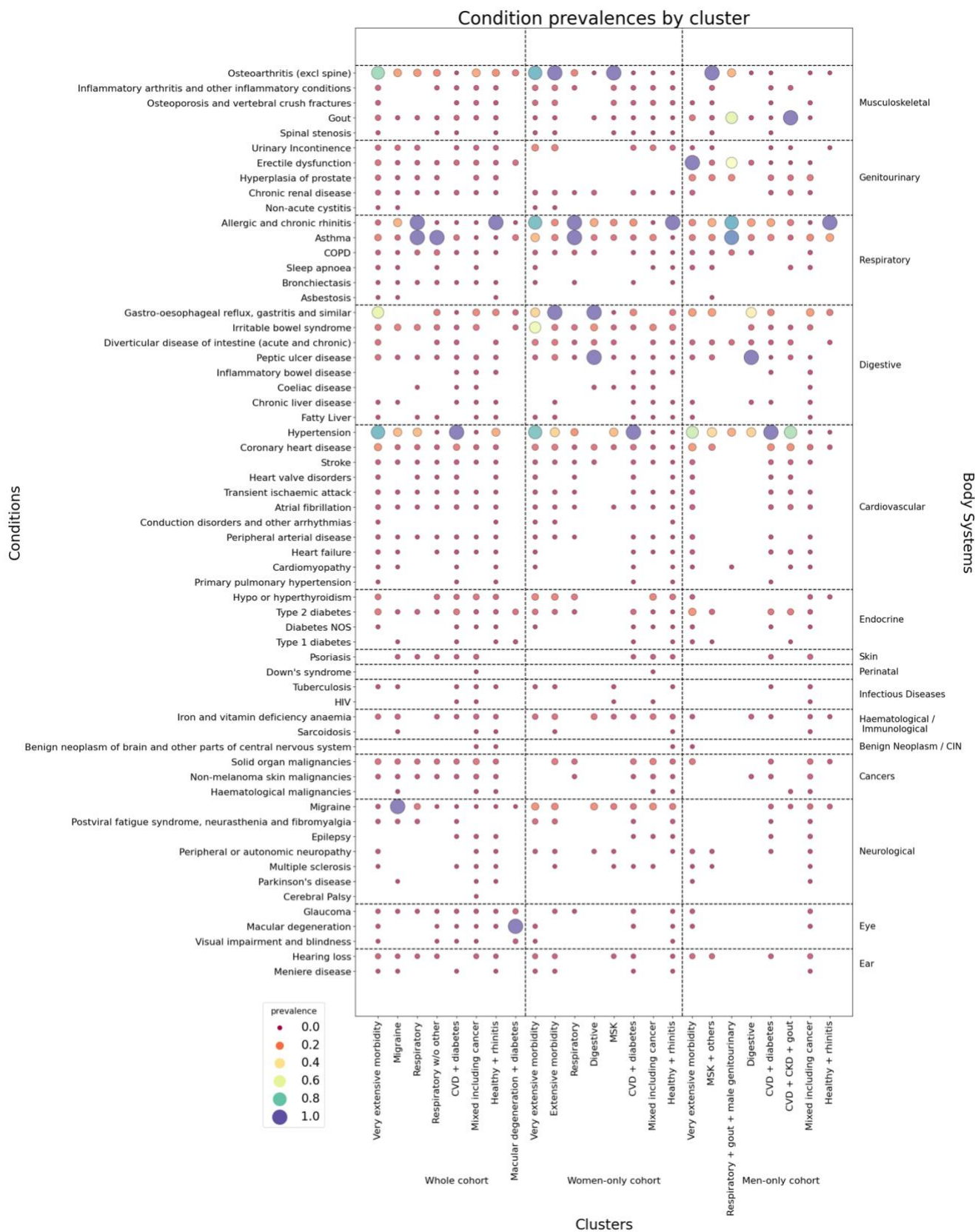
Supplementary Table 5. Cluster labels for the *men-only* cohort.

Cluster Label	No. of conditions significantly over-represented	No. of conditions significantly under-represented	Explanation
Very extensive morbidity	29	3	29 conditions are significantly over-represented with no stand-out conditions or groups. While three conditions have relatively small under-representation, there are no stand-out conditions or groups amongst them either.
MSK + others	14	7	Four musculoskeletal conditions are significantly over represented, but additionally 10 other conditions with no stand out pattern (unlike the Women-only <i>MSK</i> cluster, which is more purely MSK).
Respiratory + gout + male genitourinary	6	1	Three respiratory conditions, gout, and two genitourinary conditions affecting primarily men have large over-representation, while only one condition has relatively small under-representation.
Digestive	6	6	Four digestive conditions have large over-representation, with two other conditions significantly over-represented.
CVD + diabetes	11	20	Eight CVD conditions and two diabetes types are significantly over-represented, in addition to one other condition, but, unlike 'very extensive morbidity', many other conditions are under-represented.
CVD + CKD + gout	11	5	Seven CVD conditions are significantly over-represented. Additionally, gout and chronic renal/kidney disease (CKD) have large over-representation.
Mixed including cancer	19	18	Nineteen conditions, including cancers, are significantly over-represented with no stand-out conditions or groups, but CVD, respiratory and MSK are significantly under-represented. However, the over-represented conditions are relatively rare, so this group is relatively healthy.
Healthy + rhinitis	2	42	Allergic and chronic rhinitis as well as Asthma are significantly over-represented, with significant under-representation of 42 other conditions.

97

98

99 **Supplementary Figure 2. Prevalence values per cluster and condition.**



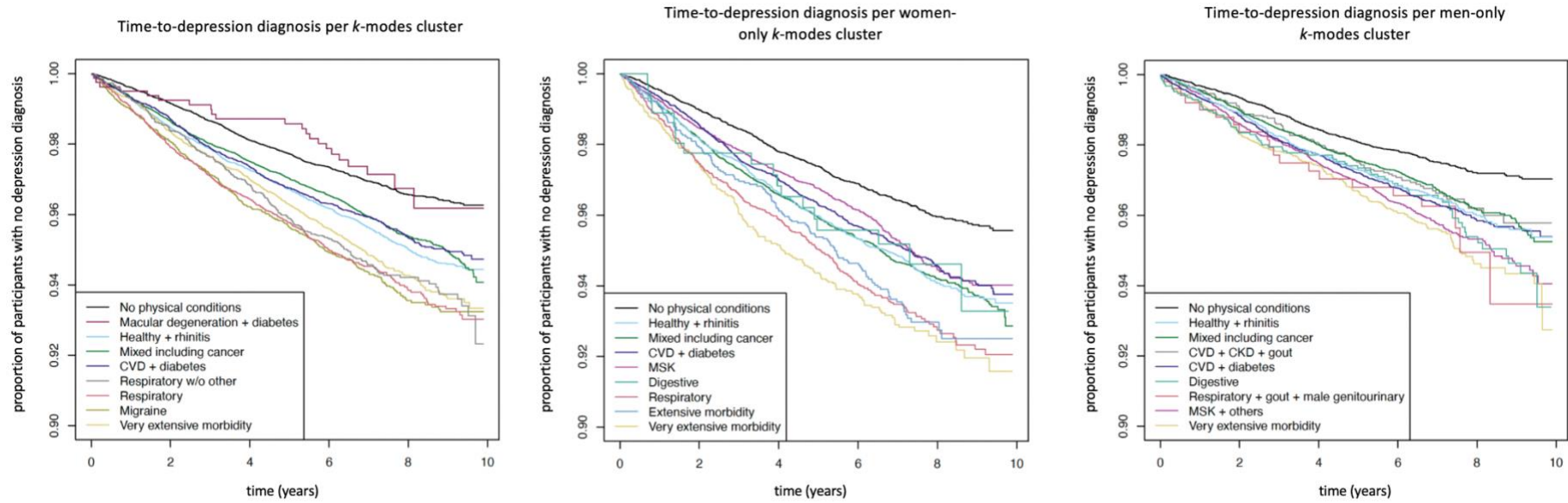
100

101 Corresponds to Fig. 1. Conditions with a prevalence of zero are omitted.

102

103

104 **Supplementary Figure 3. Time-to-depression diagnosis for each cluster in each of the *k*-modes models.**



105

106 All y -axes are cropped for better resolution between curves.

107

108

Supplementary Table 2.a. ARF values for *whole*-cohort clusters.

Condition	Very extensive morbidity	Migraine	Respiratory	Respiratory w/o other	CVD + diabetes	Mixed including cancer	Healthy + rhinitis	Macular degeneration + diabetes
Addisons disease	0.64	1.20	1.62	1.85	1.17	0.81	0.87	0.00
Allergic and chronic rhinitis	0.20	0.78	2.97	0.00	0.00	0.00	2.97	0.00
Asbestosis	2.23	0.43	0.56	0.86	0.95	1.15	0.57	0.77
Asthma	0.86	0.44	6.05	6.05	0.53	0.00	0.00	0.58
Atrial fibrillation	1.78	0.73	0.78	0.56	1.48	0.84	0.70	1.22
Benign neoplasm of brain and other parts of central nervous system	1.21	1.06	0.84	0.76	0.87	1.22	0.80	0.42
Bronchiectasis	1.31	0.48	2.29	2.72	0.71	0.85	0.58	1.37
COPD	1.88	0.55	2.43	3.41	0.87	0.55	0.32	0.95
Cardiomyopathy	1.65	0.51	0.91	0.64	1.57	0.82	0.74	1.35
Cerebral Palsy	1.16	0.54	0.70	0.46	1.10	1.34	0.80	0.96
Chronic liver disease	1.57	0.75	0.90	1.05	0.72	1.20	0.76	0.63
Chronic renal disease	2.09	0.74	0.73	0.39	1.55	0.74	0.64	1.19
Coeliac disease	0.93	1.01	1.30	0.90	0.48	1.30	0.93	0.92
Conduction disorders and other arrhythmias	1.38	1.09	0.95	0.71	1.10	1.08	0.65	1.29
Coronary heart disease	2.26	0.69	0.75	0.52	1.58	0.67	0.60	1.13
Cystic Fibrosis	0.99	0.84	0.98	0.72	0.67	1.26	1.02	1.50
Dementia	1.39	0.60	1.09	0.71	1.05	1.32	0.55	0.00
Diabetes NOS	1.78	0.78	0.76	0.78	1.71	0.72	0.60	1.91
Diverticular disease of intestine (acute and chronic)	1.86	0.93	0.93	0.78	0.81	0.98	0.77	1.15
Down's syndrome	0.67	0.49	0.00	1.93	0.27	2.09	0.56	6.07
Epilepsy	0.93	1.09	0.95	0.78	0.76	1.37	0.77	0.89

Erectile dysfunction	1.67	0.54	0.70	0.59	1.24	1.04	0.77	1.47
Fatty Liver	2.15	0.87	1.40	0.58	1.09	0.59	0.82	0.17
Gastro-oesophageal reflux, gastritis and similar	3.23	0.96	0.96	0.64	0.00	0.86	0.74	0.41
Glaucoma	1.25	0.79	0.80	0.66	1.16	1.13	0.75	2.97
Gout	1.94	0.48	0.72	0.53	1.50	0.87	0.68	1.02
HIV	0.61	0.85	1.16	0.56	0.60	1.45	1.00	1.18
Haematological malignancies	1.17	0.77	0.79	0.72	1.02	1.32	0.72	0.77
Hearing loss	1.50	0.82	0.86	0.80	0.94	1.15	0.74	1.01
Heart failure	2.34	0.56	0.79	0.48	1.74	0.60	0.56	1.50
Heart valve disorders	1.51	1.06	0.79	0.52	1.28	0.96	0.71	1.25
Hyperplasia of prostate	1.54	0.56	0.74	0.69	1.05	1.15	0.80	1.16
Hypertension	2.32	0.76	0.77	0.00	2.72	0.00	0.69	0.90
Hypo or hyperthyroidism	1.11	1.01	0.96	0.83	0.83	1.22	0.82	1.07
Immunodeficiencies	1.39	1.16	0.98	0.77	0.64	1.13	0.82	2.41
Inflammatory arthritis and other inflammatory conditions	1.69	1.00	0.90	0.69	0.71	1.11	0.79	0.74
Inflammatory bowel disease	1.05	0.87	0.98	0.88	0.68	1.31	0.89	0.60
Iron and vitamin deficiency anaemia	1.26	1.22	0.94	0.76	0.67	1.22	0.79	0.89
Irritable bowel syndrome	1.15	1.45	1.17	0.88	0.44	1.10	0.97	0.62
Macular degeneration	1.33	0.86	0.85	0.00	0.00	0.00	0.80	74.74
Meniere disease	1.40	1.29	0.83	0.73	0.71	1.11	0.85	0.94
Migraine	0.29	10.17	1.12	0.00	0.00	0.00	0.00	0.00
Motor neuron disease	1.51	1.26	0.32	0.63	0.35	1.45	0.85	0.00
Multiple sclerosis	0.57	1.12	0.81	0.76	0.71	1.59	0.77	1.20
Myasthenia gravis	1.40	0.88	1.19	1.45	0.48	1.12	0.85	1.82

Non-acute cystitis	1.41	1.50	0.69	0.89	0.89	1.03	0.77	0.47
Non-melanoma skin malignancies	1.09	0.81	0.79	0.75	0.89	1.36	0.78	1.04
Osteoarthritis (excl spine)	3.05	0.92	0.83	0.66	0.00	0.98	0.74	0.50
Osteoporosis and vertebral crush fractures	1.32	1.01	1.05	0.84	0.62	1.25	0.78	1.20
Parkinson's disease	1.27	0.45	0.83	0.96	0.94	1.39	0.69	0.93
Peptic ulcer disease	2.20	0.89	0.71	0.72	0.62	1.09	0.69	0.82
Peripheral arterial disease	2.31	0.61	0.68	0.57	1.56	0.73	0.55	1.28
Peripheral or autonomic neuropathy	1.88	1.01	0.91	0.80	0.95	0.88	0.74	1.59
Postviral fatigue syndrome, neurasthenia and fibromyalgia	1.26	1.45	1.35	0.90	0.48	0.99	0.96	0.66
Primary pulmonary hypertension	2.00	1.20	0.95	0.26	1.76	0.69	0.31	3.31
Psoriasis	1.05	0.85	0.85	0.77	0.80	1.25	0.96	0.84
Sarcoidosis	1.07	0.64	1.21	1.27	0.88	1.20	0.80	0.71
Sickle-cell anaemia	0.90	1.54	0.90	0.88	0.73	1.42	0.60	0.00
Sleep apnoea	1.74	0.76	1.12	0.61	1.12	0.76	0.93	0.74
Solid organ malignancies	1.14	0.89	0.81	0.85	0.86	1.32	0.76	1.09
Spinal stenosis	2.04	0.76	0.97	0.49	0.70	1.10	0.71	0.69
Stroke	1.97	1.25	0.72	0.43	1.65	0.66	0.53	1.33
Thalassaemia	0.92	0.84	0.86	1.34	0.98	1.16	0.94	0.00
Transient ischaemic attack	2.18	1.27	0.69	0.36	1.43	0.72	0.54	0.85
Tuberculosis	1.39	0.80	0.99	0.86	0.80	1.21	0.76	1.35
Type 1 diabetes	1.25	0.51	0.82	0.55	1.59	0.97	0.72	4.52
Type 2 diabetes	2.18	0.49	0.84	0.42	1.87	0.54	0.64	1.61

Urinary Incontinence	1.31	1.37	1.21	0.88	0.53	1.10	0.84	0.80
Visual impairment and blindness	1.50	1.25	1.09	0.35	0.66	0.79	0.78	15.17

Supplementary Table 2.b. ARF values for *women-only* clusters.

Condition	Very extensive morbidity	Extensive morbidity	Respiratory	Digestive	MSK	CVD + diabetes	Mixed including cancer	Healthy + rhinitis
Addisons disease	1.28	1.20	1.15	0.00	0.56	0.74	1.43	0.80
Allergic and chronic rhinitis	2.37	0.44	2.80	0.68	0.44	0.53	0.00	2.80
Asbestosis	0.00	0.00	0.00	0.00	0.87	2.18	0.98	0.90
Asthma	1.67	0.74	5.79	0.58	0.54	0.47	0.78	0.00
Atrial fibrillation	1.91	1.42	0.63	1.54	1.25	1.50	0.76	0.29
Benign neoplasm of brain and other parts of central nervous system	1.27	0.84	0.83	0.94	1.07	1.02	1.17	0.63
Bronchiectasis	1.61	1.20	2.29	0.56	1.01	0.74	0.96	0.49
COPD	2.03	1.81	2.26	2.08	1.01	0.86	0.83	0.20
Cardiomyopathy	2.06	1.37	1.09	1.08	0.62	1.79	0.75	0.34
Cerebral Palsy	0.94	1.96	0.00	2.47	1.09	0.91	1.40	0.45
Chronic liver disease	1.37	1.58	0.91	1.05	0.92	0.76	1.24	0.68
Chronic renal disease	1.67	1.77	0.65	1.96	1.08	1.63	0.70	0.36
Coeliac disease	1.19	1.32	1.17	2.56	0.70	0.60	1.25	1.04
Conduction disorders and other arrhythmias	1.40	1.44	0.79	1.58	0.89	1.17	1.08	0.56
Coronary heart disease	2.21	2.04	0.68	1.88	1.10	1.66	0.57	0.25
Cystic Fibrosis	0.41	1.03	0.90	1.63	0.79	0.72	1.29	1.26
Dementia	1.29	1.07	0.62	0.00	1.19	0.99	1.33	0.31
Diabetes NOS	1.83	1.50	0.62	2.25	1.00	1.57	0.78	0.39
Diverticular disease of intestine (acute and chronic)	2.24	2.14	0.84	1.73	1.10	0.94	0.86	0.51
Down's syndrome	0.00	0.00	0.00	0.00	0.32	0.27	2.69	0.66
Epilepsy	1.16	1.04	0.98	0.48	0.92	0.77	1.33	0.73

Erectile dysfunction								
Fatty Liver	3.31	3.14	0.98	1.84	0.76	1.31	0.48	0.35
Gastro-oesophageal reflux, gastritis and similar	1.87	5.43	1.00	5.43	0.00	0.81	0.93	0.61
Glaucoma	1.24	1.39	0.67	0.52	1.07	1.19	1.02	0.64
Gout	1.86	1.85	0.96	2.51	1.34	1.43	0.57	0.31
HIV	0.31	0.59	1.02	0.00	0.05	0.68	1.83	1.24
Haematological malignancies	0.91	1.06	0.90	0.95	0.98	1.06	1.27	0.51
Hearing loss	1.66	1.46	0.90	0.90	1.15	0.87	1.03	0.64
Heart failure	2.19	1.56	1.03	1.76	1.20	1.65	0.56	0.16
Heart valve disorders	1.58	1.27	0.63	1.18	0.95	1.28	1.03	0.53
Hyperplasia of prostate								
Hypertension	2.57	1.06	0.60	1.13	0.87	3.09	0.00	0.00
Hypo or hyperthyroidism	1.18	1.14	0.84	0.95	0.94	1.01	1.17	0.70
Immunodeficiencies	1.81	0.97	0.93	4.07	0.36	1.12	1.36	0.46
Inflammatory arthritis and other inflammatory conditions	1.79	1.93	0.77	1.23	1.37	0.79	0.91	0.59
Inflammatory bowel disease	1.21	1.22	0.84	0.98	0.89	0.74	1.29	0.84
Iron and vitamin deficiency anaemia	1.14	1.30	0.92	1.47	0.79	0.84	1.28	0.79
Irritable bowel syndrome	4.46	0.85	0.69	1.39	0.47	0.49	1.11	0.90
Macular degeneration	1.72	1.26	0.75	0.80	1.05	1.16	1.00	0.54
Meniere disease	2.20	1.90	0.78	1.58	0.96	0.81	0.98	0.71
Migraine	1.42	1.13	0.98	1.31	0.85	0.76	1.22	0.85
Motor neuron disease	3.20	1.99	0.00	0.00	1.11	0.46	1.24	0.57
Multiple sclerosis	0.95	0.53	0.73	1.29	0.67	0.74	1.61	0.82

Myasthenia gravis	2.20	0.91	0.79	0.00	1.15	0.53	1.14	0.92
Non-acute cystitis	1.70	2.02	0.66	2.78	0.88	0.98	0.97	0.73
Non-melanoma skin malignancies	0.97	1.07	0.77	1.07	1.07	0.85	1.28	0.69
Osteoarthritis (excl spine)	3.19	3.71	0.55	0.00	3.71	0.00	0.00	0.00
Osteoporosis and vertebral crush fractures	1.39	1.53	0.90	1.27	1.20	0.77	1.09	0.64
Parkinson's disease	1.47	1.64	0.85	2.31	1.27	0.89	1.00	0.53
Peptic ulcer disease	1.71	3.43	0.80	35.75	0.69	0.43	0.61	0.30
Peripheral arterial disease	2.05	2.07	0.62	2.26	1.01	1.37	0.79	0.37
Peripheral or autonomic neuropathy	2.00	1.91	0.93	2.03	1.22	0.87	0.89	0.51
Postviral fatigue syndrome, neurasthenia and fibromyalgia	2.35	1.84	1.13	1.30	0.90	0.61	0.96	0.85
Primary pulmonary hypertension	0.00	2.53	0.73	2.66	0.82	1.96	0.98	0.00
Psoriasis	1.16	1.18	0.86	1.34	1.01	0.77	1.21	0.83
Sarcoidosis	0.99	1.71	1.15	1.30	0.99	1.05	1.02	0.61
Sickle-cell anaemia	0.95	0.00	1.37	0.00	1.16	0.83	1.38	0.51
Sleep apnoea	2.98	1.52	1.27	2.31	0.83	1.23	0.62	0.54
Solid organ malignancies	1.02	1.16	0.76	1.17	0.95	0.92	1.27	0.71
Spinal stenosis	2.51	2.40	0.91	1.06	1.94	0.56	0.63	0.37
Stroke	1.94	1.64	0.71	1.86	1.07	1.63	0.70	0.28
Thalassaemia	0.57	1.91	0.83	0.00	0.66	0.89	1.26	1.03
Transient ischaemic attack	2.17	2.21	0.53	1.33	1.11	1.53	0.67	0.29
Tuberculosis	1.52	1.46	0.82	0.88	1.22	0.87	1.01	0.71
Type 1 diabetes	1.15	1.01	0.83	2.13	0.71	1.57	0.98	0.61
Type 2 diabetes	2.33	1.58	0.79	1.24	1.09	1.90	0.47	0.22

Urinary Incontinence	1.81	1.57	0.98	1.11	1.03	0.76	1.05	0.73
Visual impairment and blindness	2.37	1.53	0.89	0.74	1.05	0.98	0.92	0.54

Supplementary Table 2.c. ARF values for *men-only* clusters.

Condition	Very extensive morbidity	MSK + others	Respiratory + gout + male genitourinary	Digestive	CVD + diabetes	CVD + CKD + gout	Mixed including cancer	Healthy + rhinitis
Addisons disease	1.29	0.48	0.00	0.00	1.13	0.64	0.73	1.93
Allergic and chronic rhinitis	0.49	0.80	2.78	0.71	0.74	0.38	0.00	3.20
Asbestosis	1.17	1.66	0.54	1.85	0.75	0.58	1.05	0.44
Asthma	0.72	0.86	6.08	0.76	0.82	0.39	1.02	1.46
Atrial fibrillation	1.54	1.01	1.37	1.05	1.18	1.81	0.80	0.43
Benign neoplasm of brain and other parts of central nervous system	1.50	1.10	1.16	0.55	0.89	0.91	1.15	0.70
Bronchiectasis	1.04	1.16	2.20	0.55	0.83	0.94	1.23	0.81
COPD	1.30	1.44	2.82	1.48	0.91	0.96	0.85	0.46
Cardiomyopathy	1.61	0.89	2.60	1.02	1.14	2.15	0.68	0.55
Cerebral Palsy	0.44	1.14	2.06	0.88	1.12	1.10	1.40	0.42
Chronic liver disease	1.35	0.84	1.74	2.12	0.82	1.32	1.20	0.63
Chronic renal disease	1.95	0.90	1.23	1.03	1.22	2.57	0.64	0.26
Coeliac disease	0.91	0.79	2.29	0.93	0.75	0.84	1.53	0.98
Conduction disorders and other arrhythmias	1.24	0.90	0.91	1.14	1.14	1.36	1.13	0.50
Coronary heart disease	1.70	1.10	1.08	1.07	1.33	1.49	0.69	0.26
Cystic Fibrosis	0.00	0.44	0.00	2.35	1.03	5.28	0.40	1.45
Dementia	1.68	1.47	1.74	1.48	0.87	1.48	0.51	0.61
Diabetes NOS	2.61	0.84	1.30	0.74	1.38	1.06	0.66	0.20
Diverticular disease of intestine (acute and chronic)	1.26	1.29	1.64	1.30	0.89	1.29	0.99	0.52
Down's syndrome	0.00	1.05	0.00	0.00	0.00	0.00	2.89	1.16
Epilepsy	0.99	1.02	0.69	0.79	0.82	0.78	1.55	0.69

Erectile dysfunction	8.11	0.57	4.30	0.44	0.00	0.00	0.00	0.41
Fatty Liver	1.64	1.03	1.06	1.09	1.19	1.36	0.69	0.61
Gastro-oesophageal reflux, gastritis and similar	1.14	1.15	1.10	2.31	0.77	0.94	1.14	0.65
Glaucoma	1.24	0.97	1.05	0.85	1.08	1.03	1.20	0.57
Gout	1.36	0.48	8.30	0.32	0.00	13.49	0.00	0.27
HIV	1.11	0.66	1.47	1.04	0.77	0.31	1.78	0.86
Haematological malignancies	1.12	0.93	0.67	1.36	0.89	1.48	1.28	0.64
Hearing loss	1.31	1.29	1.37	1.06	0.85	0.90	1.17	0.51
Heart failure	1.82	1.03	1.40	1.22	1.29	2.29	0.55	0.24
Heart valve disorders	1.38	1.03	1.39	1.25	1.16	1.37	0.90	0.48
Hyperplasia of prostate	1.46	1.24	1.42	0.86	0.85	0.84	1.12	0.60
Hypertension	1.57	0.86	0.60	0.80	2.37	1.70	0.00	0.00
Hypo or hyperthyroidism	1.24	0.94	1.55	0.82	1.01	0.82	1.24	0.68
Immunodeficiencies	1.61	1.38	0.00	0.00	0.78	0.80	1.18	0.66
Inflammatory arthritis and other inflammatory conditions	1.02	1.62	1.33	1.04	0.75	1.31	1.02	0.51
Inflammatory bowel disease	1.03	0.93	0.94	1.28	0.76	0.76	1.37	0.96
Iron and vitamin deficiency anaemia	1.30	1.09	1.40	1.83	0.79	1.08	1.14	0.63
Irritable bowel syndrome	0.96	1.01	1.10	1.43	0.74	0.68	1.33	0.98
Macular degeneration	1.44	1.01	1.38	1.02	0.99	0.91	1.17	0.56
Meniere disease	1.02	1.22	1.54	1.39	0.87	0.82	1.31	0.48
Migraine	0.95	0.99	0.89	0.96	0.84	0.68	1.40	0.90
Motor neuron disease	1.23	1.81	0.00	0.00	0.48	0.81	0.93	1.12
Multiple sclerosis	2.32	0.43	1.03	1.17	0.73	0.44	1.52	0.78
Myasthenia gravis	1.67	1.31	0.00	0.88	0.90	0.00	1.05	0.73

Non-acute cystitis	1.52	1.08	1.57	0.45	0.98	0.67	1.38	0.37
Non-melanoma skin malignancies	1.02	1.08	0.98	0.76	0.88	1.03	1.36	0.67
Osteoarthritis (excl spine)	0.95	4.45	1.19	0.00	0.00	1.01	0.00	0.00
Osteoporosis and vertebral crush fractures	1.34	1.29	1.38	0.98	0.76	0.76	1.18	0.65
Parkinson's disease	2.08	0.88	1.52	1.01	0.70	0.76	1.43	0.48
Peptic ulcer disease	0.82	1.18	0.74	18.36	0.00	0.80	0.00	0.00
Peripheral arterial disease	2.29	1.14	0.47	1.32	1.20	1.33	0.61	0.20
Peripheral or autonomic neuropathy	2.02	1.30	0.99	1.36	0.87	1.16	0.78	0.44
Postviral fatigue syndrome, neurasthenia and fibromyalgia	1.20	1.05	0.99	1.12	0.64	0.77	1.26	1.07
Primary pulmonary hypertension	1.48	0.87	0.00	0.78	1.84	1.17	0.67	0.16
Psoriasis	1.08	0.94	1.03	0.84	0.82	1.01	1.34	0.89
Sarcoidosis	0.82	0.80	2.32	0.56	0.89	0.99	1.66	0.70
Sickle-cell anaemia	0.00	1.31	0.00	1.76	1.03	0.00	1.20	1.09
Sleep apnoea	1.62	1.21	0.95	0.95	1.02	1.34	0.80	0.57
Solid organ malignancies	1.62	1.00	1.27	1.07	0.83	0.99	1.25	0.57
Spinal stenosis	1.11	2.01	1.53	0.62	0.68	1.21	0.84	0.46
Stroke	1.73	1.02	0.61	1.07	1.44	1.67	0.64	0.23
Thalassaemia	0.91	0.78	0.00	1.80	0.70	1.80	1.33	0.99
Transient ischaemic attack	1.78	1.13	0.80	0.92	1.30	1.72	0.70	0.20
Tuberculosis	0.98	1.01	1.89	1.07	0.84	1.13	1.47	0.56
Type 1 diabetes	2.92	0.53	1.72	0.59	1.23	0.22	0.99	0.42
Type 2 diabetes	2.56	0.91	1.10	0.84	1.37	1.57	0.49	0.22

Urinary Incontinence	2.20	1.24	1.55	0.79	0.63	0.48	1.03	0.71
Visual impairment and blindness	1.38	1.21	0.48	0.68	0.82	0.81	1.18	0.75

Supplementary Table 2.d. Adjusted *p*-values values for *whole*-cohort clusters.

Condition	Very extensive morbidity	Migraine	Respiratory	Respiratory w/o other	CVD + diabetes	Mixed including cancer	Healthy + rhinitis	Macular degeneration + diabetes
Addisons disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Allergic and chronic rhinitis	0.00E+00	5.20E-58	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.40E-173
Asbestosis	1.00E-06	8.97E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.88E-02	1.00E+00
Asthma	1.90E-14	2.34E-213	0.00E+00	0.00E+00	1.40E-270	0.00E+00	0.00E+00	5.89E-08
Atrial fibrillation	4.36E-53	5.07E-05	6.81E-03	4.06E-07	3.04E-28	6.59E-07	1.16E-19	1.00E+00
Benign neoplasm of brain and other parts of central nervous system	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.99E-02	3.17E-01	1.00E+00
Bronchiectasis	5.93E-02	3.71E-05	1.07E-21	2.19E-17	1.42E-02	4.47E-01	2.46E-10	1.00E+00
COPD	3.07E-67	2.35E-15	9.73E-102	8.51E-120	1.29E-01	6.00E-62	1.24E-118	1.00E+00
Cardiomyopathy	1.76E-04	7.04E-02	1.00E+00	1.00E+00	1.29E-04	1.00E+00	2.53E-01	1.00E+00
Cerebral Palsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.57E-01	1.00E+00	1.00E+00
Chronic liver disease	2.75E-10	6.49E-01	1.00E+00	1.00E+00	1.67E-03	2.50E-03	9.20E-04	1.00E+00
Chronic renal disease	2.86E-113	4.51E-05	1.96E-05	4.12E-17	2.02E-41	1.88E-21	7.15E-33	1.00E+00
Coeliac disease	1.00E+00	1.00E+00	4.77E-01	1.00E+00	5.55E-12	1.18E-06	1.00E+00	1.00E+00
Conduction disorders and other arrhythmias	7.24E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.14E-10	1.00E+00
Coronary heart disease	0.00E+00	4.70E-28	5.24E-17	9.89E-36	1.39E-171	7.11E-119	5.50E-137	1.00E+00
Cystic Fibrosis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Dementia	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Diabetes NOS	6.64E-11	1.00E+00	1.00E+00	1.00E+00	7.33E-12	2.15E-04	3.84E-07	1.00E+00
Diverticular disease of intestine (acute and chronic)	9.34E-139	1.00E+00	1.00E+00	5.54E-03	2.26E-09	1.00E+00	1.90E-23	1.00E+00

Down's syndrome	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.44E-01	1.00E+00	1.00E+00
Epilepsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.71E-05	1.06E-26	7.46E-08	1.00E+00
Erectile dysfunction	2.55E-104	3.35E-41	1.45E-15	1.77E-15	3.70E-19	2.46E-01	2.32E-27	5.01E-02
Fatty Liver	9.12E-28	1.00E+00	6.44E-02	4.42E-01	1.00E+00	3.28E-13	3.21E-01	1.00E+00
Gastro-oesophageal reflux, gastritis and similar	0.00E+00	1.00E+00	1.00E+00	5.25E-44	0.00E+00	7.78E-28	4.08E-119	1.70E-20
Glaucoma	6.27E-06	9.19E-03	2.85E-02	6.19E-04	4.75E-03	4.24E-05	9.70E-14	5.77E-13
Gout	7.73E-131	2.32E-37	4.51E-09	2.28E-14	2.46E-52	7.52E-07	3.29E-37	1.00E+00
HIV	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.24E-01	1.27E-03	1.00E+00	1.00E+00
Haematological malignancies	1.00E+00	6.07E-01	1.00E+00	1.00E+00	1.00E+00	4.93E-11	8.21E-07	1.00E+00
Hearing loss	2.12E-46	3.62E-04	4.36E-02	3.38E-02	1.00E+00	6.38E-14	1.69E-27	1.00E+00
Heart failure	1.14E-63	2.80E-06	1.00E+00	3.33E-04	2.28E-27	1.94E-21	2.26E-19	1.00E+00
Heart valve disorders	6.49E-20	1.00E+00	3.89E-02	4.96E-07	6.86E-08	1.00E+00	2.83E-15	1.00E+00
Hyperplasia of prostate	1.91E-56	1.02E-30	3.24E-09	2.32E-06	1.00E+00	6.54E-14	6.21E-16	1.00E+00
Hypertension	0.00E+00	1.04E-85	5.84E-71	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00
Hypo or hyperthyroidism	5.80E-06	1.00E+00	1.00E+00	3.97E-03	1.52E-14	1.96E-65	2.95E-22	1.00E+00
Immunodeficiencies	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Inflammatory arthritis and other inflammatory conditions	7.01E-77	1.00E+00	1.00E+00	1.29E-05	1.15E-21	1.71E-06	9.72E-16	1.00E+00
Inflammatory bowel disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.49E-11	3.32E-22	2.09E-01	1.00E+00
Iron and vitamin deficiency anaemia	4.11E-16	6.14E-07	1.00E+00	2.97E-04	1.33E-35	8.77E-36	3.23E-19	1.00E+00
Irritable bowel syndrome	1.70E-12	7.68E-57	2.34E-09	1.71E-01	3.08E-214	1.51E-20	1.00E+00	4.78E-03
Macular degeneration	8.53E-06	1.00E+00	1.00E+00	3.79E-35	1.90E-138	1.02E-270	1.31E-04	0.00E+00
Meniere disease	1.37E-04	3.15E-01	1.00E+00	1.00E+00	1.18E-03	1.00E+00	7.08E-01	1.00E+00

Migraine	3.69136621242e-312	0.00E+00	7.75E-05	4.04E-280	0.00E+00	0.00E+00	0.00E+00	1.19E-42
Motor neuron disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Multiple sclerosis	2.03E-04	1.00E+00	1.00E+00	1.00E+00	3.30E-02	9.49E-20	5.51E-02	1.00E+00
Myasthenia gravis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Non-acute cystitis	6.79E-01	7.45E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Non-melanoma skin malignancies	7.32E-01	9.62E-04	5.58E-05	2.55E-03	3.77E-02	2.71E-62	7.53E-17	1.00E+00
Osteoarthritis (excl spine)	0.00E+00	8.17E-03	1.94E-24	1.21E-54	0.00E+00	2.03E-08	1.08E-153	5.19E-20
Osteoporosis and vertebral crush fractures	1.74E-15	1.00E+00	1.00E+00	1.00E+00	1.10E-32	4.49E-28	4.10E-16	1.00E+00
Parkinson's disease	1.00E+00	1.11E-01	1.00E+00	1.00E+00	1.00E+00	2.10E-03	1.88E-01	1.00E+00
Peptic ulcer disease	4.57E-216	9.20E-01	4.88E-10	2.31E-04	2.08E-37	2.07E-04	6.11E-38	1.00E+00
Peripheral arterial disease	1.02E-67	3.78E-05	4.44E-03	6.16E-03	7.62E-18	4.61E-10	8.13E-23	1.00E+00
Peripheral or autonomic neuropathy	3.17E-47	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.85E-02	4.69E-10	1.00E+00
Postviral fatigue syndrome, neurasthenia and fibromyalgia	2.63E-07	3.09E-13	5.80E-08	1.00E+00	4.27E-47	1.00E+00	1.00E+00	1.00E+00
Primary pulmonary hypertension	1.20E-01	1.00E+00	1.00E+00	1.00E+00	3.13E-01	1.00E+00	1.12E-02	1.00E+00
Psoriasis	1.00E+00	1.23E-02	2.74E-02	3.60E-03	2.60E-10	1.22E-37	1.00E+00	1.00E+00
Sarcoidosis	1.00E+00	3.37E-01	1.00E+00	1.00E+00	1.00E+00	1.13E-01	6.02E-01	1.00E+00
Sickle-cell anaemia	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Sleep apnoea	2.00E-26	1.25E-01	1.00E+00	1.78E-02	1.00E+00	4.34E-09	1.00E+00	1.00E+00
Solid organ malignancies	1.28E-08	2.36E-02	2.27E-09	1.78E-02	1.31E-08	6.29E-129	9.89E-45	1.00E+00
Spinal stenosis	3.18E-23	1.00E+00	1.00E+00	5.99E-02	6.10E-03	1.00E+00	2.59E-04	1.00E+00

Stroke	5.88E-82	8.94E-04	2.38E-05	4.64E-13	3.73E-50	1.11E-33	6.00E-52	1.00E+00
Thalassaemia	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Transient ischaemic attack	3.56E-72	1.77E-02	4.14E-04	1.71E-10	7.73E-14	9.68E-15	1.74E-31	1.00E+00
Tuberculosis	2.44E-07	5.74E-01	1.00E+00	1.00E+00	2.71E-02	5.21E-06	2.43E-06	1.00E+00
Type 1 diabetes	1.00E+00	1.51E-02	1.00E+00	1.00E+00	1.98E-06	1.00E+00	2.06E-02	2.67E-05
Type 2 diabetes	1.83E-302	6.56E-52	6.79E-04	7.88E-35	2.13E-226	7.51E-160	8.32E-70	8.81E-04
Urinary Incontinence	5.02E-23	5.95E-20	2.53E-06	1.00E+00	8.15E-79	8.00E-08	9.51E-11	1.00E+00
Visual impairment and blindness	2.88E-02	1.00E+00	1.00E+00	2.66E-01	1.82E-01	6.64E-01	1.00E+00	5.24E-33

Supplementary Table 2.e. Adjusted *p*-values values for *women-only* clusters.

Condition	Very extensive morbidity	Extensive morbidity	Respiratory	Digestive	MSK	CVD + diabetes	Mixed including cancer	Healthy + rhinitis
Addisons disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Allergic and chronic rhinitis	0.00E+00	1.75E-156	0.00E+00	8.91E-07	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Asbestosis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Asthma	5.35E-89	5.87E-12	0.00E+00	8.88E-05	4.08E-157	9.84E-263	3.64E-48	0.00E+00
Atrial fibrillation	1.79E-08	3.79E-01	2.40E-02	1.00E+00	4.84E-02	7.02E-12	4.44E-05	4.27E-27
Benign neoplasm of brain and other parts of central nervous system	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.79E-02
Bronchiectasis	1.17E-01	1.00E+00	1.35E-11	1.00E+00	1.00E+00	3.02E-01	1.00E+00	7.75E-06
COPD	2.00E-16	2.62E-08	9.57E-34	9.94E-02	1.00E+00	4.85E-01	1.95E-03	5.93E-54
Cardiomyopathy	4.66E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.16E-03	1.00E+00	4.06E-02
Cerebral Palsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Chronic liver disease	1.00E+00	4.20E-01	1.00E+00	1.00E+00	1.00E+00	4.77E-01	3.63E-02	8.21E-02
Chronic renal disease	3.76E-10	2.50E-10	1.41E-04	6.98E-02	1.00E+00	1.98E-39	1.90E-17	1.01E-42
Coeliac disease	1.00E+00	1.00E+00	1.00E+00	2.14E-01	2.93E-02	5.10E-06	3.11E-03	1.00E+00
Conduction disorders and other arrhythmias	7.53E-01	9.66E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.91E-06
Coronary heart disease	7.94E-67	8.72E-40	4.46E-09	3.74E-04	8.49E-02	3.18E-97	1.09E-77	1.19E-135
Cystic Fibrosis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Dementia	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Diabetes NOS	7.72E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.63E-04	8.21E-01	2.76E-05
Diverticular disease of intestine (acute and chronic)	1.41E-62	2.61E-42	2.00E-01	2.34E-02	2.58E-01	1.00E+00	1.06E-05	1.21E-44

Down's syndrome	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.42E-04	1.00E+00
Epilepsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.90E-03	8.04E-11	2.28E-03
Erectile dysfunction	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Fatty Liver	3.51E-17	9.03E-12	1.00E+00	1.00E+00	1.00E+00	2.85E-01	2.73E-10	7.92E-08
Gastro-oesophageal reflux, gastritis and similar	4.43E-157	0.00E+00	1.00E+00	0.00E+00	0.00E+00	1.14E-26	1.00E+00	3.35E-109
Glaucoma	1.00E+00	5.89E-02	3.75E-03	1.00E+00	1.00E+00	1.05E-02	1.00E+00	3.04E-09
Gout	5.08E-05	8.97E-04	1.00E+00	2.47E-01	1.17E-02	8.67E-06	3.64E-12	2.58E-16
HIV	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.20E-06	1.00E+00	6.82E-06	1.00E+00
Haematological malignancies	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.66E-03	2.47E-06
Hearing loss	6.49E-14	1.91E-05	1.00E+00	1.00E+00	1.76E-02	7.82E-02	1.00E+00	1.43E-16
Heart failure	2.07E-05	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.04E-07	3.05E-07	5.41E-16
Heart valve disorders	3.28E-05	1.00E+00	1.40E-03	1.00E+00	1.00E+00	1.11E-05	1.00E+00	2.14E-14
Hyperplasia of prostate	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Hypertension	0.00E+00	6.66E-02	1.56E-111	1.00E+00	1.21E-11	0.00E+00	0.00E+00	0.00E+00
Hypo or hyperthyroidism	6.75E-05	4.74E-02	1.78E-04	1.00E+00	1.00E+00	1.00E+00	1.60E-47	2.91E-41
Immunodeficiencies	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Inflammatory arthritis and other inflammatory conditions	3.72E-26	1.48E-27	2.58E-03	1.00E+00	9.10E-22	1.08E-08	2.18E-01	5.84E-29
Inflammatory bowel disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.41E-05	1.00E-10	5.59E-01
Iron and vitamin deficiency anaemia	4.65E-01	5.20E-05	1.00E+00	2.67E-01	1.69E-10	1.96E-07	2.51E-48	1.37E-10
Irritable bowel syndrome	0.00E+00	5.81E-02	1.81E-21	4.93E-02	1.61E-152	5.78E-170	1.14E-23	1.80E-02
Macular degeneration	7.88E-06	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.52E-01	1.00E+00	2.34E-10
Meniere disease	4.18E-10	1.82E-04	1.00E+00	1.00E+00	1.00E+00	9.00E-01	1.00E+00	4.13E-02
Migraine	9.45E-27	5.56E-02	1.00E+00	2.97E-01	1.65E-08	3.45E-34	8.61E-77	8.28E-09

Supplementary Table 2.f. Adjusted *p*-values values for *men-only* clusters.

Condition	Very extensive morbidity	MSK + others	Respiratory + gout + male genitourinary	Digestive	CVD + diabetes	CVD + CKD + gout	Mixed including cancer	Healthy + rhinitis
Addisons disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Allergic and chronic rhinitis	6.13E-183	3.76E-37	2.01E-194	5.98E-20	9.71E-92	4.23E-140	0.00E+00	0.00E+00
Asbestosis	1.00E+00	3.06E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.76E-02
Asthma	3.09E-20	6.44E-09	0.00E+00	1.49E-05	3.80E-19	5.37E-59	1.85E-01	9.33E-123
Atrial fibrillation	2.75E-13	1.00E+00	1.00E+00	1.00E+00	3.04E-05	1.90E-13	1.14E-04	4.53E-41
Benign neoplasm of brain and other parts of central nervous system	9.12E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Bronchiectasis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
COPD	4.40E-03	5.04E-17	5.98E-09	1.62E-02	1.00E+00	1.00E+00	7.74E-02	1.41E-28
Cardiomyopathy	1.00E-01	1.00E+00	8.86E-01	1.00E+00	1.00E+00	4.34E-03	2.48E-01	1.24E-02
Cerebral Palsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.59E-01
Chronic liver disease	4.81E-01	1.00E+00	1.00E+00	1.31E-04	7.70E-01	1.00E+00	6.82E-01	8.08E-04
Chronic renal disease	9.25E-29	1.00E+00	1.00E+00	1.00E+00	6.25E-06	1.14E-33	5.35E-14	2.38E-56
Coeliac disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.30E-04	1.00E+00
Conduction disorders and other arrhythmias	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.64E-07
Coronary heart disease	3.44E-93	2.63E-07	1.00E+00	1.00E+00	2.37E-82	4.01E-22	1.85E-49	3.54E-284
Cystic Fibrosis	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Dementia	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Diabetes NOS	3.02E-16	1.00E+00	1.00E+00	1.00E+00	2.82E-03	1.00E+00	3.20E-02	4.10E-15
Diverticular disease of intestine (acute and chronic)	2.44E-04	1.29E-13	6.12E-02	1.08E-01	8.16E-02	3.60E-02	1.00E+00	5.96E-37

Down's syndrome	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Epilepsy	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.43E-02	1.00E+00	1.33E-18	3.89E-05
Erectile dysfunction	0.00E+00	9.81E-90	1.66E-138	6.37E-27	0.00E+00	5.60E-177	0.00E+00	3.22E-167
Fatty Liver	2.03E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.17E-02	4.08E-03
Gastro-oesophageal reflux, gastritis and similar	6.79E-08	3.57E-28	1.00E+00	5.06E-173	1.46E-41	1.00E+00	3.20E-29	7.92E-83
Glaucoma	1.18E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.55E-03	1.63E-16
Gout	4.19E-14	1.37E-81	4.61E-279	8.45E-25	0.00E+00	0.00E+00	0.00E+00	3.34E-159
HIV	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.07E-03	1.00E+00
Haematological malignancies	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.61E-01	7.23E-03	2.17E-04
Hearing loss	2.99E-07	1.83E-16	1.00E+00	1.00E+00	1.02E-04	1.00E+00	1.82E-06	1.16E-45
Heart failure	1.74E-13	1.00E+00	1.00E+00	1.00E+00	9.80E-06	1.37E-14	1.23E-13	1.86E-38
Heart valve disorders	1.98E-03	1.00E+00	1.00E+00	1.00E+00	4.55E-02	2.83E-01	1.00E+00	5.66E-19
Hyperplasia of prostate	2.12E-32	9.04E-25	4.11E-02	1.00E+00	4.16E-09	2.67E-01	4.42E-08	3.16E-55
Hypertension	0.00E+00	1.33E-13	1.05E-17	9.06E-14	0.00E+00	3.42E-274	0.00E+00	0.00E+00
Hypo or hyperthyroidism	3.29E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.54E-06	9.20E-10
Immunodeficiencies	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Inflammatory arthritis and other inflammatory conditions	1.00E+00	3.60E-39	1.00E+00	1.00E+00	7.73E-09	1.32E-01	1.00E+00	1.28E-26
Inflammatory bowel disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.15E-05	1.00E+00	3.58E-09	1.00E+00
Iron and vitamin deficiency anaemia	6.99E-03	1.00E+00	1.00E+00	2.43E-07	1.76E-04	1.00E+00	1.32E-01	4.74E-11
Irritable bowel syndrome	1.00E+00	1.00E+00	1.00E+00	1.05E-04	9.36E-18	5.31E-04	1.26E-23	1.00E+00
Macular degeneration	4.97E-03	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.55E-01	2.50E-08
Meniere disease	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.37E-01	4.38E-05
Migraine	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.04E-06	4.40E-04	5.62E-33	5.49E-01

