Journal of Clinical Epidemiology 194 (2026) 112209

ORIGINAL RESEARCH

Cluster separation outperforms other metrics in validating multimorbidity patterns: statistical simulation study

Thamer Ba Dhafari[a], Alexander Pate[a], Glen P. Martin[a], James Rafferty[b], Farideh Jalali-najafabadi[c,d], Marlous Hall[e,f], Niels Peek[a,g,*]

[a]*Division of Informatics, Imaging & Data Sciences, School of Health Sciences, The University of Manchester, Manchester M13 9PL UK*
[b]*Faculty of Medicine, Health & Life Science, Population Data Science, Swansea University Medical School, Swansea University, Singleton Park, Swansea SA2 8PP, UK*
[c]*Faculty of Biology, Medicine and Health, Division of Cardiovascular Sciences, School of Medical Sciences, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK*
[d]*Manchester University NHS Foundation Trust, Manchester Heart Institute, Manchester, UK*
[e]*Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK*
[f]*Leeds Institute for Data Analytics, University of Leeds, Leeds, UK*
[g]*THIS Institute, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK*

Accepted 2 March 2026; Published online 6 March 2026

Abstract

Background and Objectives: Multimorbidity, defined as the presence of multiple long-term health conditions within an individual, remains a growing challenge in healthcare. Identifying frequently occurring multimorbidity clusters may help to develop targeted interventions and optimize care pathways. However, the validation of multimorbidity clusters derived from real-world data is complicated by the lack of a known ''ground truth.'' We conducted a statistical simulation study that aimed to evaluate the performance of three common validation approaches (cluster separation, clustering stability, and strength of association with health outcomes) in assessing the quality of multimorbidity clusters, where performance was measured by agreement with known ground truth clusters.

Methods: Simulated datasets with predefined clusters were generated across 25 scenarios, varying parameters such as disease prevalence, sample size, and noise levels. Latent class analysis was applied to derive clusters from the simulated data, which were compared to the predefined clusters using the adjusted rand index (ARI). The ARI served as our gold standard quality assessment of derived clusters.

Results: Cluster separation, measured by the Calinski−Harabasz index, showed the strongest agreement with our gold standard in most scenarios (median correlation: 0.641, IQR: 0.505−0.728). Clustering stability—assessed using resampling—had mixed performance, with a median correlation of 0.421 (IQR: 0.127−0.526). The strength of association with health outcomes, assessed using Nagelkerke's R^2, consistently showed poor agreement (median correlation: −0.424, IQR: −0.543 to −0.173) with the ARI.

Conclusion: Cluster separation seems to be the most reliable approach to validate multimorbidity clusters. Clustering stability can sometimes be used for validation but has limitations. Assessing the strength of association of multimorbidity clusters with health outcomes, though valuable for understanding clinical relevance, appears to not validate cluster quality despite being commonly used in published literature. © 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Multimorbidity; Analytical method; Cluster analysis; Validation; Simulation study; Latent class analysis

Funding: This study was supported by the UK Medical Research Council. This work was partially supported by the Medical Research Council (MR/S027750/1) and the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre (IS-BRC-1215-20007).

Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

* Corresponding author. The Healthcare Improvement Studies Institute (THIS Institute), Department of Public Health and Primary Care, University of Cambridge Strangeways Research Laboratory, 2 Worts Causeway, Cambridge CB1 8RN, UK.

E-mail address: niels.peek@thisinstitute.cam.ac.uk (N. Peek).

https://doi.org/10.1016/j.jclinepi.2026.112209
0895-4356/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

What is new?

Key findings

- Our simulation study evaluated how three common validation metrics for assessing the quality of multimorbidity clusters.
- Cluster separation emerged as the most reliable metric, showing strong agreement with our gold standard assessment.
- The strength of association with health outcomes, a commonly used cluster validation metric, consistently showed poor performance.

What this adds to what is known?

- First systematic evaluation of validation metrics against known ground truth in multimorbidity clustering

What is the implication and what should change now?

- Researchers should prioritize cluster separation rather than associations with health outcomes when validating multimorbidity clusters.

1. Introduction

Multimorbidity, defined as the coexistence of two or more long-term health conditions in an individual, presents a significant challenge for clinical management due to the complex needs of these patients and the associated resource burdens on health systems [1]. The UK Academy of Medical Sciences [1] suggested that long-term conditions often emerge in commonly recurring patterns or clusters across patients. Understanding these clusters could reveal relationships between long-term conditions and may help to redesign clinical services [2–5].

One of the clinical areas where multimorbidity clustering has proven to be useful is diabetes [6]. Complications affecting the cardiovascular system, nerves, eyes, kidneys, and skin frequently co-occur in diabetic patients due to shared underlying mechanisms. Recognizing these patterns of co-occurring complications took many years but eventually had a profound influence on clinical approaches toward prevention and management of diabetes. Similar clustering approaches have shown value in cardiovascular disease management and mental health care, where identifying patterns of co-occurring conditions could lead to improved prevention and management, reducing the burden on patients and healthcare services [7,8].

Various analytical methods exist to identify multimorbidity clusters, including latent class analysis (LCA), factor analysis, and hierarchical clustering [9]. A concerning finding is that these different methods can lead to varying and sometimes conflicting results [10], where different clustering methods applied to the same data identify different multimorbidity patterns [9]. This variability highlights the importance of evaluating how well clustering methods capture meaningful patterns in the data. In real-world multimorbidity research, there is no observable ground truth due to the complex nature of disease co-occurrence. This absence of ground truth is precisely why we conducted a simulation study, which allowed us to create artificial datasets with predefined ''true'' clusters. Like any data-driven approach, multimorbidity clustering analyses carry a risk of overfitting, potentially capturing random variation rather than meaningful clinical associations. Robust validation approaches are essential to ensure that derived clusters reflect important disease relationships and are not merely algorithmic artifacts. Our recent review [11] found that multimorbidity research has used a considerable variety of methods. These include assessing association of the clusters with clinical outcomes; considering the clinical plausibility of patterns; evaluating stability across different subsamples and methods; and exploring common determinants. We therefore focused out evaluation on the three most frequently used validation metrics: clustering stability, cluster separation, and associations with health outcomes.

To date, no studies have assessed how well these validation metrics discriminate between poor- and high-quality clusterings. Therefore, we conducted a simulation study to evaluate how accurately different validation metrics can assess the quality of empirically derived multimorbidity clusters. We assumed that true clusters exist in the data and that closeness to these true clusters indicates good clustering quality. The simulated datasets embedded predefined clusters which served as the ground truth, thus enabling assessments that are not possible with real-world data. The aim of this simulation study was to determine whether three commonly used multimorbidity validation metrics: clustering stability, cluster separation, and associations with health outcomes can accurately identify whether empirically derived clusters represent the predefined clusters in the simulated datasets. If these validation metrics show strong agreement with the ground truth, they can be reliably used for assessing the quality of empirically derived multimorbidity clusters in real-world studies where the ground truth is unknown.

2. Simulation study design

In this section, we describe the design of the simulation study which we report according to the aims, data-generating mechanisms, estimands/target of analysis, methods, and performance structure [12]. All simulation

code is available on GitHub[1]. Simulations were conducted on high-performance computers (the Computational Shared Facility) at the University of Manchester.

2.1. Data-generating mechanisms

2.1.1. Underlying structures

We designed the simulation to mimic a cross-sectional extract of a patient's health record; that is, we simulated data capturing the binary status of long-term conditions (0 for absent and 1 for present) at a given point in time. This simulation design allows for patients to have zero or more conditions, ensuring the data include individuals both with and without multimorbidity, which is defined as having two or more long-term conditions.

To generate the simulated data, we replicated the procedures described in Nichols et al [13] and used their source code. Nichols et al validated their simulation approach using real-world UK primary care data from over 23,000 patients with multimorbidity, ensuring the procedures generate realistic multimorbidity patterns. These procedures generate k simulated disease clusters that consist of N patients and 26 diseases (denoted a,b,…,z) resulting in a $N \times 26$ matrix of disease observations. In our study, each of the N patients were probabilistically assigned as a member of one of k disease clusters. We used a uniform model where each cluster had an equal probability of being selected, providing a baseline for uniform distribution across clusters. This is represented by the probability:

$$P(k) = \frac{1}{K}$$

Additionally, we included N_{out} patients who did not belong to any of the predefined clusters to better mimic the variability seen in real patient data. For these patients, the probability of having each disease was set to the overall prevalence of that disease in the simulated population, representing background disease prevalence independent of cluster membership. The total number of patients can thus be expressed as $N = N_{in} + N_{out}$, where N_{in} is the number of patients assigned to one of the k clusters, and N_{out} is the number of patients not assigned to any cluster.

Once patients were assigned to clusters, the number of diseases allocated to each cluster was randomly drawn from a Poisson distribution with an expected mean of 5 diseases per cluster, adjusted to ensure that there were at least two diseases per cluster. This reflected the variability typically found in published clusters related to multimorbidity. Diseases were assigned to clusters with equal probability, and it was allowed for single diseases to occur in multiple clusters.

For each patient in a disease cluster, the presence of diseases was simulated using a multinomial probit model with a 26×26 correlation matrix. Within each disease cluster k, the off diagonals of the correlation matrix between diseases were set to a positive correlation coefficient ρ_k. This correlation coefficient ρ_k was one of the parameters used to generate the simulated data and varied across different simulation scenarios. For example, diseases within cluster 1 may have a correlation of $\rho_k = 0.2$.

To add realism and account for potential data inaccuracies such as misdiagnosis or data entry errors, random noise P_{noise} was incorporated into the disease status assignment. Real-world clinical datasets often contain such imperfections, which can affect how well different validation methods perform. This noise was implemented by adjusting the probit threshold for each disease, simulating the likelihood of misclassification. Further details are provided in Appendix 1.

In line with terminology used in the unsupervised machine learning literature, we refer to the results obtained from applying a clustering method as "a clustering." That is, a clustering is a partitioning of a dataset, where each partition is called a cluster. The clustering that was used in the generation of simulated datasets was referred to as the "original clustering." This clustering varied from one simulated dataset to the next.

To each simulated dataset, we added a binary outcome variable (denoted by Y) which represented an adverse outcome related to multimorbidity. We assumed that the risk of adverse outcome was driven by the number of diseases (minimum 0, maximum 26) at patient level, and that there were no other (confounding) factors driving this risk. We calculated a risk score R_j for each patient, $j = 1,\ldots, N$, in the simulated dataset which quantified their disease burden:

$$R_j = \sum_{m=a}^{z} D_{jm}$$

where D_{jm} denotes the presence (1) or absence (0) of disease m for patient j. The risk scores were converted into probabilities of experiencing adverse outcomes using a logistic function:

$$P_j = \frac{1}{1 + e^{-(b_0 + b_1 R_i)}}$$

We chose b_0(intercept) $= -4$ and $b_1(slope) = -0.8$ to ensure the model captured the typically low risk of adverse outcomes for most patients in real-world datasets. This choice set a realistic baseline risk for healthy individuals and ensured the risk for most patients remained under 0.05, allowing for a consistent, moderate increase in risk with each additional disease. Finally, we performed a Bernoulli trial for each patient, using $E[Y_j] = P_j$ to determine their health outcome Y_j.

[1] https://github.com/thamer-badhafari/MVMs-SimulationStudy.

Importantly, this outcome model was deliberately designed to be driven by overall disease burden rather than specific cluster membership patterns. This design choice allows us to test a critical question: can outcome associations misleadingly suggest disease clustering even when the outcome-generating mechanism is independent of such clusters? In real-world multimorbidity research, the true relationship between disease patterns and outcomes is unknown, making it impossible to verify whether observed outcome associations reflect genuine cluster-outcome relationships or spurious correlations.

2.1.2. Process for simulation

Our evaluation of the validation metrics used for multimorbidity clusters involved several steps, as outlined in Table 1 (further details in Appendix 1). After generating simulated datasets, we selected patients with multimorbidity (ie, patients with two or more conditions) for analysis. We then applied LCA clustering algorithm to these datasets to derive empirical clusters. We selected LCA based on findings from Nichols et al [13], who compared four clustering algorithms and found that LCA achieved the highest adjusted rand index (ARI) when evaluated against predefined clusters in simulated datasets. Full computational details are provided in Appendix 2. The resulting clusters were evaluated using three validation metrics: clustering stability, cluster separation, and strength of association with health outcomes (see Section 2.4).

To explore a range of possible scenarios for our study, we started with a baseline set of default parameter values, ensuring a controlled environment for each simulation scenario. Then, we introduced variability by adjusting one parameter at a time while keeping others constant (Table 2). Each parameter had a predefined range of possible values, including the default value. Disease prevalence P_k and proportion of random disease observations P_{noise} were given using probit values and transformed to probabilities. For P_k, probit values $\{-2.0, -1.5, -1.0, 0, 1, 1.5, 2.0\}$ corresponded to prevalences $\{0.023, 0.067, 0.159, 0.500, 0.841, 0.933, 0.977\}$, with negative values yielding lower than average prevalence. For P_{noise}, probit values $\{-5.0, -4.0, -3.0, -2.0, -1.0\}$ corresponded to noise probabilities $\{0.0000003, 0.00003, 0.001, 0.023, 0.159\}$, with more negative values indicating less random noise. Throughout the results, we report both probit values and their corresponding probabilities for these two parameters.

For each parameter set (scenario), we generated 1000 simulated datasets. Within each simulated dataset, we applied the clustering algorithm to derive empirical clusters. There were 24 unique parameter sets for the simulated datasets, along with one default scenario, resulting in a total of 25 scenarios.

Table 1. Full simulation procedure

Algorithm 1
1: Algorithm assess validation metrics
2: Input: simulation parameters (Table 1), number of simulations (m = 1000), clustering algorithm (LCA).
3: For each parameter setting in Input:
4: For i = 1,…,m:
5: Generate simulated dataset d_i using the approach by Nichols et al [13] outlined in section 2.1.1.
6: Apply clustering algorithm to d_i (section 2.1.2)
7: Calculate correspondence between the original and derived clusters using ARI (section 2.2)
8: Apply validation metrics (section 2.3)
9: Compute summary statistics for quality scores for all validation metrics (section 2.4)
10: Assess the correlation between the correspondence scores (ARI) and clustering quality scores (section 2.4)

ARI, adjusted rand index.

2.2. Estimands and target of analysis

The target of the simulation study is to evaluate the agreement between (a) the measure of correspondence between original (ground truth) clusters and empirically derived clusters and (b) the quality scores from three validation metrics (see section 2.3). For (a), we used the ARI, a statistical measure used to evaluate the similarity between two clustering results. The ARI serves as our gold standard quality assessment measure as it leverages the known ground truth cluster assignments.

ARI values range from −1 to 1. A value of 1 indicates identical clusterings, 0 indicates independent clusterings (agreement by chance), and negative values suggest less agreement than expected by chance (see Appendix 2 for more details).

To quantify agreement between the ARI and the quality scores from the validation metrics (section 2.3), we use the Pearson correlation coefficient. A strong positive correlation indicates that the validation metric in question effectively reflects correspondence between original (ground truth) clusters and empirically derived clusters. Conversely, a weak or negative correlation would suggest that the validation metrics does not reflect this. See section 2.4 for more details on this comparison.

2.3. Methods for analysis

2.3.1. Clustering stability

Clustering stability is a common validation method that evaluates the consistency of clusters across different subsamples [11]. We implemented this using a two-stage resampling approach. First, we created B=400 bootstrap samples from our simulated dataset (S), where each bootstrap sample is denoted as S_b *and* $b \in \{1,2,\cdots,B\}$. Second, we randomly split each bootstrap sample S_b into two equal subsets $S_{b,1}$ and $S_{b,2}$, applied clustering algorithm to both subsets, and

Table 2. A summary of the parameters that were varied in the simulations

Parameter	Description	Ranges and default values	Number of options
1. General simulation parameters: these parameters define the overall structure and scale of the simulated datasets.			
Number of ''true'' clusters k:	Represents the number of disease clusters in the dataset.	{2,3,...,8}; default value: 3	7
Total number of patients N	Total number of patients in the dataset.	{1000, 10000, 50000, 100000}; default value: 10,000	4
Proportion of patients not in a cluster N_{out}	Proportion of patients that do not belong to any cluster.	{0.001, 0.05, 0.25}, where 0·001% means that almost all patients in the data belong to clusters; default value: 0.05	3
2. Cluster-specific parameters: these parameters are specific to each cluster, determining the internal characteristics within clusters.			
Within-cluster disease correlation ρ_k	Indicates how often diseases co-occur within clusters.	{0.3, 0.5, 0.7, 0.9}; default value: 0.5	4
Prevalence of diseases P_k	Indicates how common each disease is within a cluster.	{−2.0, −1.5, −1.0, 0, 1, 1.5, 2.0} (probit scale); negative values mean less common, zero means average, and positive values means more common than average; default value: 0	7
The proportion of random, uncorrelated disease observations P_{noise}	Indicates the level of unexplained disease occurrences, or ''noise,'' in the data.	{−5.0, −4.0, −3.0, −2.0, −1.0} (probit scale); where 5.0 means high randomness and −1.0 means low randomness; default value: −3.0	5

transferred the resulting models between them (see Appendix 3 for more details). We then used the ARI to compare the initial and transferred cluster assignments for each subset, averaging these values to quantify overall stability.

While our analysis uses ARI both to assess stability and to compare clusters against ground truth, these applications measure distinct properties: internal stability and external validity, respectively. There is no information leakage between these applications as they evaluate independent aspects of clustering performance. Using ARI for both purposes provides complementary insights into cluster quality without unfairly favoring the stability metric over the other validation measures.

2.3.2. *Cluster separation*

The second validation metric was assessing the separation between the empirically derived clusters. We used the Calinski−Harabasz index (CHI) measure to evaluates the nonoverlapping nature of the identified clusters. CHI is a quality measure as it compares the coherence of clusters (between-cluster variance) to their separation (within-cluster variance). It is calculated by dividing the sum of the squared deviations between the cluster means and the grand mean by the sum of the squared deviations within each cluster. The CHI is a non-negative value, with higher values indicating better clustering. We refer to Appendix 5 for further details.

2.3.3. *Association of multimorbidity clusters with outcomes*

The third method of validation we evaluated in this study involves examining the strength of the relationship between multimorbidity clusters and health outcomes, an approach that is increasingly used in multimorbidity studies [11]. This relationship was assessed through the binary outcome variable Y added to each simulated dataset, representing an adverse outcome related multimorbidity severity (number of conditions), as described in Section 2.1.1.

We then applied logistic regression analysis to assess the association between these simulated outcomes and the empirically derived clusters in our study. The logistic regression model is formulated as follows:

$$logit(P(Y = 1)) = \beta_0 + \beta_1 X_i$$

where X_i is a categorical variable representing the cluster to which patient i belongs; and β_1 is the coefficient for the cluster membership variable, providing a risk estimate for each cluster.

Since the clusters do not overlap, each patient is assigned to exactly one cluster, and this membership is represented by the categorical variable X_i. This nonoverlapping nature of clusters allows us to estimate one risk value per cluster directly. For instance, if there are five clusters, the model estimates the risk associated with each of these clusters relative to the baseline.

We note that this simulated analysis assumed that adjustment for confounders was not necessary (see Section 2.1.1). In a real-world study, one would probably have to adjust for confounders such as age and sex. Without loss of generality, we omitted this step from our study.

To quantify the strength of these associations, we used Nagelkerke's R^2, a metric that ranges from 0 to 1, providing a measure of how much of the observed variance in

outcomes is explained by the model, with higher values indicating stronger associations.

2.4. Performance metrics

Each of three validation metrics described above produces a score (a number between 0 and 1) that is intended to reflect the quality of the empirically derived clustering being validated. We assessed the agreement between these scores and our gold standard quality assessment of the derived clusterings, that is, the ARI value quantifying the correspondence between the original and derived clusterings. To establish this agreement, we calculated the Pearson correlation coefficient (r) within each simulation scenarios, across the 1000 iterations. To compare correlation coefficients, we used the Zou's confidence interval method [14] in the ''cocor'' R package [15].

3. Results

3.1. Default simulation scenario

Figure presents the results from the default scenario. Across 1000 datasets, the average number of multimorbid patients was 7165 (SD = 795), with an average execution time was 10.77 minutes (SD = 2.45) per dataset (all execution times are provided in Appendix 5). Clear patterns emerged in how validation metrics related to cluster quality.

The gold standard quality measure (correspondence ARI) had a mean of 0.773 (SD = 0.091), indicating that empirically derived clusters captured approximately 77% of the true cluster structure on average. Cluster separation (CHI), clustering stability (ARI), and outcome association (Nagelkerke's R^2) had means of 1587.83 (SD = 397.26), 0.876 (SD = 0.052), and 0.032 (SD = 0.024), respectively. Figure shows the relationships between correspondence ARI and each validation metric, displaying positive trends for cluster separation (r = 0.790) and stability (r = 0.435), and a negative trend for outcome association (r = −0.507).

3.2. Performance of validation metrics across simulation scenarios

Table 3 presents the performance of the three validation metrics across all 25 simulation scenarios, with detailed results in Appendices 6 and 7. Cluster separation demonstrated the strongest agreement with our gold standard quality assessment (correspondence ARI), with a median correlation of 0.641 (IQR: 0.505 − 0.728). It performed best in scenarios with a low proportion of patients not assigned to clusters and in the default scenario. Clustering stability showed weaker agreement with the gold standard than cluster separation, with significantly lower correlations than separation in 21 out of 25 scenarios (Table S5). Stability performed best in scenarios where disease prevalence was low (negative P_k). The strength of association with outcomes demonstrated poor agreement with ARI across all scenarios, with a median correlation of −0.424 (IQR: −0.543 to −0.173). Notably, 22 out of 25 scenarios exhibited negative correlations. The correlation between Nagelkerke's R^2 and ARI was consistently weaker than both CHI and stability ARI across all scenarios (Table S5).

4. Discussion

Our study evaluated three approaches for validating multimorbidity clusters derived from simulated patient data. We assessed cluster separation, clustering stability, and association with outcomes across 25 diverse scenarios,

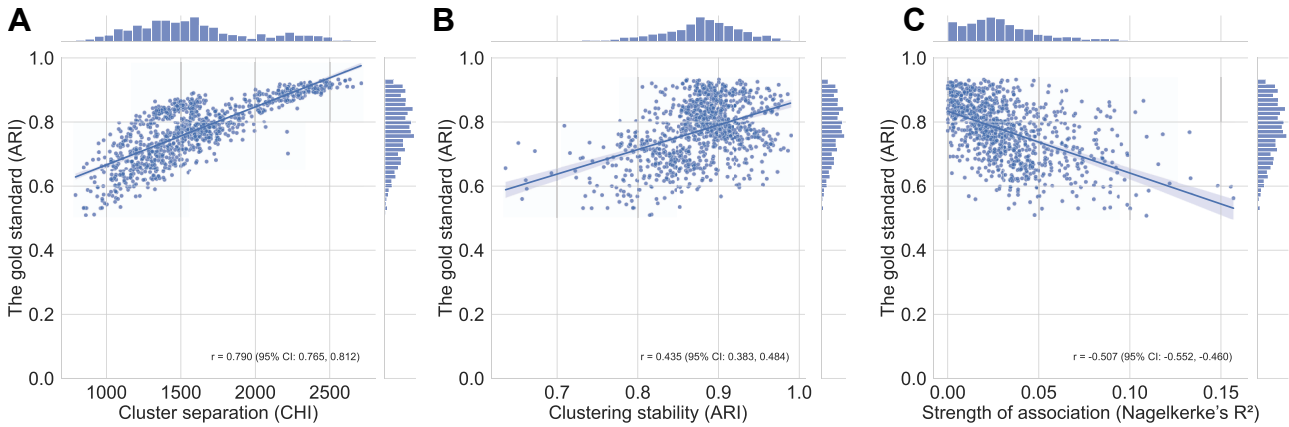

Figure. Relationships between validation metrics and ground truth cluster quality (ARI) for the default simulation scenario. A, shows a strong positive relationship between cluster separation (CHI) and ARI (r = 0.790; 95% CI: 0.765–0.812), indicating that higher separation reliably corresponds to better cluster quality. B, displays a moderate positive relationship between clustering stability (ARI) and ARI (r = 0.435; 95% CI: 0.383–0.484). C, presents a moderate negative relationship between outcome association strength (Nagelkerke's R^2) and ARI (r = −0.507; 95% CI: −0.552 to −0.460), demonstrating that stronger outcome associations paradoxically indicate lower cluster quality. Each figure includes marginal histograms showing distributions and a fitted regression line (blue) with 95% confidence bands. Each scatter plot represents 1000 iterations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3. Summary of statistics across 25 simulation scenarios

Parameter name	Parameter value	Correlation between ARI and separation CHI [%95 CI]	Correlation between ARI and stability ARI [%95 CI]	Correlation between ARI and Nag R^2 [%95 CI]
Default scenario		0.790 [0.765–0.812]	0.435 [0.383–0.484]	−0.507 [−0.552 to −0.460]
Within-cluster disease correlation ρ_k	0·3	0.771 [0.745–0.795]	0.554 [0.509–0.595]	−0.424 [−0.474 to −0.372]
	0·7	0.735 [0.705–0.762]	0.506 [0.458–0.551]	−0.548 [−0.590 to −0.504]
	0·9	0.728 [0.697–0.756]	0.506 [0.459–0.551]	−0.530 [−0.573 to −0.484]
Number of ''true'' clusters k	2	0.747 [0.719–0.773]	0.546 [0.500–0.588]	−0.554 [−0.595 to −0.509]
	4	0.587 [0.544–0.626]	0.290 [0.232–0.345]	−0.368 [−0.421 to −0.313]
	5	0.557 [0.512–0.598]	0.177 [0.116–0.236]	−0.222 [−0.280 to −0.162]
	6	0.542 [0.496–0.584]	0.103 [0.041–0.164]	−0.185 [−0.245 to −0.125]
	7	0.440 [0.389–0.489]	0.048 [−0.014 to 0.110]	−0.173 [−0.233 to −0.112]
	8	0.468 [0.418–0.515]	0.115 [0.054–0.176]	−0.184 [−0.243 to −0.123]
Prevalence of diseases P_k	−2 (P = 0·023)	0.445 [0.394–0.493]	0.526 [0.479–0.569]	0.030 [−0.032 to 0.092]
	−1.5 (P = 0·067)	0.595 [0.553–0.633]	0.751 [0.723–0.777]	−0.101 [−0.162 to −0.039]
	−1 (P = 0·159)	0.708 [0.676–0.738]	0.739 [0.709–0.766]	−0.134 [−0.195 to −0.073]
	1 (P = 0·841)	0.760 [0.732–0.785]	0.405 [0.352–0.456]	−0.716 [−0.745 to −0.685]
	1.5 (P = 0·933)	0.692 [0.658–0.723]	0.240 [0.181–0.298]	−0.649 [−0.684 to −0.612]
	2 (P = 0·977)	0.562 [0.517–0.604]	0.127 [0.064–0.188]	−0.496 [−0.542 to −0.447]
Patients not in a cluster N_{out}	0·001	0.794 [0.770–0.816]	0.433 [0.382–0.482]	−0.521 [−0.565 to −0.474]
	0·25	0.649 [0.611–0.683]	0.638 [0.599–0.673]	−0.356 [−0.409 to −0.300]
Total number of patients N	1000	0.505 [0.448–0.559]	0.457 [0.396–0.514]	0.016 [−0.059 to 0.090]
	50,000	0.388 [0.334–0.439]	0.086 [0.024–0.147]	−0.605 [−0.643 to −0.564]
	100,000	0.136 [0.074–0.196]	−0.110 [−0.171 to −0.048]	−0.689 [−0.720 to −0.655]
The proportion of random, uncorrelated disease observations P_{noise}	−5 (P = 0.0000003)	0.141 [0.080–0.201]	−0.676 [−0.709 to −0.641]	−0.543 [−0.586 to −0.498]
	−4 (P = 0.00003)	0.654 [0.617–0.688]	0.421 [0.369–0.471]	−0.496 [−0.541 to −0.448]
	−2 (P = 0.023)	0.725 [0.695–0.753]	0.597 [0.556–0.636]	−0.084 [−0.146 to −0.022]
	−1 (P = 0.159)	0.641 [0.598–0.681]	0.292 [0.225–0.355]	0.162 [0.092–0.230]
Median (IQR) across all scenarios		0.641 (0.505–0.728)	0.421 (0.127–0.526)	−0.424 (−0.543 to −0.173)

ARI, adjusted rand index; CHI, Calinski–Harabasz index.
Each row represents a different parameter setting, with the default scenario shown first.

varying parameters such as disease prevalence, sample size, and noise levels. To mimic real-world conditions, we created simulated datasets with predefined patient clusters as our ''ground truth'' for comparison.

Using ARI as our gold standard quality measure, we found varying performance among the three approaches. Cluster separation showed strongest agreement with our gold standard, outperforming other methods in 21 out of 25 scenarios, with a median correlation of 0.641 (range: 0.136–0.794). Clustering stability showed more variable performance, with a median correlation of 0.421 (range: −0.676 to 0.751), occasionally outperforming cluster separation in specific contexts. Association of patient clusters with health outcomes consistently showed poor agreement with our gold standard, with a median correlation of −0.424 (range: −0.716 to 0.162), with 22 out of 25 scenarios exhibiting negative correlations.

From a clinical perspective, our findings demonstrate that deriving multimorbidity clusters (for instance, in the field of diabetes) without robust validation can lead to false positive results—leading to erroneous beliefs about the co-occurrence of long-term conditions, and therefore potentially to unnecessary preventative treatment and management. Moreover, this risk still exists when a validation method is used that poorly correlates to the ground truth, such as considering the association of clusters with health

outcomes. Such an approach may incorrectly reinforce the belief that certain conditions are likely to co-occur, when this was actually a chance finding. By prioritizing validation methods like cluster separation, researchers and clinicians can be more confident that the clusters they find reflect genuine associations between conditions.

Our study extends existing research on validating multimorbidity clusters, providing a more comprehensive comparison in multimorbidity research than previous studies [16–18]. Our direct comparison of multiple validation metrics within a controlled simulation environment expands on studies investigating individual validation approaches [19–21]. While some researchers have suggested outcome-based measures for cluster validation [22–24], our results demonstrate that Nagelkerke's R^2 shows consistently poor agreement with our gold standard, challenging previous assumptions about its reliability.

A key consideration in interpreting these findings is the relationship between simulated and real-world clustering. While our simulation assumes the existence of ''true'' clusters drive the underlying data generation mechanisms, the reality of multimorbidity clustering is more complex. In clinical practice, multimorbidity clustering results from complex disease interactions, with one condition causing another or multiple conditions sharing a common cause. The concept of ''true'' underlying clusters in such systems is debatable, but this abstraction was necessary to evaluate validation metrics in a controlled environment. The negative correlations observed between outcome association strength and our gold standard measure highlight complex relationships between these metrics, challenging the suitability of using outcome associations for validation.

It is also important to consider the design of our outcome generation model when interpreting our findings. We generated outcomes based on total disease count rather than cluster-specific patterns, which may appear to disadvantage outcome-based validation. However, this design choice was intentional and reveals a fundamental problem with using outcome associations for cluster validation. Our results showed that even when outcomes follow disease burden (which well-separated clusters should capture), Nagelkerke's R^2 values varied independently of cluster quality. This demonstrates that researchers will not be able to distinguish between genuine cluster-outcome relationships and spurious associations without knowledge of the true outcome-generating mechanism knowledge that is unavailable in real-world studies. We investigated whether the strongly negative $NagR^2$ correlations related to empirical cluster selection. However, the correlation between empirically selected k and $NagR^2$ performance was weak ($r = -0.085$), suggesting the negative relationship reflects a fundamental limitation of outcome-based validation rather than an artifact of model selection. The paradoxical finding that stronger outcome associations can indicate worse cluster quality likely occurs because outcome relationships can arise from factors unrelated to cluster validity such as overall disease burden, shared risk factors, or spurious correlations. This demonstrates that without knowledge of the true outcome-generating mechanism (unavailable in real-world research), researchers cannot reliably use outcome associations to validate cluster quality. While more complex outcome models incorporating cluster-specific risks might show different patterns, our findings highlight that outcome associations alone cannot validate cluster quality, as they may reflect factors unrelated to how well clusters capture true disease patterns. This limitation is intrinsic to outcome-based validation rather than specific to our simulation design.

These insights have implications for multimorbidity clustering research. Our findings discourage using the association with health outcomes to validate multimorbidity clusters, as this approach consistently performed poorly. However, examining such associations remains important for demonstrating clinical relevance. Recent studies have shown that multimorbidity clusters can offer more comprehensive risk assessments than individual disease evaluations by uncovering distinct health risks [25–27]. This highlights the clinical utility of multimorbidity clusters while confirming that association strength does not indicate how well derived clusters reflect true underlying disease clusters. Instead, cluster separation emerged as more suitable for validation, while stability measures showed variability in performance across different dataset characteristics.

Practical interpretation of cluster separation requires context-dependent interpretation. While cluster separation (CHI) emerged as the most reliable validation metric in our study, CHI values are context-dependent and vary with sample size, number of clusters, and data characteristics. Universal thresholds are therefore impractical. Instead, we recommend comparative interpretation within each analysis. When evaluating different clustering solutions for the same dataset, compare CHI values to identify solutions with substantially better separation. Quantify uncertainty by calculating bootstrap confidence intervals around CHI values. Check minimum class proportions to ensure no clusters are vanishing. Finally, compare observed CHI against null benchmarks (eg, random permutations) to assess whether separation exceeds chance levels. This relative approach provides more meaningful guidance than seeking absolute cutoffs.

Practically, validated multimorbidity clusters can inform integrated care pathways and health policy decisions. For example, if a specific cluster is found to include cardiovascular disease, clinicians could proactively target cardiovascular prevention in those patients that belong to that cluster. Health systems can also decide to allocate resources to multimorbidity clusters, for instance, by setting up integrated services that address all conditions in a prevalent cluster—such as cardiovascular prevention in diabetes clinics.

The study strengths include using simulated datasets with predefined clusters, enabling controlled evaluation of

different validation metrics. By assuming predefined ''true'' clusters, we established an objective benchmark for assessing correspondence between derived clusters and original structures.

Limitations include our binary representation of disease status, which simplifies disease complexity but aligns with common practices in multimorbidity studies. We used only LCA, though our evaluation of final assignments (rather than the clustering process itself) should generalize to other methods since all widely used algorithms fundamentally optimize cluster separation. We fixed the number of diseases at 26, which is typical for multimorbidity research (systematic reviews report median 17, IQR 11−23 [28]). Varying sample sizes (1000−100,000) with fixed diseases tests equivalent dimensional relationships to varying disease counts with fixed samples. The consistent findings across 25 diverse scenarios indicate our results are robust to parameter variations most relevant for validation metrics. The different scales of the metrics complicate interpretation: CHI values range from 0 to over 1,000, stability ARI values from −1 to 1, and Nagelkerke's R^2 from 0 to 1. Other metrics like the Dunn index and Davies−Bouldin Index also have unbounded ranges and pose similar challenges. Our stability assessment used bootstrap samples half the size of the original dataset. This may reduce the number of patients per cluster and can lower stability estimates, particularly for rare clusters. Despite this conservative approach potentially disadvantaging the stability metric, cluster separation still outperformed stability in most scenarios, especially in large-sample settings where even half-samples provided adequate cluster representation.

In conclusion, this study examined validation metrics for multimorbidity clusters using simulated datasets. Cluster separation seems to be the most reliable approach to validate multimorbidity clusters, showing stronger correspondence with ground truth clustering than other metrics across most scenarios. Clustering stability can sometimes be used for validation but has limitations, as it showed variable performance depending on dataset characteristics. Assessing the strength of association of multimorbidity clusters with health outcomes, though valuable for understanding clinical relevance, appears to not validate cluster quality despite being commonly used in published literature. These insights may inform future multimorbidity research, particularly regarding the widespread practice of using outcome associations for validation. These findings contribute to ongoing research on multimorbidity methodologies, highlighting the importance of carefully selecting validation methods.

CRediT authorship contribution statement

Thamer Ba Dhafari: Writing − review & editing, Writing − original draft, Software, Methodology, Formal analysis, Conceptualization. **Alexander Pate:** Writing − review & editing, Methodology. **Glen P. Martin:** Writing − review & editing, Methodology. **James Rafferty:** Writing − review & editing. **Farideh Jalali-najafabadi:** Writing − review & editing. **Marlous Hall:** Writing − review & editing, Methodology. **Niels Peek:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors have no conflicts of interest to declare.

Acknowledgments

The authors would like to acknowledge the assistance given by Research IT and the use of the Computational Shared Facility at The University of Manchester.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112209.

Data availability

This study uses simulated data from custom code available on GitHub. The public code enables reproduction of all datasets and analyses. No real-world patient data was used.

References

[1] Academy of Medical Sciences. Multimorbidity: a priority for global health research. London: Academy of Medical Sciences; 2018. Available at: https://acmedsci.ac.uk/file-download/82222577. Accessed March 27, 2026.

[2] Crowe F, Zemedikun DT, Okoth K, Adderley NJ, Rudge G, Sheldon M, et al. Comorbidity phenotypes and risk of mortality in patients with ischaemic heart disease in the UK. Heart 2020;106(11):810−6. https://doi.org/10.1136/heartjnl-2019-316091.

[3] Juul-Larsen HG, Christensen LD, Bandholm T, Andersen O, Kallemose T, Jørgensen LM, et al. Patterns of multimorbidity and differences in healthcare utilization and complexity among acutely hospitalized medical patients (≥65 years) - a latent class approach. Clin Epidemiol 2020;12:245−59. https://doi.org/10.2147/CLEP.S226586.

[4] Kuan V, Denaxas S, Patalay P, Nitsch D, Mathur R, Gonzalez-Izquierdo A, et al. Identifying and visualising multimorbidity and comorbidity patterns in patients in the English National Health Service: a population-based study. Lancet Digit Health 2023;5(1):e16−27. https://doi.org/10.1016/S2589-7500(22)00187-X.

[5] Rafferty J, Lee A, Lyons RA, Akbari A, Peek N, Jalali-Najafabadi F, et al. Using hypergraphs to quantify importance of sets of diseases by healthcare resource utilisation: a retrospective cohort study. PLoS One 2023;18(12):1−12. https://doi.org/10.1371/journal.pone.0295300.

[6] Nowakowska M, Zghebi SS, Ashcroft DM, Buchan I, Chew-Graham C, Holt T, et al. The comorbidity burden of type 2 diabetes

mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC Med 2019;17(1):145. https://doi.org/10.1186/s12916-019-1373-y.

[7] Suh JW, Floud S, Reeves GK, Cairns BJ, Wright FL. Multimorbidity of cardiovascular disease subtypes in a prospective cohort of 1.2 million UK women. Open Heart 2023;10(2):e002552. https://doi.org/10.1136/openhrt-2023-002552.

[8] DeLong LN, Fleetwood K, Prigge R, Galdi P, Guthrie B, Fleuriot JD. Cluster and survival analysis of UK biobank data reveals associations between physical multimorbidity clusters and subsequent depression. Commun Med 2025;5(1):156. https://doi.org/10.1038/s43856-025-00825-7.

[9] Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018;47(5):1687−704. https://doi.org/10.1093/ije/dyy134.

[10] Busija L, Lim K, Szoeke C, Sanders KM, McCabe MP. Do replicable profiles of multimorbidity exist? Systematic review and synthesis. Eur J Epidemiol 2019;34(11):1025−53. https://doi.org/10.1007/s10654-019-00568-5.

[11] Dhafari TB, Pate A, Azadbakht N, Bailey R, Rafferty J, Jalali-Najafabadi F, et al. A scoping review finds a growing trend in studies validating multimorbidity patterns and identifies five broad types of validation methods. J Clin Epidemiol 2024;165:111214. https://doi.org/10.1016/j.jclinepi.2023.11.004.

[12] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38(11):2074−102. https://doi.org/10.1002/sim.8086.

[13] Nichols L, Taverner T, Crowe F, Richardson S, Yau C, Kiddle S, et al. In simulated data and health records, latent class analysis was the optimum multimorbidity clustering algorithm. J Clin Epidemiol 2022;152:164−75. https://doi.org/10.1016/j.jclinepi.2022.10.011.

[14] Zou GY. Toward using confidence intervals to compare correlations. Psychol Methods 2007;12(4):399−413. https://doi.org/10.1037/1082-989X.12.4.399.

[15] Diedenhofen B, Musch J. Cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 2015;10(4):1−12. https://doi.org/10.1371/journal.pone.0121945.

[16] Nguyen T, Viehman J, Yeboah D, Olbricht GR, Obafemi-Ajayi T. Statistical comparative analysis and evaluation of validation indices for clustering optimization. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia:3081−90. https://doi.org/10.1109/SSCI47803.2020.9308412.

[17] Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit 2013;46(1):243−56. https://doi.org/10.1016/j.patcog.2012.07.021.

[18] Akbarpour N, Akbari E, Motameni H. External clustering validity index based on extended similarity measures. J Comput Sci 2023;72:102116. https://doi.org/10.1016/j.jocs.2023.102116.

[19] Roso-Llorach A, Violán C, Foguet-Boreu Q, Rodriguez-Blanco T, Pons-Vigués M, Pujol-Ribera E, et al. Comparative analysis of methods for identifying multimorbidity patterns: a study of "real-world" data. BMJ Open 2018;8(3):e018986. https://doi.org/10.1136/bmjopen-2017-018986.

[20] Violán C, Fernández-Bertolín S, Guisado-Clavero M, Foguet-Boreu Q, Valderas JM, Vidal Manzano J, et al. Five-year trajectories of multimorbidity patterns in an elderly mediterranean population using hidden markov models. Sci Rep 2020;10(1):16879. https://doi.org/10.1038/s41598-020-73231-9.

[21] Guisado-Clavero M, Roso-Llorach A, López-Jimenez T, Pons-Vigués M, Foguet-Boreu Q, Muñoz MA, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. BMC Geriatr 2018;18(1):16. https://doi.org/10.1186/s12877-018-0705-7.

[22] Grant RW, McCloskey J, Hatfield M, Uratsu C, Ralston JD, Bayliss E, et al. Use of latent class analysis and k-Means clustering to identify complex patient profiles. JAMA Netw Open 2020;3(12):e2029068. https://doi.org/10.1001/jamanetworkopen.2020.29068.

[23] Zhong Y, Qin G, Xi H, Cai D, Wang Y, Wang T, et al. Prevalence, patterns of multimorbidity and associations with health care utilization among middle-aged and older people in China. BMC Public Health 2023;23(1):537. https://doi.org/10.1186/s12889-023-15412-5.

[24] Carrasco-Ribelles LA, Cabrera-Bean M, Danés-Castells M, Zabaleta-Del-Olmo E, Roso-Llorach A, Violán C. Contribution of frailty to multimorbidity patterns and trajectories: longitudinal dynamic cohort study of aging people. JMIR Public Health Surveill 2023;9:e45848. https://doi.org/10.2196/45848.

[25] Tran T, Bliuc D, Abrahamsen B, Chen W, Eisman JA, Hansen L, et al. Multimorbidity clusters potentially superior to individual diseases for stratifying fracture risk in older people: a nationwide cohort study. Age Ageing 2024;53(7):afae164. https://doi.org/10.1093/ageing/afae164.

[26] Zheng W, Huang X, Wang X, Suo M, Yan Y, Gong W, et al. Impact of multimorbidity patterns on outcomes and treatment in patients with coronary artery disease. Eur Heart J Open 2024;4(2):oeae009. https://doi.org/10.1093/ehjopen/oeae009.

[27] Krauth SJ, Steell L, Ahmed S, McIntosh E, Dibben GO, Hanlon P, et al. Association of latent class analysis-derived multimorbidity clusters with adverse health outcomes in patients with multiple long-term conditions: comparative results across three UK cohorts. eClinicalMedicine 2024;74:102703. https://doi.org/10.1016/j.eclinm.2024.102703.

[28] Ho IS-S, Azcoaga-Lorenzo A, Akbari A, Black C, Davies J, Hodgins P, et al. Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies. Lancet Public Health 2021;6(8):e587−97. https://doi.org/10.1016/S2468-2667(21)00107-9.