Unique genetic and risk-factor profiles in multimorbidity clusters of depression-related disease trajectories from a study of 1.2 million subjects

Andras Gezsi[1 *], Sandra Van der Auwera[2,3 *], Hannu Mäkinen[4], Nora Eszlari[5,6], Gabor Hullam[1,5], Tamas Nagy[1,5,6], Sarah Bonk[2], Rubèn González-Colom[7], Xenia Gonda[5,6,8], Linda Garvert[2], Teemu Paajanen[4], Zsofia Gal[5,6], Kevin Kirchner[2], Andras Millinghofer[9], Carsten O. Schmidt[10], Bence Bolgar[1], Josep Roca[7], Isaac Cano[7], Mikko Kuokkanen[4,11,12], Peter Antal[1 #], Gabriella Juhasz[5,6 # +]

[*] These authors contributed equally to the manuscript

[#] These authors jointly supervised the research

[+] Corresponding author: Gabriella Juhasz

[1] Department of Measurement and Information Systems, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

[2] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, 17475 Greifswald, Germany.

[3] German Centre for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, 17475 Greifswald, Germany.

[4] Department of Public Health and Welfare, Finnish Health and Welfare Institute. Biomedicum 1, Haartmaninkatu 8, 00290 Helsinki.

[5] Department of Pharmacodynamics, Faculty of Pharmacy, Semmelweis University, Nagyvárad tér 4., H-1089 Budapest, Hungary.

[6] NAP3.0-SE Neuropsychopharmacology Research Group, Hungarian Brain Research Program, Semmelweis University, Üllői út 26., H-1085 Budapest, Hungary.

[7] Hospital Clínic de Barcelona. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). Universitat de Barcelona. Villarroel 170. Barcelona, 08036. Spain.

[8] Department of Psychiatry and Psychotherapy, Semmelweis University, Gyulai Pál utca 2., H-1085 Budapest, Hungary

[9] Abiomics Europe Ltd., Zólyomi út. 23., H-1118 Budapest, Hungary

[10] Institute for Community Medicine, University Medicine Greifswald, 17475 Greifswald, Germany.

[11] Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine at University of Texas Rio Grande Valley, Brownsville, TX, United States.

[12] Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Finland

Supplementary File 1

Determination of the optimal number of clusters

Based on the cross-cohort relevance scores, we computed the *weighted direct MDD-related multimorbidity scores* for each participant in each cohort and for each cumulative time interval (see Online Methods). The score for the *i*-th participant in the *t*-th time interval is denoted as:

$$multimorbidity\text{-}score^{(t)}(i)$$

These weighted direct MDD-related multimorbidity scores reflect the temporal MDD-related multimorbidity burden of a participant. These scores define a 4-dimensional space of the participants. More specifically, we used the raw score for the first time interval and the difference between the consecutive scores for the subsequent time intervals, i.e.,

$$d^{(1)}(i) = multimorbidity\text{-}score^{(1)}(i)$$

$$d^{(2)}(i) = multimorbidity\text{-}score^{(2)}(i) - multimorbidity\text{-}score^{(1)}(i)$$

$$d^{(3)}(i) = multimorbidity\text{-}score^{(3)}(i) - multimorbidity\text{-}score^{(2)}(i)$$

$$d^{(4)}(i) = multimorbidity\text{-}score^{(4)}(i) - multimorbidity\text{-}score^{(3)}(i)$$

Together, $D(i) = \{d^{(i)}; i = 1..4\}$ define a 4 dimensional vector for the *i*-th participant.

We clustered all participants whose observed multimorbidity scores were complete (25% of the participants) using the k-means clustering algorithm in this 4-dimensional space. In the case of younger participants, one or more multimorbidity scores were unavailable because there were no observations of their future disease onsets. We assigned them to the nearest cluster (of which cluster center was nearest).

The determination of the optimal number of clusters in our study was conducted through a systematic manual process, employing multiple analytical techniques to ensure accuracy and reliability. During this process, we computed the Adjusted Rand index between consecutive cluster numbers, performed a principal component analysis of participants and disease over- and underrepresentation analysis, compared weighted disease burden among clusters, and computed Silhouette scores of the resulting clusters. The multimorbidity burden is calculated for each individual by summing the relevance score for each diagnosed disease multiplied by the number of years spent in that disease.

More specifically, we followed a heuristic approach: with an increasing number of clusters, (1) we inspected the change in cluster structure by visualizing the switching of individuals between clusters while increasing the number of clusters, (2) we calculated disease risk profiles in the clusters, and (3) assessed the quality of the actual number of clusters based on the other metrics and analyses.

Finally, our investigation led us to select 7 clusters as the optimal solution. The transitions from 5 to 6 clusters and from 6 to 7 clusters demonstrated significant changes in cluster structure, indicating meaningful distinctions between participant groups. However, beyond 7 clusters, the observed changes were insignificant from all viewpoints, which is also reflected by the high Adjusted Rand indices between 7 and 8 clusters (Adj. Rand index = 0.967) and between 8 and 9 (Adj. Rand index = 0.936). Besides, when using 8 clusters, the smallest cluster consisted of only 1.6% of all participants. This highlights the limitations of using a higher number of clusters, as it would lead to exceedingly small clusters that may not provide meaningful insights or statistical robustness.

In the following, we demonstrate the heuristic process of finding the optimal number of clusters through various types of figures.

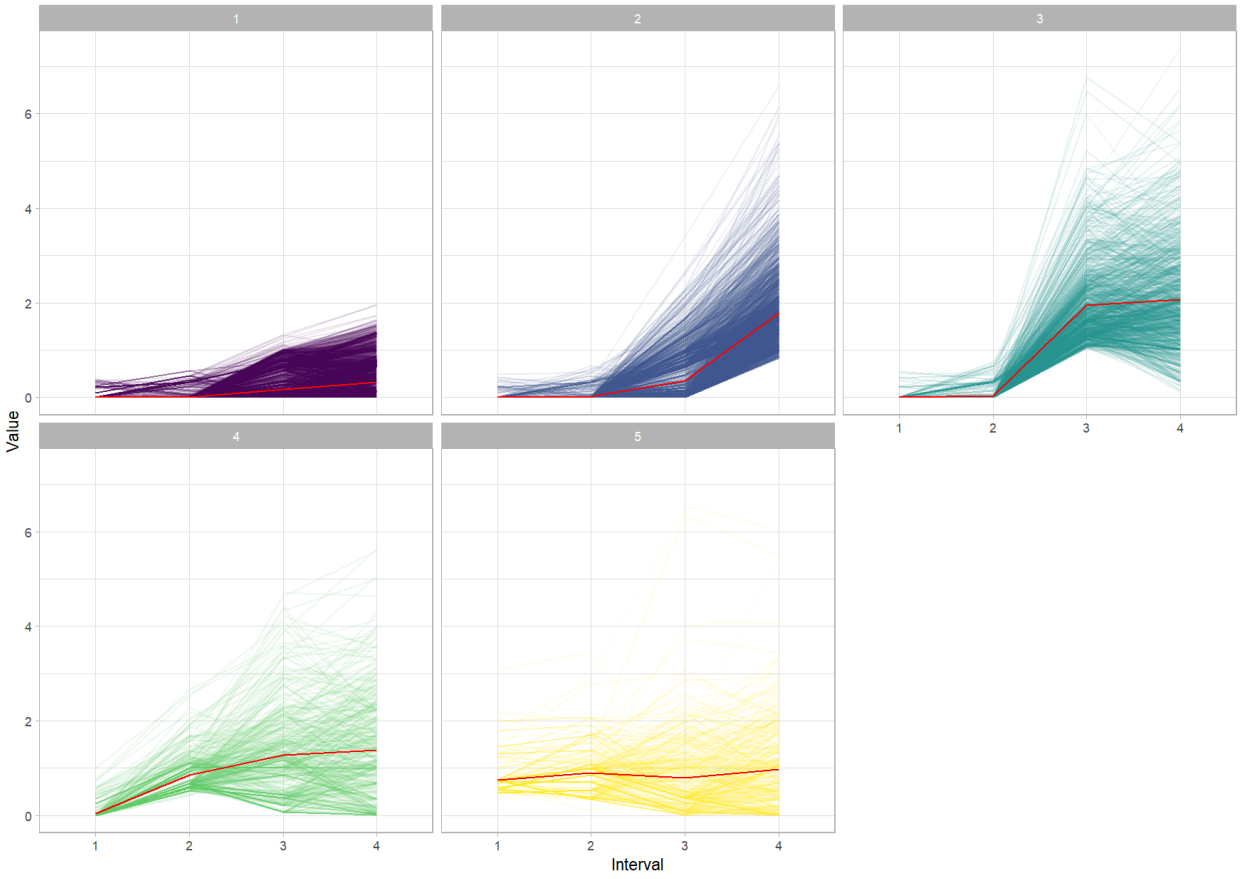


Figure 1A. Trajectories of the weighted direct MDD-related multimorbidity score over time using 5 clusters. Each box corresponds to a cluster of participants in which the trajectories of the scores are similar. Each colored line corresponds to the trajectory of a single individual. The red lines show the mean trajectory in a cluster. The x-axis corresponds to the discrete cumulative time intervals (1: 0-20, 2: 0-40, 3: 0-60, and 4: 0-70), and the y-axis shows the value of the weighted direct MDD-related multimorbidity score.

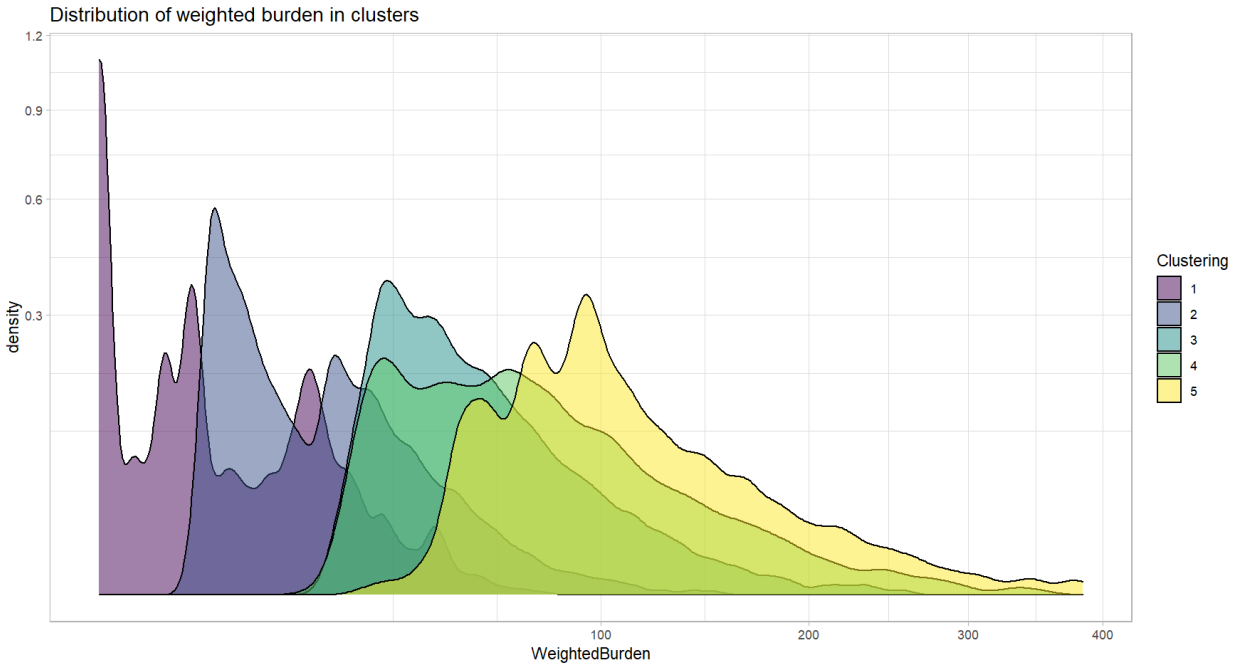


Figure 1B. Distribution of weighted direct MDD-related multimorbidity burden in the clusters where the number of clusters is set to 5. The x-axis corresponds to the weighted disease burden, and the y-axis shows the density of the distribution. Clusters 1 and 2 show multimodal distributions.

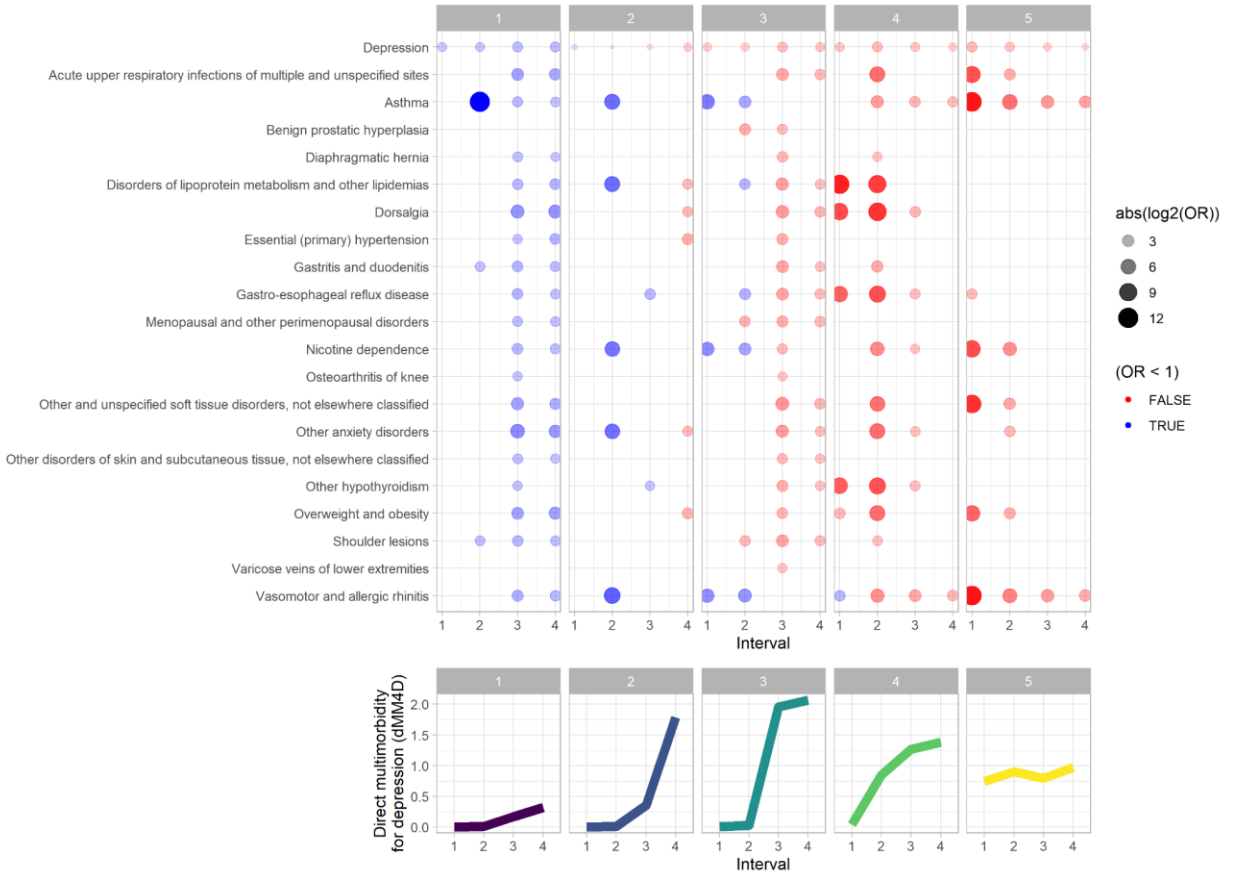


Figure 1C. The top over- and underrepresented diseases in each discrete time interval in the clusters where the number of clusters is set to 5. In the upper panel, the colored dots show the odds ratio of a disease in a specific time interval (indicated by the columns) for the individuals belonging to a given cluster (indicated by the boxes). Fisher's exact test was used to assess the statistical significance of the association between the disease status and the cluster membership. The color and the size of the dots show the direction and the size of the effect (blue: decreased risk, red: increased risk), respectively. For each visible dot, the Bonferroni-adjusted p-value < 0.05, the OR > 3, or OR < ⅓, and the disease prevalence in the time interval exceeds 0.05. Otherwise, the dot is not shown. The lower panel shows the mean trajectory of the weighted direct MDD-related multimorbidity score in the clusters.

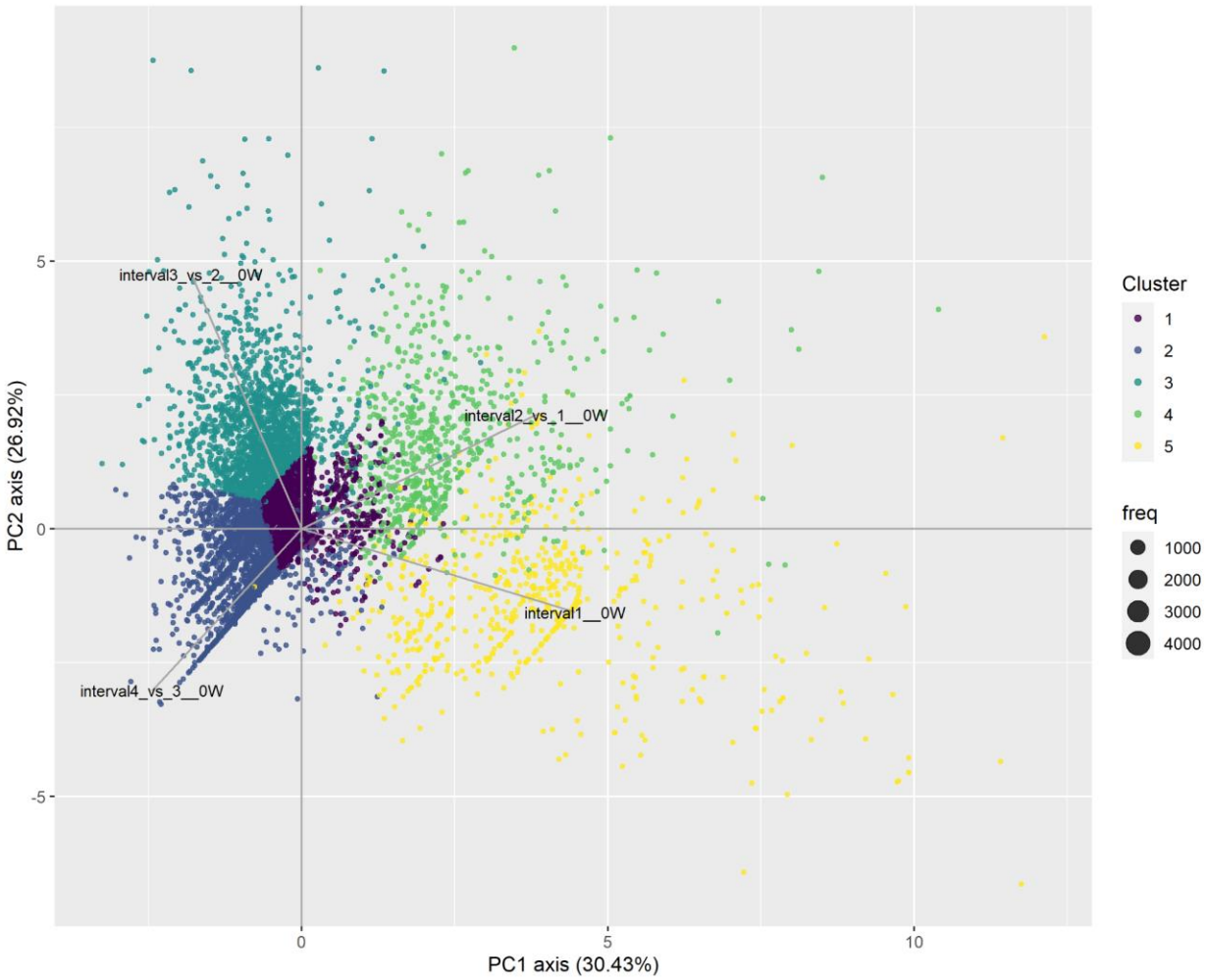


Figure 1D. The first two principal components of the 4-dimensional space defined by the temporal MDD-related multimorbidity burden score of the participants where the number of clusters is set to 5. Dots correspond to individuals, where the size of the dots indicates the number of individuals in that specific region of space. The labeled loading vectors show the strength and direction of the association between each variable in the 4-dimensional space and each principal component.

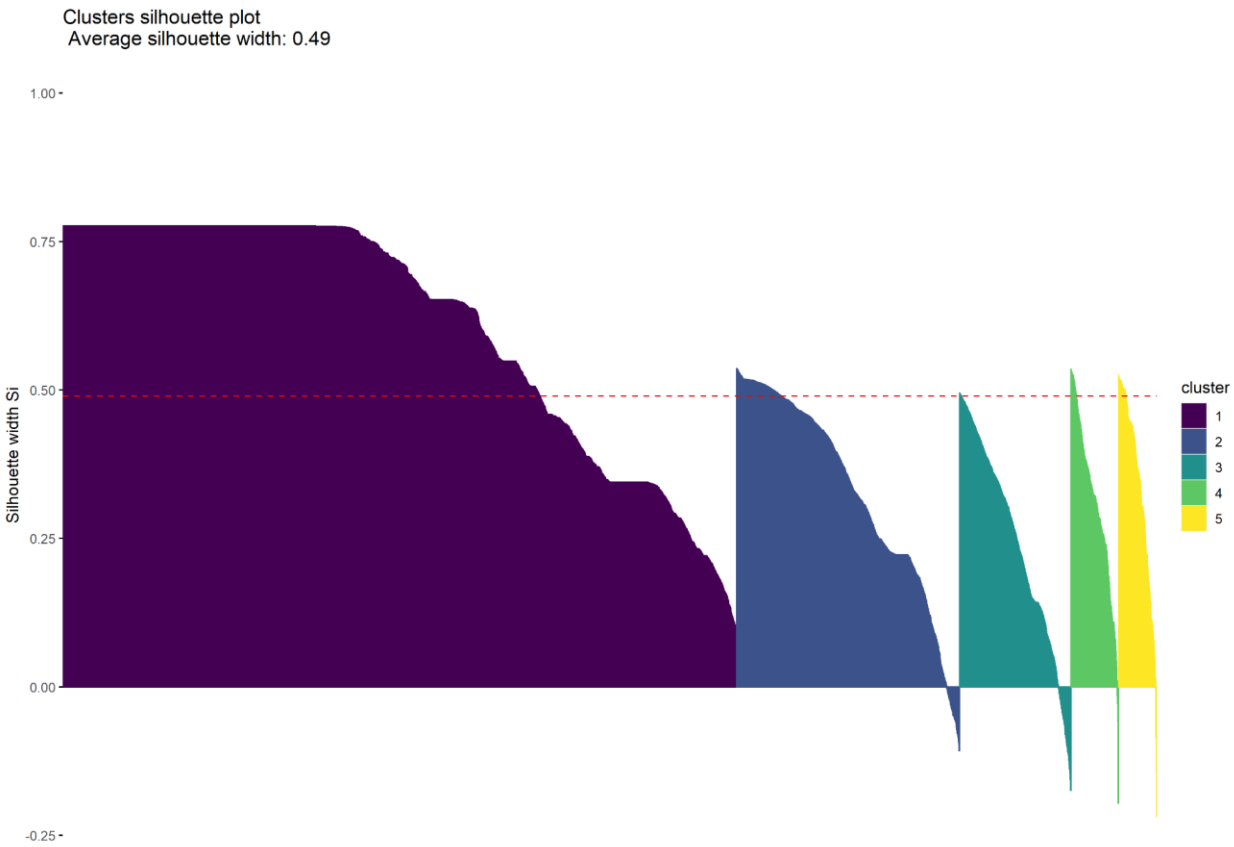


Figure 1E. Silhouette plot of the clusters where the number of clusters is set to 5. Each individual is represented by a vertical line, with its position indicating the cluster it belongs to and its length representing the Silhouette coefficient for that individual.

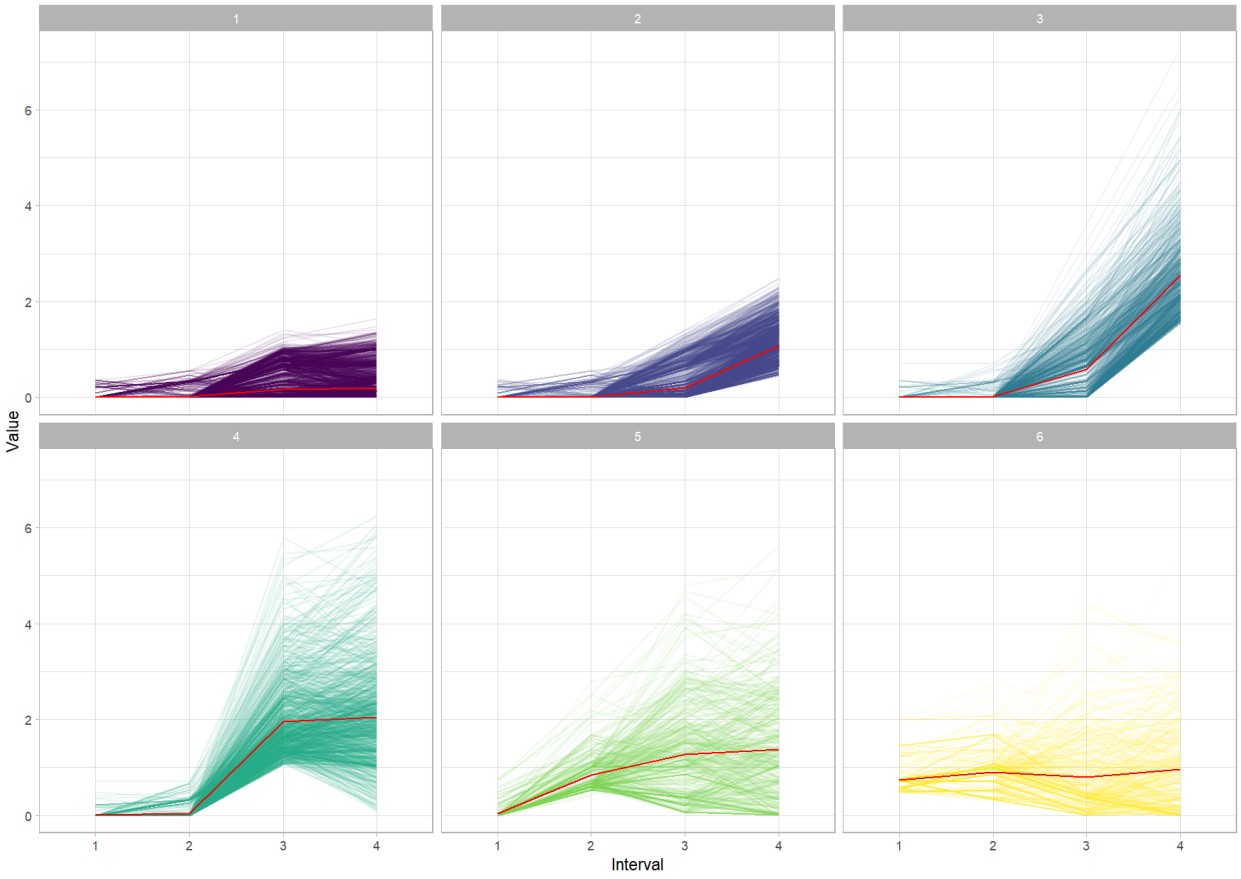


Figure 2A. Trajectories of the weighted direct MDD-related multimorbidity score over time using 6 clusters. See the caption of Figure 1A for details.

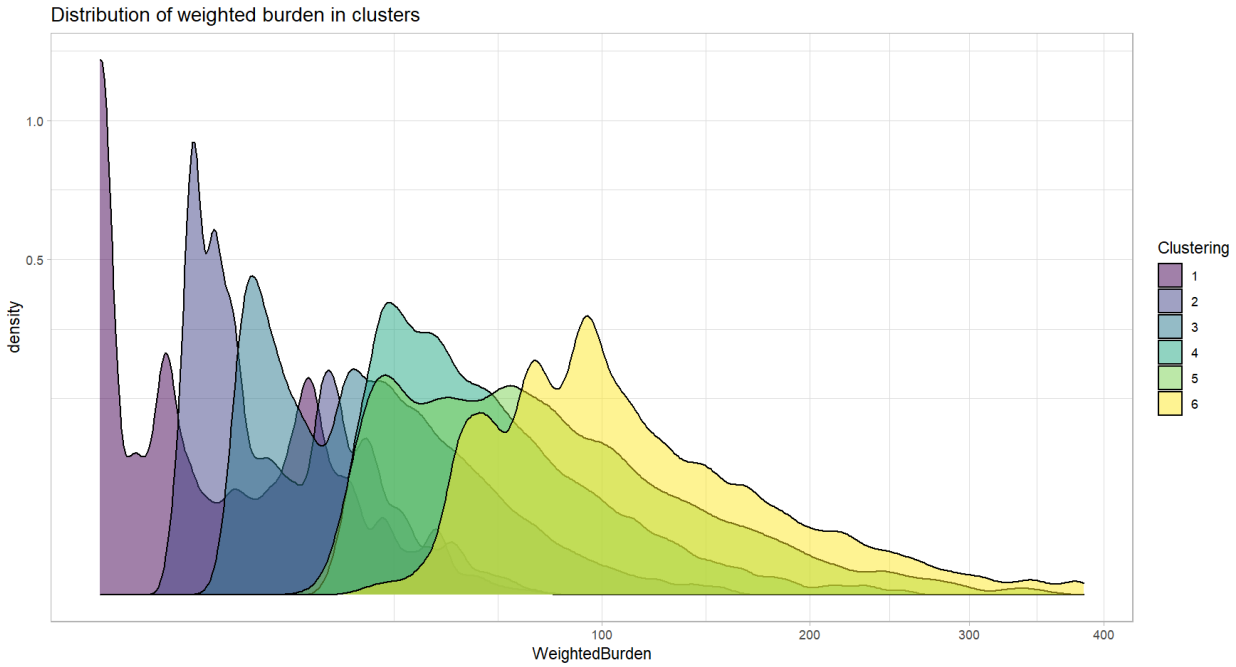
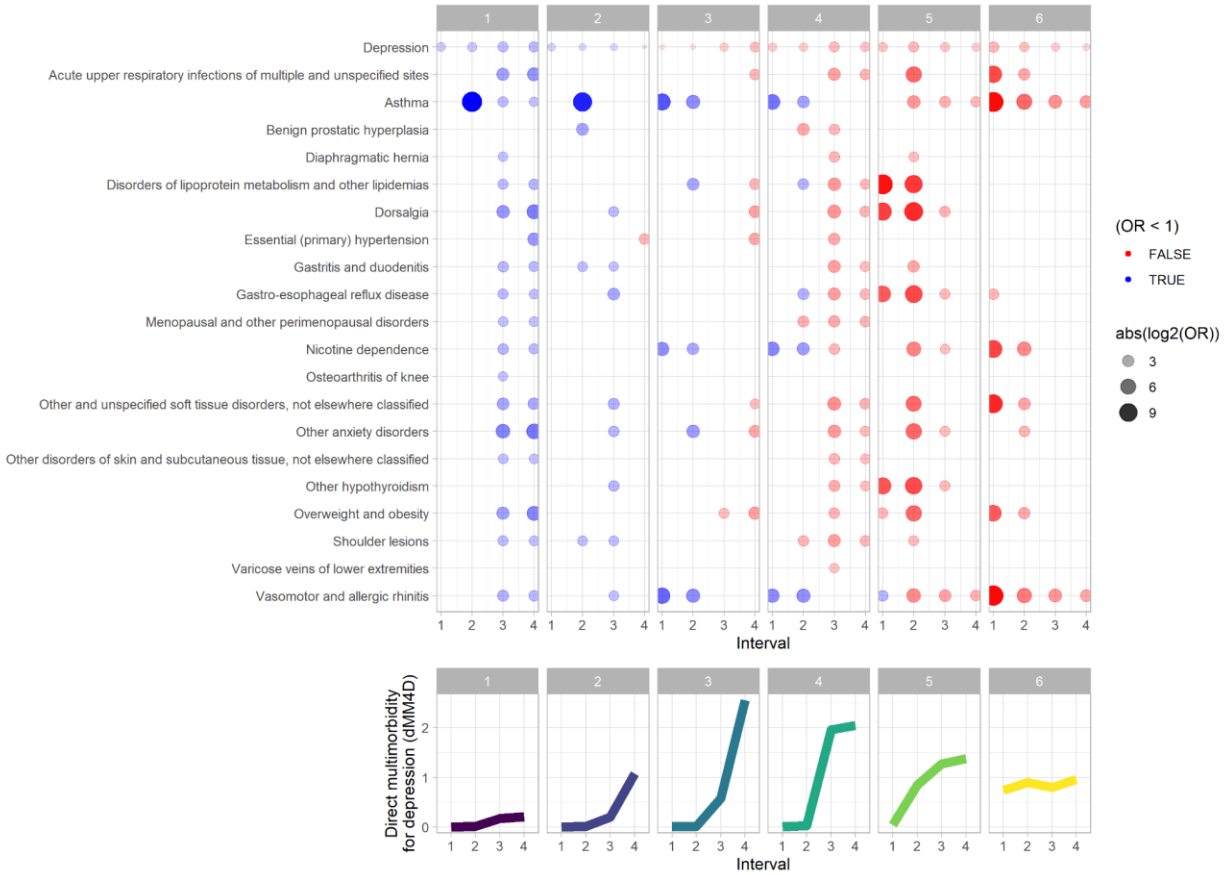


Figure 2B. Distribution of weighted direct MDD-related multimorbidity burden in the clusters where the number of clusters is set to 6. See the caption of Figure 1B for details.



125

126 **Figure 2C. The top over- and underrepresented diseases in each discrete time interval in the**

127 **clusters where the number of clusters is set to 6.** See the caption of Figure 1C for details.

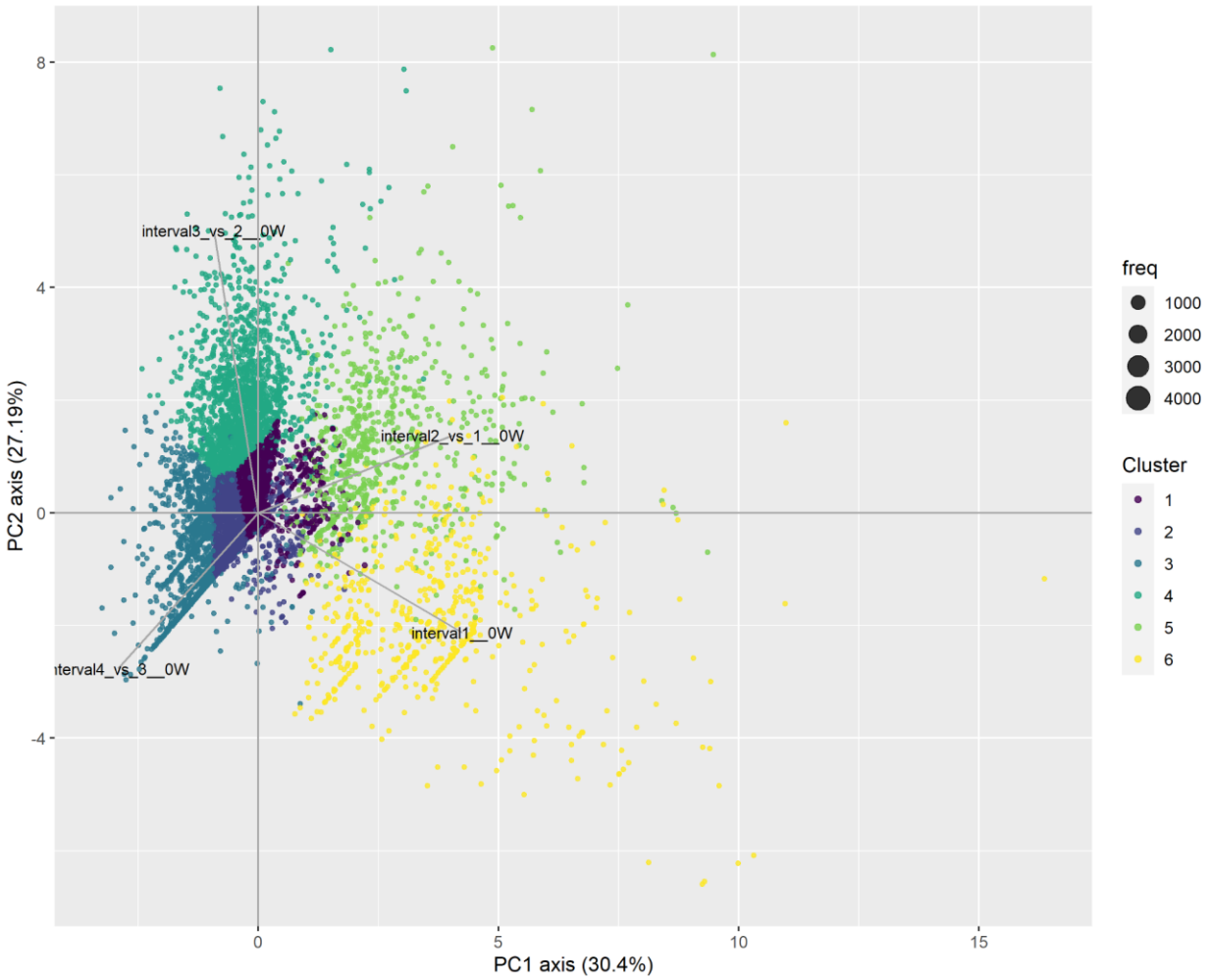


Figure 2D. The first two principal components of the 4-dimensional space defined by the temporal MDD-related multimorbidity burden score of the participants where the number of clusters is set to 6. See the caption of Figure 1D for details.

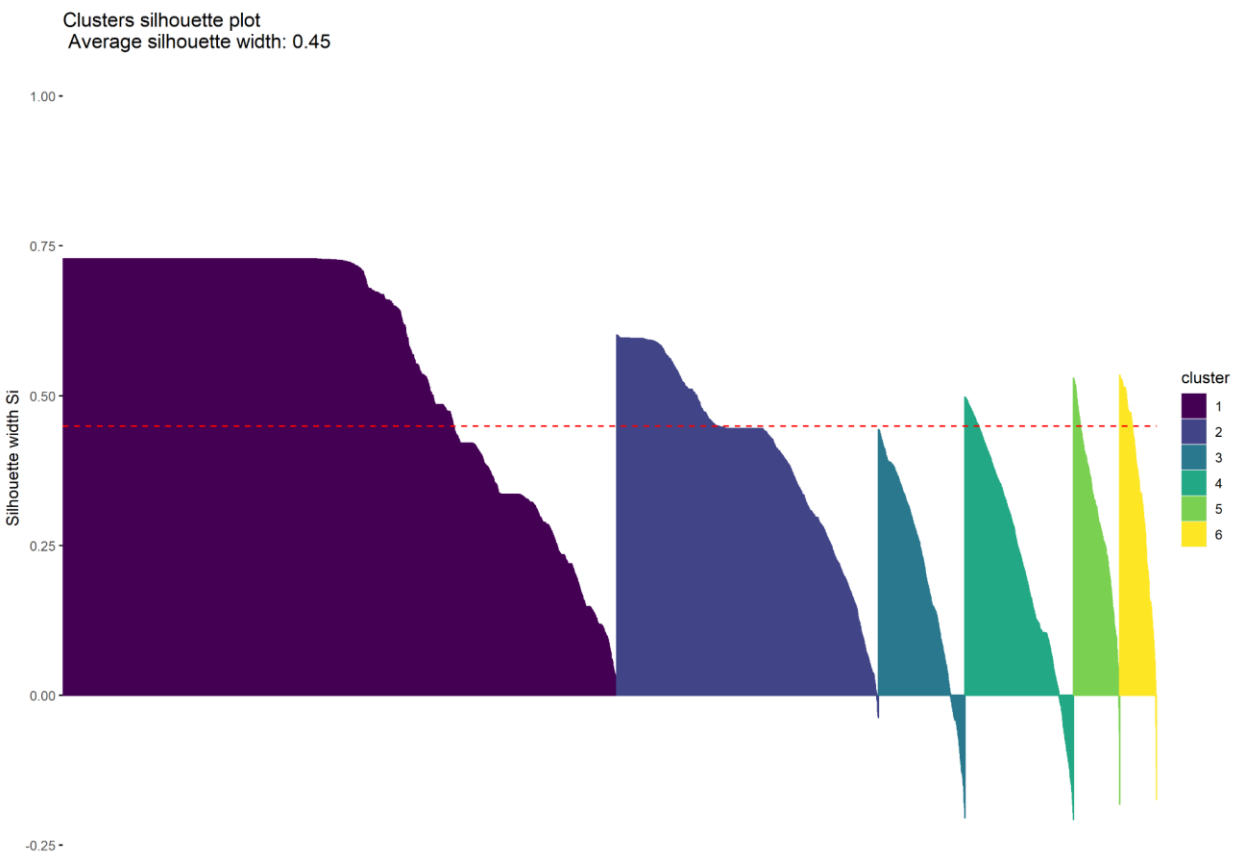


Figure 2E. Silhouette plot of the clusters where the number of clusters is set to 6. See the caption of Figure 1E for details.

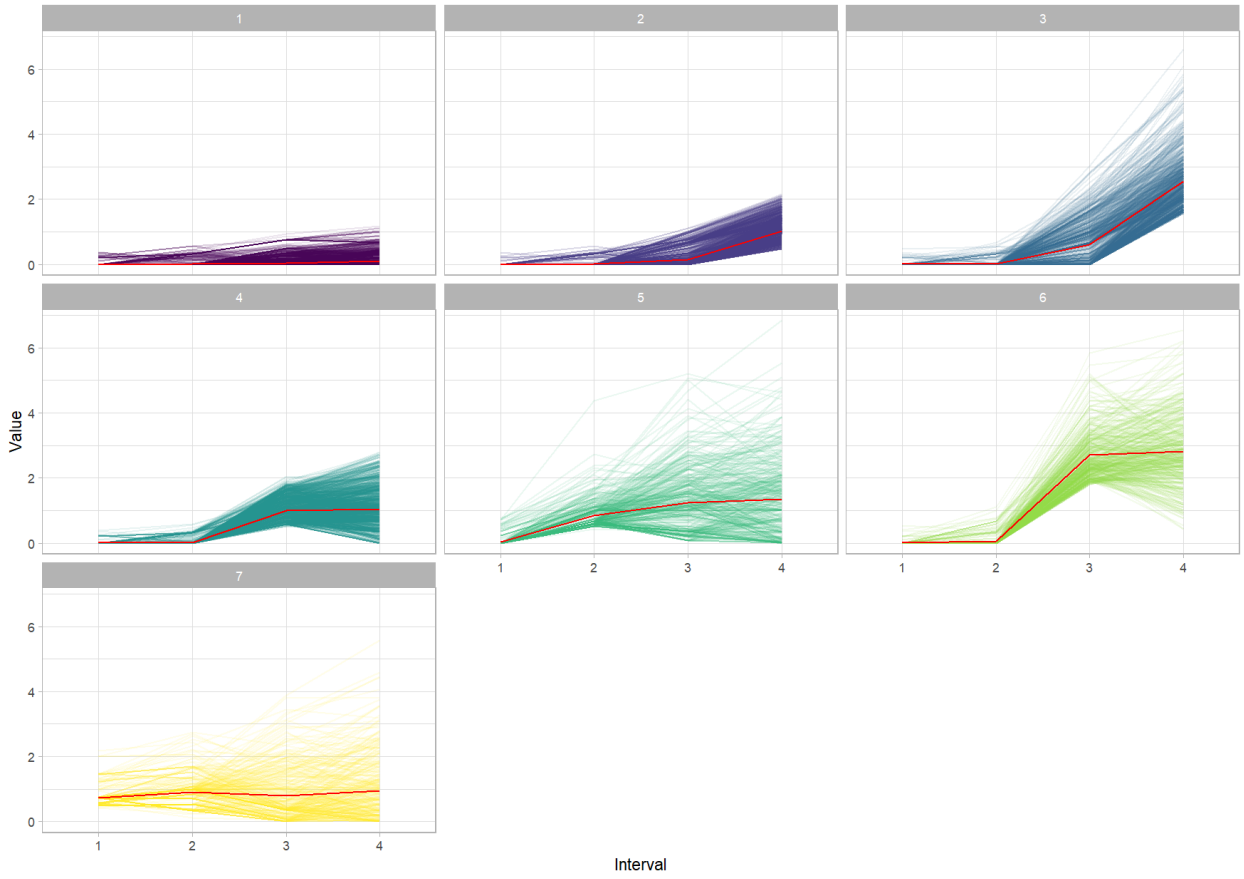


Figure 3A. Trajectories of the weighted direct MDD-related multimorbidity score over time using 7 clusters. See the caption of Figure 1A for details.

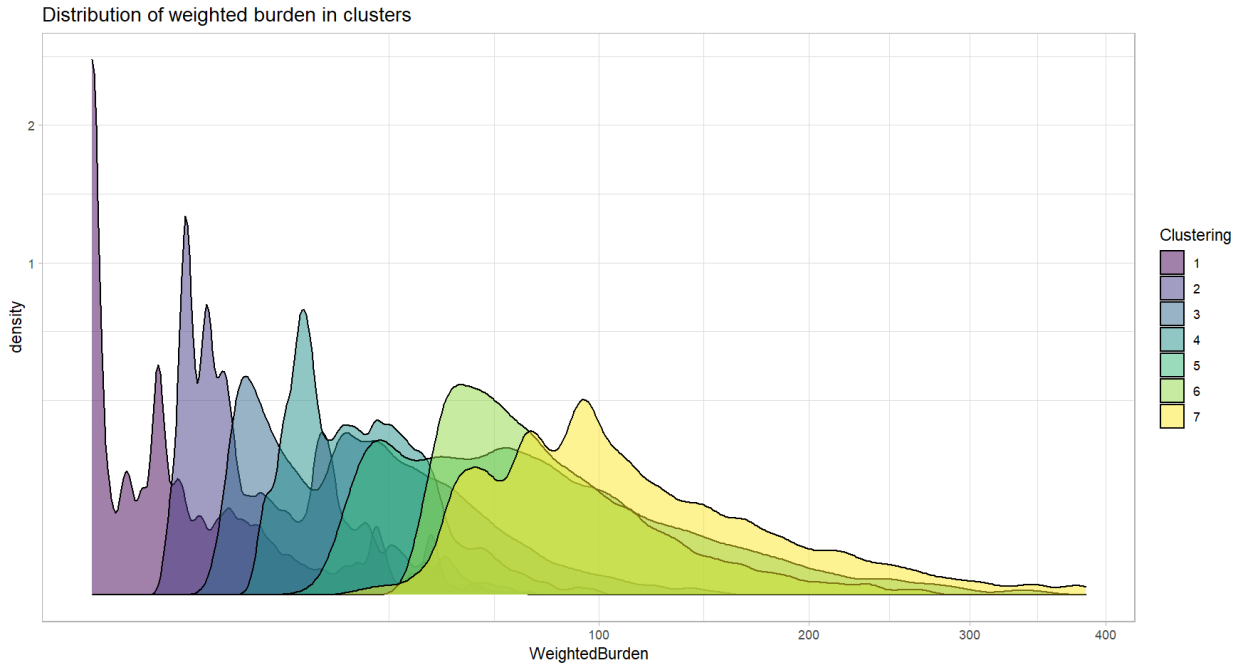
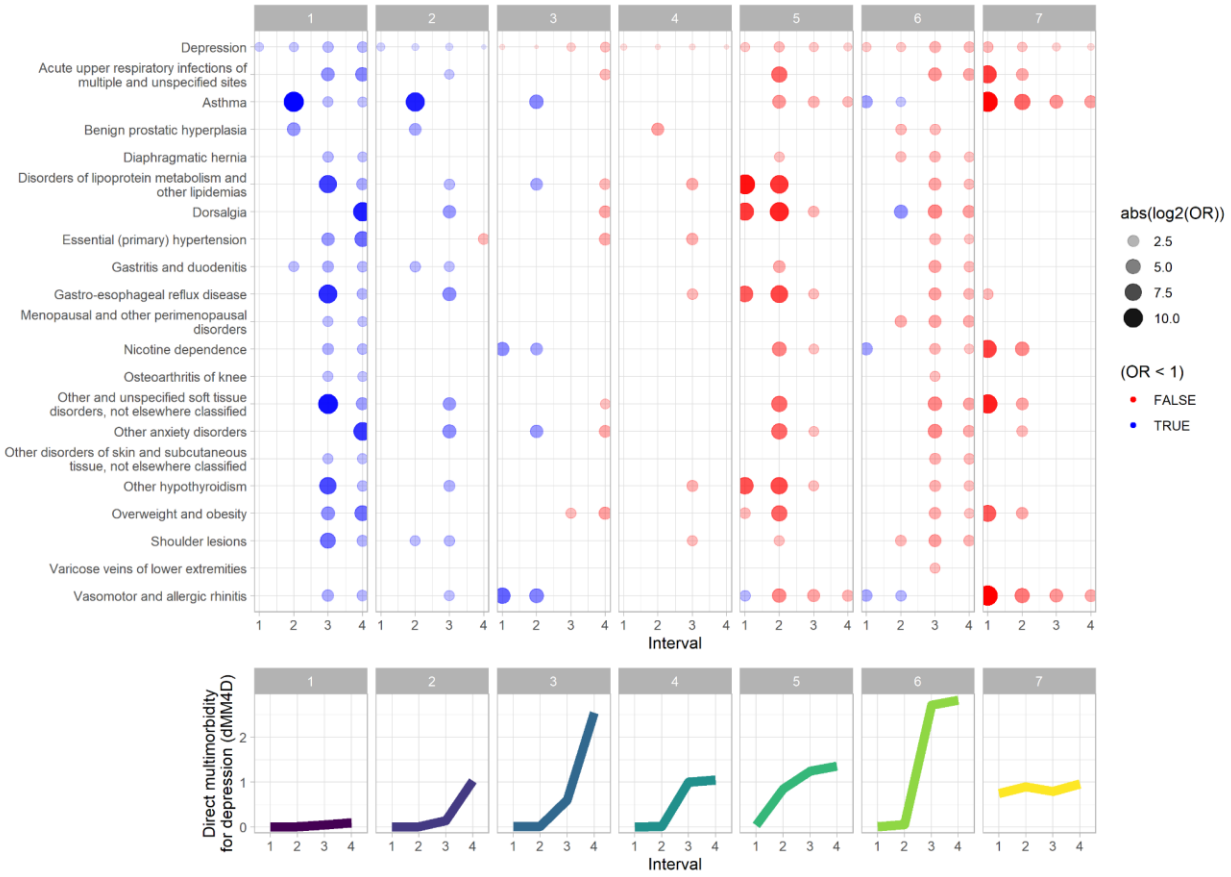


Figure 3B. Distribution of weighted direct MDD-related multimorbidity burden in the clusters where the number of clusters is set to 7. See the caption of Figure 1B for details.



141

142 **Figure 3C. The top over- and underrepresented diseases in each discrete time interval in the**

143 **clusters where the number of clusters is set to 7.** See the caption of Figure 1C for details.

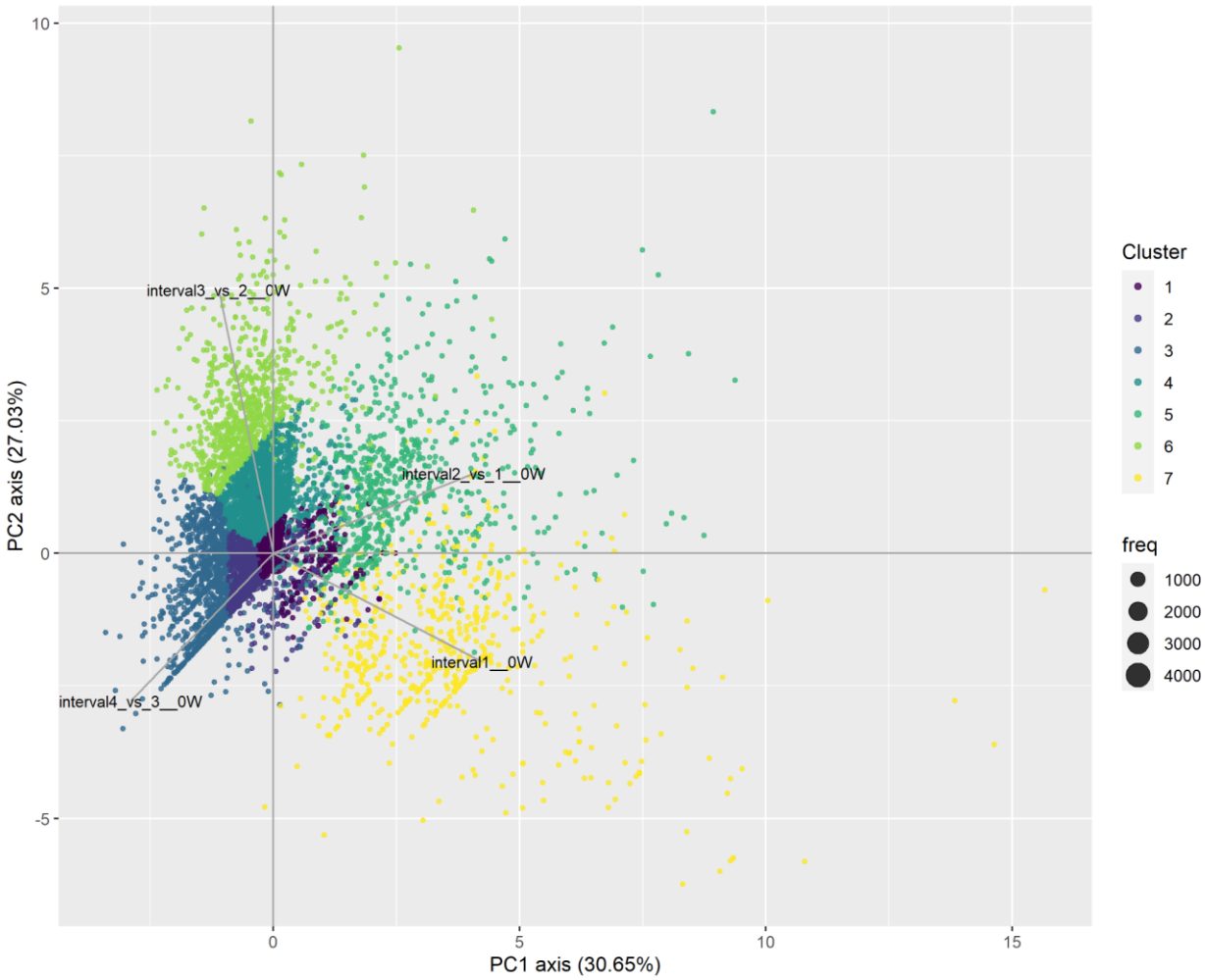


Figure 3D. The first two principal components of the 4-dimensional space defined by the temporal MDD-related multimorbidity burden score of the participants where the number of clusters is set to 7. See the caption of Figure 1D for details.

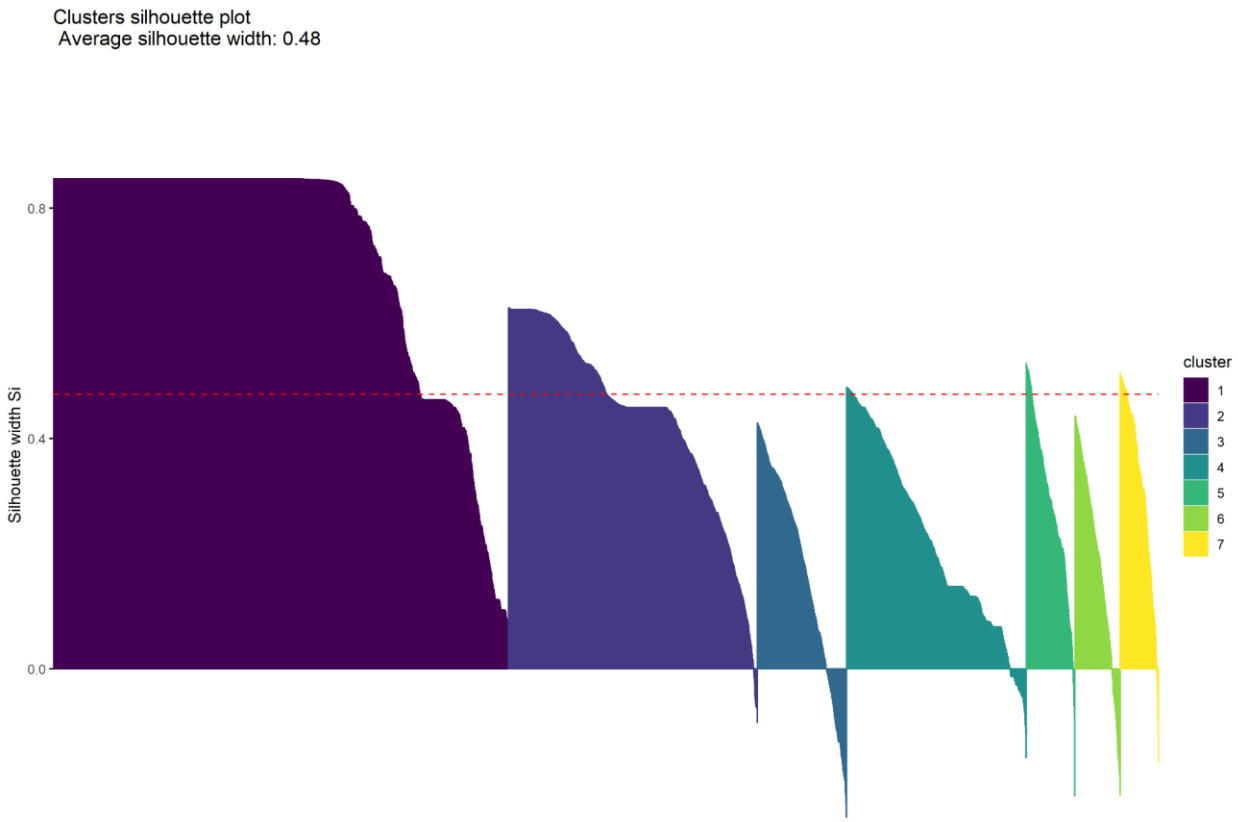


Figure 3E. Silhouette plot of the clusters where the number of clusters is set to 7. See the caption of Figure 1E for details.

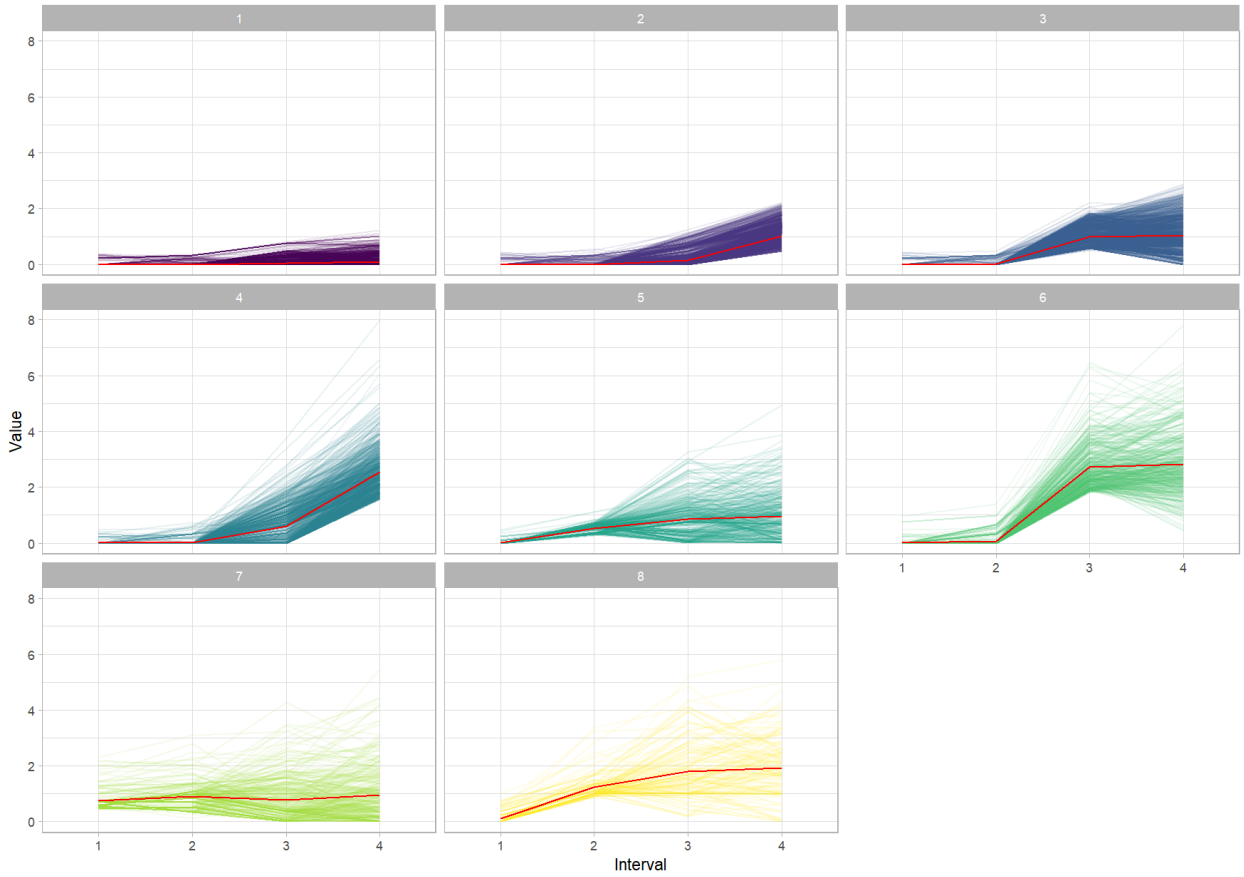


Figure 4A. Trajectories of the weighted direct MDD-related multimorbidity score over time using 8 clusters. See the caption of Figure 1A for details.

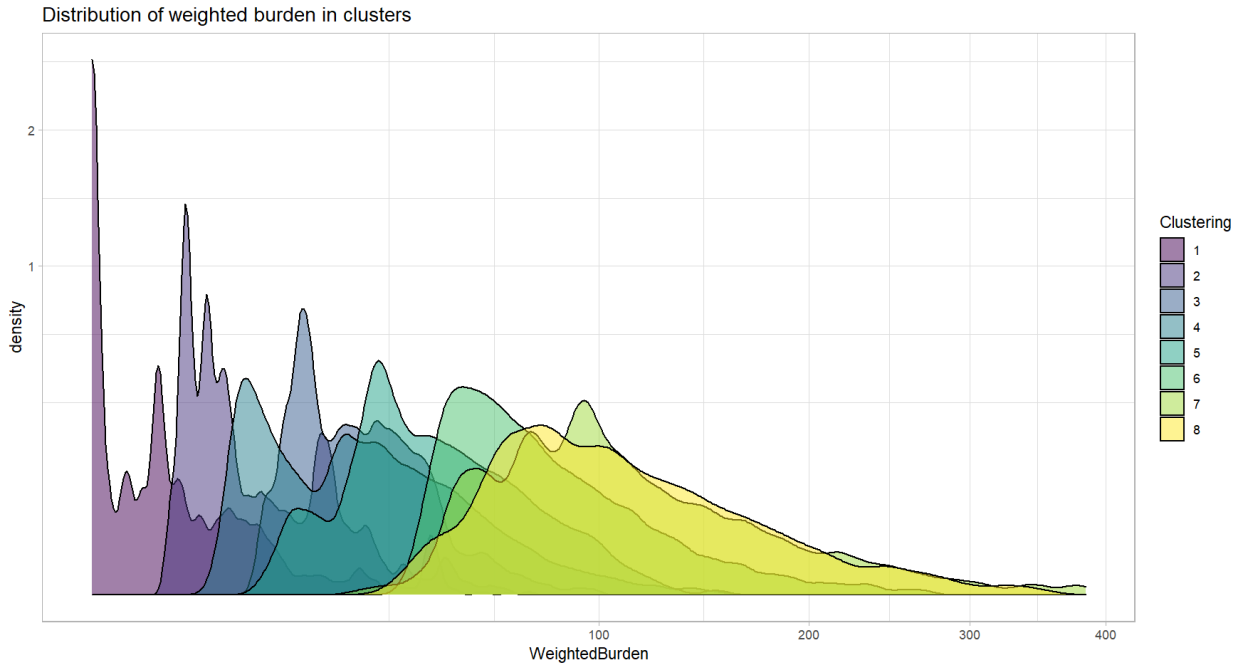
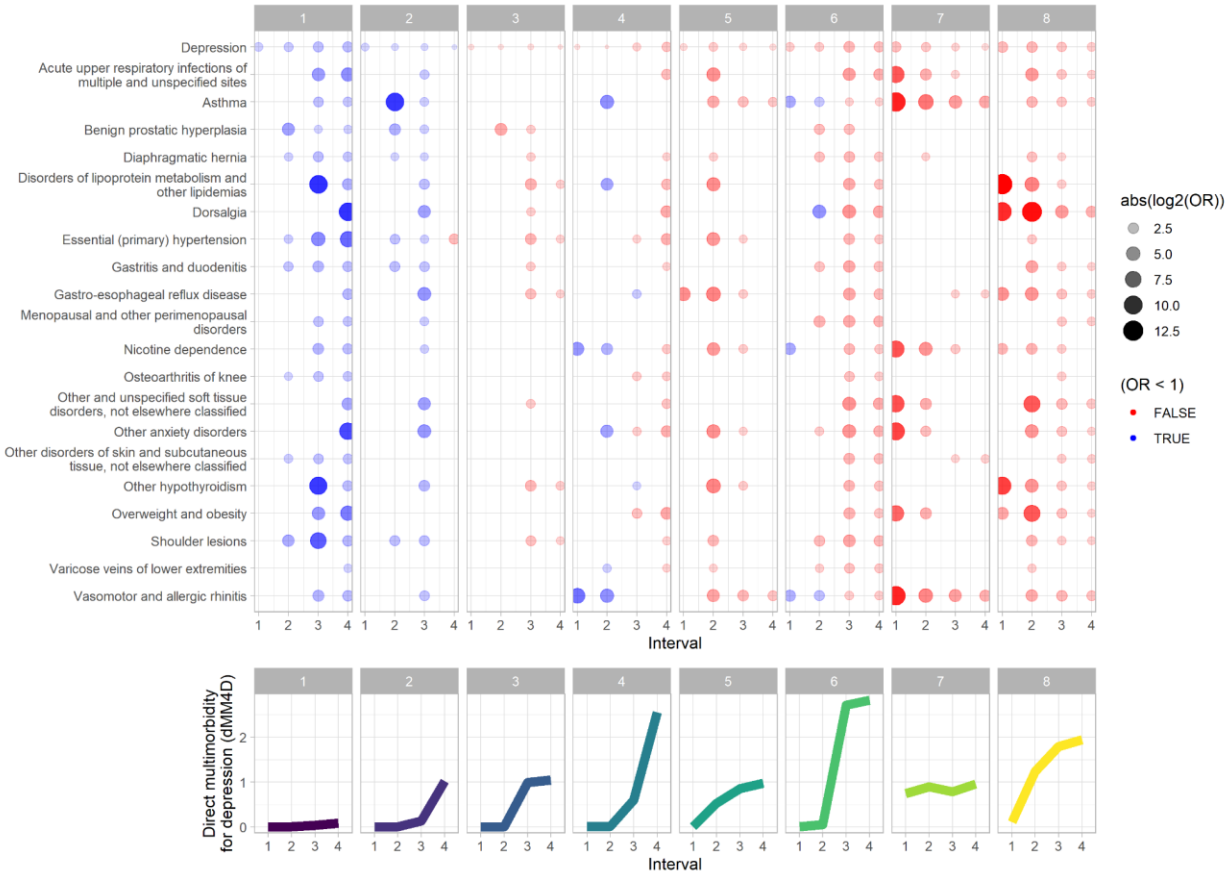


Figure 4B. Distribution of weighted direct MDD-related multimorbidity burden in the clusters where the number of clusters is set to 8. See the caption of Figure 1B for details.



158

159 **Figure 4C. The top over- and underrepresented diseases in each discrete time interval in the**

160 **clusters where the number of clusters is set to 8.** See the caption of Figure 1C for details.

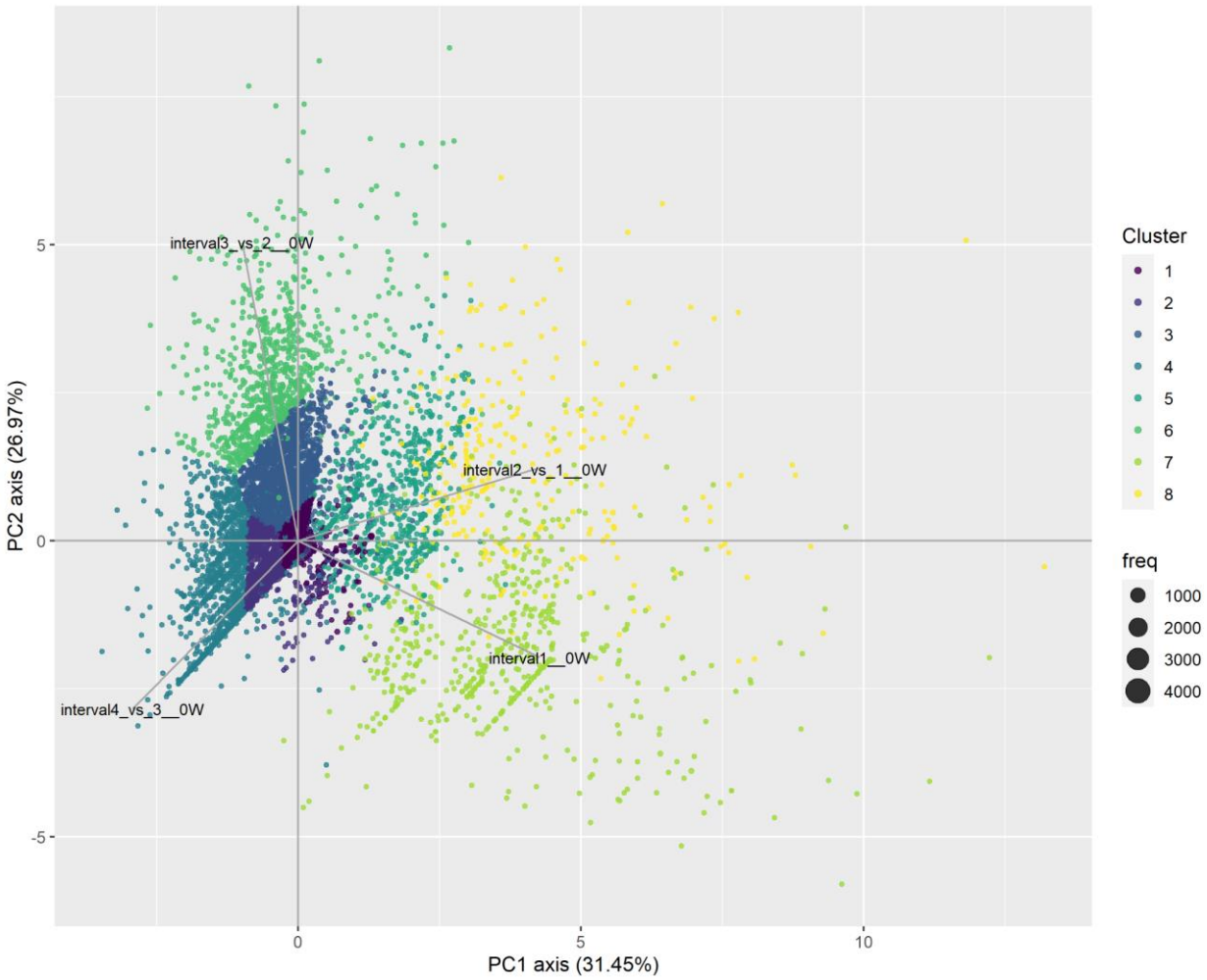
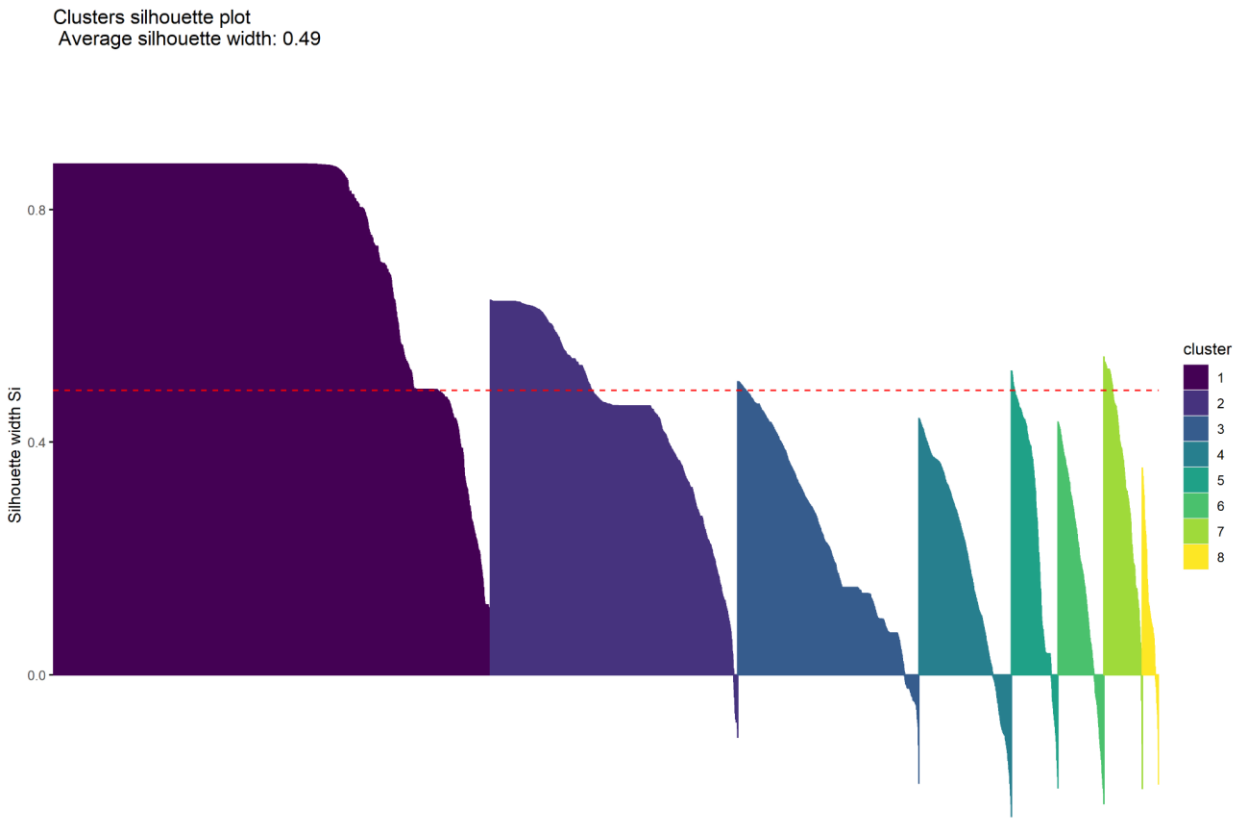


Figure 4D. The first two principal components of the 4-dimensional space defined by the temporal MDD-related multimorbidity burden score of the participants where the number of clusters is set to 8. See the caption of Figure 1D for details.



166

167 **Figure 4E. Silhouette plot of the clusters where the number of clusters is set to 8.** See the
168 caption of Figure 1E for details.

169

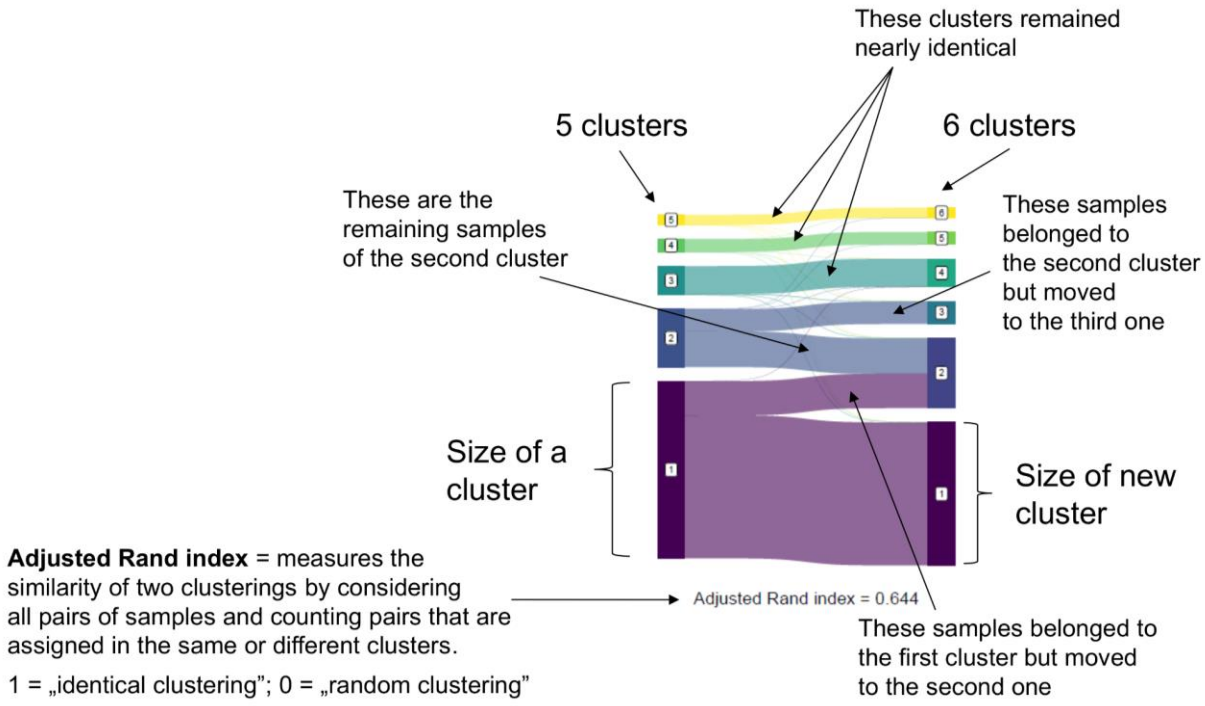


Figure 5A. Explanation of the Sankey plots visualizing the difference between the increasing cluster numbers. See Figure 5B for the differences between 5-6, 6-7, 7-8 and 8-9 cluster numbers.

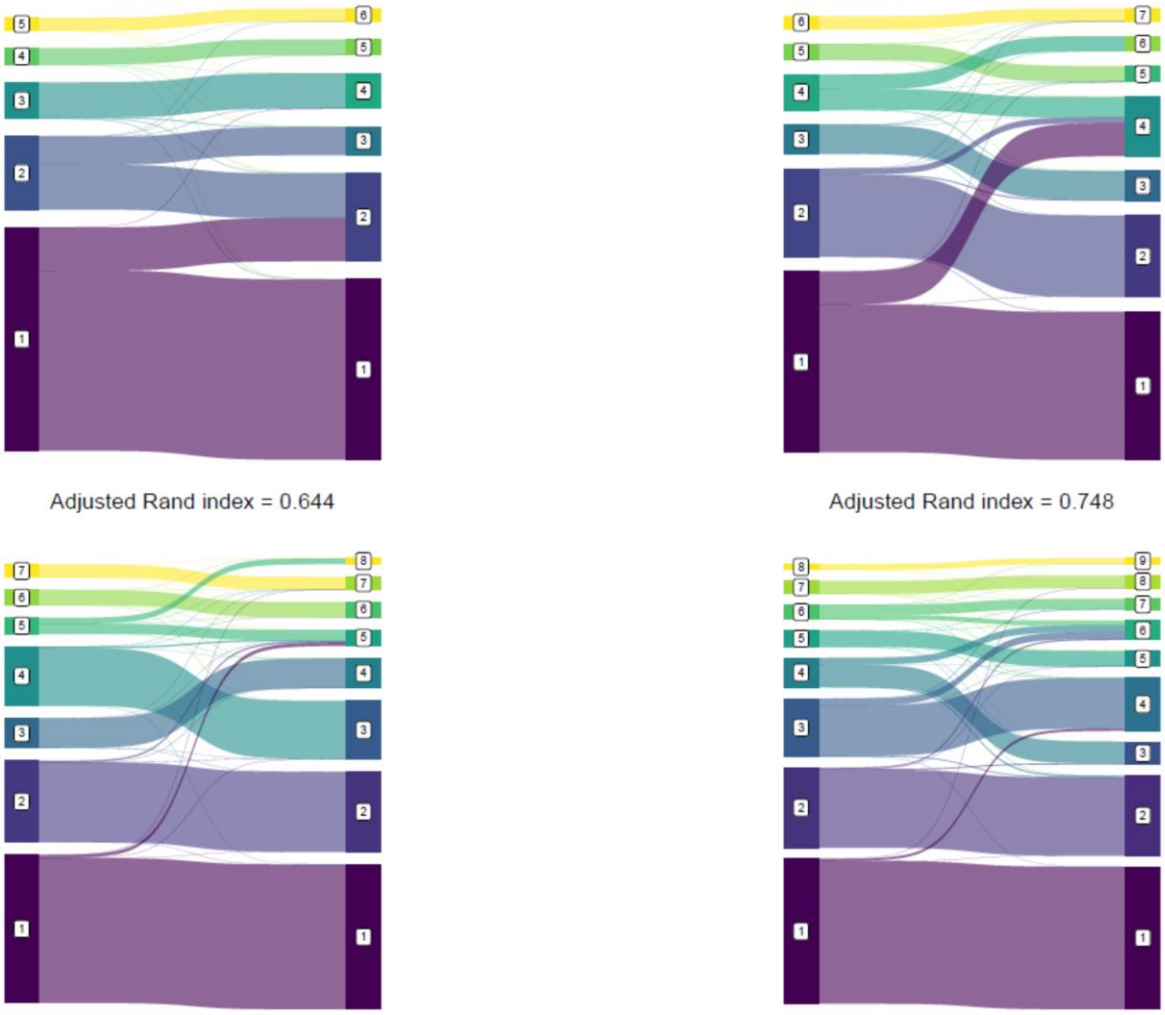


Figure 5B. The observed difference between the clusters by increasing the number of clusters. Top left: Increasing the number of clusters from 5 to 6. Top right: Increasing the number of clusters from 6 to 7. Bottom left: Increasing the number of clusters from 7 to 8. Bottom right: Increasing the number of clusters from 8 to 9.