STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–11.
Published online in International Academic Press (www.IAPress.org)

Integrating Statistical Clustering Methods and Machine Learning to Uncover Latent Population Profiles in Childhood Disability

Ali Satty [1,*], Zakariya M. S. Mohammed [1,2]

[1] *Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia*
[2] *Center for Scientific Research and Entrepreneurship, Northern Border University, Arar, Saudi*

Abstract Childhood disability in low-resource settings is shaped by intersecting socioeconomic and geographic disadvantages, yet traditional regression approaches cannot capture how these factors combine to form distinct population subgroups. Using nationally representative MICS6 data from the Central African Republic, this study integrates two complementary unsupervised learning methods, Latent Class Analysis (LCA) and Partitioning Around Medoids (PAM) with Gower distance, to identify latent vulnerability profiles among 6,167 children aged 5–17 years. All analyses were conducted within a survey-weighted framework to account for the complex sampling design. Model selection using BIC/AIC (LCA) and weighted silhouette and Calinski–Harabasz indices (PAM) consistently supported a six-cluster structure. These clusters reflected meaningful combinations of household wealth, maternal education, residential setting, and region, with disability prevalence ranging from 28.5% to 43.3% across clusters. Although individual-level agreement between LCA and PAM assignments was minimal, both methods revealed consistent high-level population patterns, such as the coexistence of socioeconomically advantaged low-risk groups and disadvantaged high-risk groups. Rather than serving as mutual validation, this agreement is interpreted as a form of sensitivity analysis, indicating that these dominant patterns are robust to the choice of clustering framework. The findings demonstrate the methodological value of triangulating model-based and distance-based clustering to uncover hidden structures in complex survey data and provide new insights into the multidimensional nature of childhood disability in fragile settings.

Keywords Childhood Disability, Clustering Methods, Latent Class Analysis, Partitioning Around Medoids

DOI: 10.19139/soic-2310-5070-3308

1. Introduction

Childhood disability is a major but often under-recognised global public-health challenge. More than 240 million children worldwide live with disabilities, with the burden disproportionately concentrated in low- and middle-income countries (LMICs), where access to healthcare, rehabilitation, inclusive education, and social protection remains limited [1]. Functional difficulties, including impairments in cognition, mobility, hearing, vision, communication, or self-care, are associated with developmental delays, reduced school readiness, social exclusion, and long-term socioeconomic disadvantage [2]. In West and Central Africa, these challenges are compounded by chronic poverty, malnutrition, forced displacement, and weak health systems, which together elevate the risk of functional limitations among children [3, 4].

The Central African Republic (CAR) is among the most fragile settings in the region, where prolonged conflict, insecurity, and widespread deprivation have severely constrained child health and development. Nearly half of children lack access to essential healthcare and approximately 40% experience chronic malnutrition [5]. Despite

*Correspondence to: Ali Satty (Email: alisatty1981@gmail.com). Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia

ISSN 2310-5070 (online) ISSN 2311-004X (print)

these conditions, empirical evidence on the sociodemographic and household patterns associated with childhood disability in CAR remains limited.

The introduction of the UNICEF/Washington Group Child Functioning Module (CFM) in the sixth round of the Multiple Indicator Cluster Surveys (MICS6) provides an unprecedented opportunity to measure functional difficulties using standardised, internationally comparable indicators [6]. Existing research in LMICs has primarily relied on logistic or multinomial regression models to examine independent associations between predictors and disability [7, 3, 4]. While informative, such approaches cannot capture how overlapping socioeconomic, geographic, and maternal factors combine to form distinct vulnerability profiles.

Unsupervised machine learning offers a powerful framework for uncovering these hidden structures. Latent Class Analysis (LCA) models the joint distribution of multiple categorical indicators to identify unobserved subgroups that share similar characteristics [8, 9]. Partitioning Around Medoids (PAM), using Gower's distance for mixed data [10], provides a complementary distance-based approach that does not rely on distributional assumptions and performs well with categorical survey data [11]. Using both approaches together provides complementary perspectives on population heterogeneity rather than direct validation. LCA identifies latent probabilistic structures based on shared response patterns, whereas PAM captures similarity based on empirical dissimilarities in mixed-type data. Comparing results across these fundamentally different frameworks can be viewed as a form of sensitivity analysis: if both methods reveal similar high-level patterns, this suggests that these patterns reflect dominant structural features of the data rather than artifacts of a single modelling assumption.

To our knowledge, no prior study has combined LCA with distance-based PAM clustering to analyse nationally representative MICS data. This dual-framework design introduces a methodological advance by integrating a likelihood-based latent class model with a nonparametric, optimization-oriented medoid clustering algorithm under a complex survey design. By incorporating sampling weights throughout and evaluating whether two fundamentally different clustering paradigms converge on similar population structures, this study strengthens the reliability of unsupervised learning applications in public-health datasets. The approach demonstrates how distribution-free Gower dissimilarities and model-based latent structures can jointly reveal robust vulnerability patterns that are not detectable using traditional regression or a single clustering method. This methodological triangulation provides a rigorous foundation for uncovering hidden population heterogeneity in childhood disability.

Motivated by the need to move beyond variable-by-variable regression analyses, this study aims to uncover data-driven profiles of childhood disability in the Central African Republic using survey-weighted unsupervised machine learning. Specifically, it seeks to: (1) identify latent subgroups of children exhibiting shared patterns of socioeconomic, geographic, and maternal disadvantage; (2) assess whether these profiles are consistently recovered across two complementary frameworks, model-based LCA and distance-based PAM with Gower dissimilarity; and (3) evaluate how the resulting clusters map onto key contextual determinants, including household wealth, maternal education, residential setting, and geographic region. By integrating multiple unsupervised learning approaches, this study provides a deeper understanding of population-level heterogeneity in childhood disability and generates evidence to guide equitable and targeted interventions in LMIC settings.

2. Methods

2.1. Survey data extraction and variable specification

This study used nationally representative microdata from the MICS6–CAR survey [12], collected using a stratified two-stage cluster sampling design. The analytic sample included 6,167 children aged 5–17 years who completed the household and child functioning modules. The outcome variable was functional disability, constructed as a binary indicator based on the UNICEF/Washington Group Child Functioning Module, a validated tool assessing difficulties in vision, hearing, cognition, mobility, communication, and behaviour [6]. Functional difficulty of the mother was defined based on self-reported responses to the survey instrument and reflects the presence of functional limitations as captured by the MICS questionnaire. The covariates incorporated into the analysis reflected well-established social determinants of child health and disability in LMICs and included: child sex (male, female), age

group (5–9, 10–14, 15–17 years), level of education of the child (preschool or none, fundamental, higher), level of maternal education (preschool or none, fundamental, higher), functional difficulty of the maternal mother (yes/no), wealth of the household (poorest, poor, middle, rich, richest), residential area (urban, rural) and region of residence (seven administrative regions code. These factors have been widely documented as key structural determinants of childhood disability, inequality, and vulnerability in low- and middle-income settings [13, 14, 4].

2.2. Pre-processing

Pre-processing ensured that all variables were suitable for clustering and that both algorithms operated on a consistent dataset. Because clustering is sensitive to category definitions, missingness, and mixed data types, all sociodemographic variables were cleaned and recoded as fully labelled categorical factors to harmonise categories across observations. The functional disability variable was likewise standardised into a binary indicator using a uniform coding scheme. The extent of missing data across clustering variables was assessed prior to analysis. Observations with missing values were excluded under a complete-case approach to ensure consistency between LCA and PAM analyses. The proportion of excluded observations is reported in the Results section.

As the clustering variables comprised both nominal and ordered categorical data, no numeric scaling or standardisation was applied. Mixed data were instead accommodated using Gower's dissimilarity measure,

$$\delta_{ij} = \frac{1}{p}\sum_{k=1}^{p}\delta_{ijk}, \tag{1}$$

where $\delta_{ijk} = 0$ for matching categories and 1 for mismatches [10]. In this expression, i and j denote individual survey respondents, k indexes the clustering variables, and p is the total number of variables included in the computation of the dissimilarity. Accordingly, δ_{ij} represents the average dissimilarity between individuals i and j across all p variables.

2.3. Clustering methods

To better understand patterns of risk for child disability, we applied two complementary unsupervised machine-learning methods: LCA and PAM. LCA provides a statistical model-based representation of population heterogeneity, while PAM offers a data-driven, assumption-free perspective based on empirical dissimilarities. Using both approaches enables methodological triangulation, strengthening confidence that the identified clusters represent meaningful vulnerability profiles rather than artifacts of a single modelling framework.

2.3.1. LCA It is a model-based clustering method for categorical data that assumes the existence of an unobserved discrete latent variable assigning individuals to one of K mutually exclusive and exhaustive classes [8]. Under this framework, the observed categorical indicators $(Y_{i1}, \ldots, Y_{ip})$ are assumed to be conditionally independent given latent class membership. The joint probability distribution for individual i factorises as:

$$P(Y_{i1}, \ldots, Y_{ip}) = \sum_{c=1}^{K}\pi_c \prod_{j=1}^{p} P(Y_{ij} \mid C_i = c), \tag{2}$$

where π_c is the prevalence (prior probability) of latent class c, satisfying $\sum_{c=1}^{K}\pi_c = 1$, and $P(Y_{ij} \mid C_i = c)$ denotes the class-specific conditional response probabilities. LCA simultaneously estimates both the latent class proportions and the conditional response probabilities, thereby uncovering hidden population subgroups with shared characteristics. Model adequacy was assessed using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

$$AIC = -2\ell + 2K, \qquad BIC = -2\ell + K\log(n), \tag{3}$$

where ℓ is the log-likelihood, K is the number of estimated parameters, and n is the sample size [15]. Lower values indicate improved parsimony-adjusted model fit.

2.3.2. PAM with Gower distance It is a distance-based clustering method that partitions observations into K clusters by selecting representative objects, called medoids, that minimise the total dissimilarity within each cluster [11]. Unlike k-means, which relies on numerical centroids, PAM identifies actual data points as cluster centres, making it especially suitable for categorical or mixed-type data when paired with Gower's distance metric. Formally, PAM seeks to solve:

$$\min_{m_1,\ldots,m_K} \sum_{c=1}^{K} \sum_{i \in C_c} d(i, m_c), \tag{4}$$

where $d(i, m_c)$ is the dissimilarity between observation i and medoid m_c, and C_c denotes the set of observations assigned to cluster c. Gower's distance is advantageous for mixed data because it accommodates nominal, binary, and ordinal variables without transformation [10]. Since PAM does not natively incorporate sampling weights, weighted silhouette widths were computed to evaluate cluster quality:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{5}$$

where a_i is the average dissimilarity of observation i to its assigned cluster, and b_i is the minimum average dissimilarity to any other cluster [16]. Weighted silhouette values were obtained by multiplying each s_i by its corresponding survey weight. It should be noted that the PAM algorithm was applied to an unweighted dissimilarity matrix, as standard implementations do not natively incorporate sampling weights. While survey weights were used in the evaluation of clustering performance (e.g., weighted silhouette widths) and in post-clustering summaries, the cluster assignments themselves reflect patterns in the unweighted data.

2.4. Model selection and validation

2.4.1. Selecting the optimal number of clusters To ensure robust and interpretable clustering solutions, we compared results across both frameworks:

- LCA: AIC and BIC were used to identify the optimal number of latent classes, with additional consideration given to interpretability.
- PAM: The width of the weighted silhouette served as the primary criterion to select the number of clusters, supported by unweighted silhouettes and Calinski–Harabasz values.

2.4.2. Stability and internal validation To evaluate the robustness of the PAM clustering solution, a split-sample stability analysis was conducted. The complete-case dataset was randomly partitioned into a training subset (70%) and a testing subset (30%), and PAM was re-estimated independently on both subsets. The similarity between cluster assignments was quantified using the Adjusted Rand Index (ARI):

$$ARI = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}, \tag{6}$$

where values near zero indicate agreement no better than chance [17].

2.5. Agreement between LCA and PAM

Concordance between LCA and PAM cluster assignments was evaluated using two metrics: (1) cross-tabulation of class/cluster membership with row and column percentages, and (2) Cohen's Kappa statistic:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{7}$$

where p_0 is the observed agreement and p_e is the agreement expected by chance [18].

2.6. Analytical approach

All statistical procedures were implemented in R version 4.5.1. The survey design object was defined using `svydesign()`, specifying primary sampling units, strata, and sampling weights, and this design was used throughout for all survey-weighted estimation. LCA models with $K = 2$–6 classes were estimated using the `poLCA` package on the numeric-coded complete-case dataset. Each model was fitted with multiple random starts to reduce the influence of local optima, and AIC/BIC values were stored to select the optimal number of classes. Posterior probabilities were converted into hard class assignments and merged back into the survey dataset for weighted estimation of cluster-specific disability prevalence and sociodemographic profiles. For PAM clustering, Gower dissimilarities were computed using `daisy()`, and PAM models for $K = 2$–6 clusters were fitted with `pam(., diss = TRUE)`. Model quality was evaluated using unweighted and survey-weighted silhouette widths and the Calinski–Harabasz index from `clusterCrit`. The optimal PAM solution was chosen using the maximum weighted silhouette width, and the resulting assignments were appended to the survey design. Robustness of PAM was examined through a 70/30 split-sample stability analysis, with agreement quantified via ARI. Concordance between LCA and PAM classifications was similarly assessed using ARI, cross-tabulations, and Cohen's Kappa. Post-clustering characterisation included survey-weighted summaries and visualisations constructed using `ggplot2`. To enhance reproducibility, all LCA models were estimated using multiple random starts to reduce sensitivity to local optima. Where applicable, random seeds were set to ensure consistent results across runs.

3. Results

3.1. Baseline characteristics of children with functional disability

Among the 6,167 children included in the analysis, 1,891 (30.7%) met the criteria for functional disability. The sex distribution was balanced, with 51.2% male and 48.8% female, indicating no major gender disparity in disability prevalence. Age patterns showed that 41.5% of children with disability were aged 5–9 years, 34.6% were 10–14 years, and 23.9% were 15–17 years, suggesting higher vulnerability in younger age groups. Educational indicators revealed that approximately 44–48% of disabled children had no preschool exposure, while the majority were enrolled in fundamental education (about 52–55%). Socioeconomic disadvantage was pronounced: 57–60% of children with disability lived in poor or poorest households, with far fewer represented in the wealthier categories. Rural residence was similarly overrepresented, with about 78% of disabled children residing in rural areas. Maternal education patterns mirrored these inequalities, with more than half (approximately 53%) of mothers reporting no formal schooling and fewer than 15% having completed secondary or higher education. Disability was disproportionately concentrated in disadvantaged geographic areas, with Regions 3, 4, and 5 together accounting for more than 45% of all disabled children. Further descriptive statistics can be found in [12]. The proportion of missing data across the clustering variables was low (less than 8%), suggesting that the impact of complete-case exclusion on the analytical sample is minimal.

3.2. Cluster solution

Table 1 summarises the model selection criteria for the LCA and PAM clustering solutions estimated for $K = 2$–6 clusters. For LCA, both BIC and AIC declined steadily as the number of classes increased, with the lowest BIC (66,508) and AIC (65,707) obtained for the six-class solution. This pattern supports a six-class model as the best trade-off between fit and parsimony. For PAM, both unweighted and survey-weighted silhouette widths increased monotonically with larger K, peaking at 0.307 (unweighted) and 0.303 (weighted) for $K = 6$. These values indicate improved within-cluster cohesion and clearer between-cluster separation. Together, the model-based LCA and distance-based PAM frameworks converged on a six-cluster solution as the most appropriate representation of heterogeneity in children's sociodemographic and disability characteristics.

Table 1. Fit statistics for LCA (BIC/AIC) and PAM (silhouette scores) for $K = 2$–6.

K	LCA BIC	LCA AIC	Silhouette (unweighted)	Silhouette (weighted)
2	67,423	67,160	0.221	0.215
3	67,021	66,624	0.261	0.262
4	66,747	66,215	0.276	0.274
5	66,524	65,858	0.269	0.257
6	66,508	65,707	0.307	0.303

3.3. Disability prevalence by cluster

Table 2 presents the survey-weighted disability prevalence for the six clusters identified by LCA and PAM, and Figure 1 provides a graphical comparison. For LCA, prevalence ranged from 29.4% (Cluster 3) to 37.9% (Cluster 5), while for PAM the range was 28.5% (Cluster 2) to 43.3% (Cluster 6). Both frameworks consistently identified one or two high-risk clusters alongside several moderate-risk groups. Clusters characterised by rural residence, socioeconomic disadvantage, or low maternal education exhibited the highest disability burden.

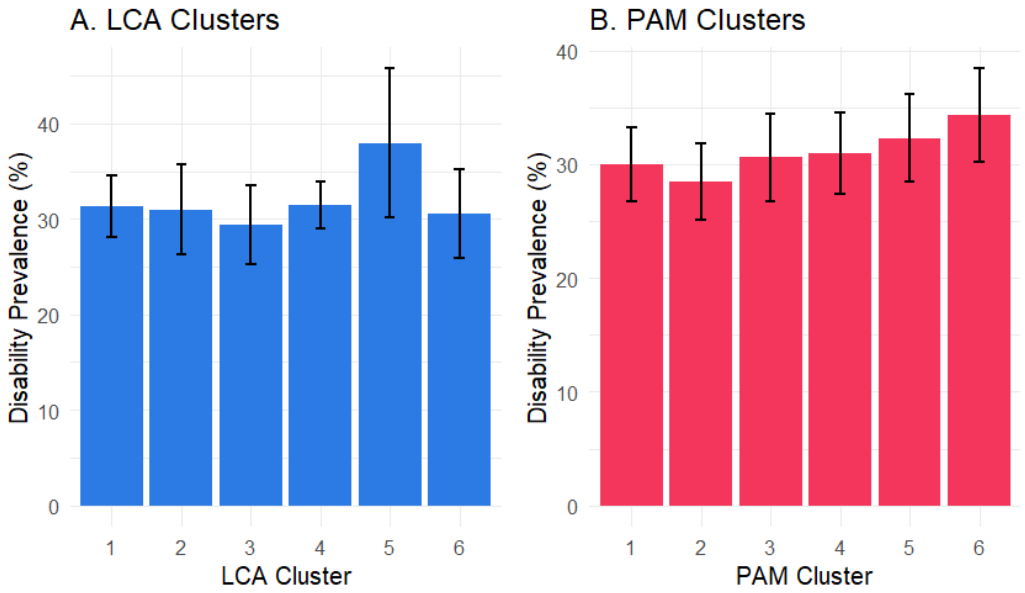

Figure 1. Comparison of survey-weighted disability prevalence (%) across the six clusters identified by LCA and PAM.

Table 2. Survey-weighted disability prevalence (%) by cluster for LCA and PAM models.

Cluster	LCA Prev.	LCA SE	PAM Prev.	PAM SE
1	31.3	0.016	30.0	0.016
2	30.9	0.024	28.5	0.017
3	29.4	0.021	30.6	0.019
4	31.5	0.013	31.0	0.018
5	37.9	0.040	32.3	0.020
6	30.5	0.024	43.3	0.021

Note: *SE denotes the standard error of the estimated prevalence.

3.4. Description of the disability profiles

Cluster labels such as "mixed socioeconomic status" and "moderate socioeconomic status" are defined based on the distribution of observations across wealth quintiles and related indicators within each cluster, as presented in Table 3, which summarises the defining characteristics of the clusters produced by LCA and PAM. Both methods reveal six distinct sociodemographic profiles reflecting different combinations of household conditions, maternal attributes, and geographic context.

3.4.1. LCA-derived profiles

- Cluster 1 (21%): Predominantly rural, high poverty, low maternal education; moderate disability prevalence.
- Cluster 2 (8%): Mostly urban, relatively affluent, higher maternal education; moderate disability prevalence.
- Cluster 3 (14%): Concentrated in Regions 2–3; moderate SES; lowest disability prevalence.
- Cluster 4 (37%): Largest cluster; rural households with average SES; moderate disability prevalence.
- Cluster 5 (3%): Small, highly vulnerable subgroup with mixed SES; highest disability prevalence.
- Cluster 6 (17%): Predominantly rural with concentration in Region 3; average disability.

3.4.2. PAM-derived profiles

- Cluster 1 (15%) and Cluster 2 (17%): Urban and wealthier households; Cluster 2 had the lowest disability prevalence.
- Cluster 3 (16%) and Cluster 4 (18%): Rural and socioeconomically disadvantaged; moderate disability prevalence.
- Cluster 5 (18%): Mixed rural SES; moderately elevated disability.
- Cluster 6 (16%): Poorest rural households; highest disability prevalence; most vulnerable PAM profile.

3.4.3. Cross-method comparison Despite methodological differences, both algorithms yielded similar structural patterns:

- Urban, wealthy clusters consistently showed the lowest disability burden.
- Large rural clusters with moderate socioeconomic disadvantage exhibited moderate disability risk.
- A small, high-poverty subgroup emerged in both methods as the most vulnerable (LCA Cluster 5; PAM Cluster 6).

These convergent findings underscore that vulnerability is multidimensional, shaped by intersecting socioeconomic, educational, and geographic factors.

To further assess the similarity between clusters identified by LCA and PAM, we compared the survey-weighted distributions of key covariates across corresponding high-risk and low-risk clusters. These comparisons confirm that, although individual-level assignments differ substantially, both methods consistently identify clusters characterized by higher poverty, rural residence, and lower maternal education as having elevated disability prevalence. This supports the interpretation that both methods capture similar structural gradients in the data, despite differences in individual classification.

3.5. Agreement between LCA and PAM classifications

Table 4 reports agreement statistics between the LCA and PAM cluster assignments. The agreement between LCA and PAM at the individual level was extremely low, with an ARI close to zero and near-zero Cohen's Kappa values. This indicates that the two methods assign individuals to clusters in fundamentally different ways. Consequently, the identified clusters should not be interpreted as stable or reproducible classifications at the individual level. Instead, they reflect alternative representations of underlying population structure, highlighting different aspects of heterogeneity depending on the methodological framework used. The split-sample validation for PAM yielded an ARI of 0.42, indicating moderate stability of the clustering solution. While this suggests some consistency in cluster structure, it also implies that the PAM-derived profiles are not perfectly robust across samples and should be interpreted with caution.

Table 3. Cluster sizes and sociodemographic profiles identified by LCA and PAM.

Method	Cluster	Unweighted n	Weighted %	Description
LCA	1	1202	20.95	Rural; high poverty; low maternal education; moderate disability.
LCA	2	890	7.95	Urban; higher wealth; more educated mothers; moderate disability.
LCA	3	744	13.97	Regions 2–3; moderate SES; lowest disability.
LCA	4	2078	36.99	Rural households; moderate disability.
LCA	5	240	2.85	Mixed SES; high disability; most vulnerable.
LCA	6	1013	17.30	Rural; Region 3; average disability.
PAM	1	1126	14.64	Urban wealthy cluster; low disability.
PAM	2	1230	16.79	Fully urban; richest; lowest disability.
PAM	3	946	16.34	Rural, poorest; moderate disability.
PAM	4	1007	18.38	Rural; lowest SES; moderate disability.
PAM	5	999	18.14	Mixed rural SES; moderately elevated disability.
PAM	6	859	15.70	Fully rural; poorest; highest disability.

Table 4. Agreement between LCA and PAM cluster solutions.

Metric	Result
ARI	0.00007
Cohen's Kappa (unweighted)	0.0007 (ASE = 0.0055, $p = 0.90$)
Cohen's Kappa (weighted)	0.0092 (ASE = 0.0084, $p = 0.27$)

3.6. Relationship of profiles with covariates

Clear and consistent associations emerged between the identified clusters and key sociodemographic characteristics. Across both LCA and PAM, four structural factors, household wealth, residential area, maternal education, and geographic region, were the strongest predictors of profile membership. A pronounced socioeconomic gradient was observed: affluent, urban clusters contained the highest proportions of children from rich households, whereas the most vulnerable clusters were overwhelmingly concentrated in the poorest categories. Figure 2 provides an illustrative example of this pattern by showing the wealth distribution across clusters. Rural residence was similarly concentrated within the high-risk groups, with some clusters (e.g., PAM Cluster 6) composed entirely of rural households. Maternal education also showed strong stratification: clusters characterised by low maternal schooling exhibited higher disability prevalence, while profiles with better-educated mothers showed the lowest levels. Regional patterns further reinforced these inequalities, with certain administrative regions dominating the low-risk clusters and others contributing disproportionately to the most disadvantaged groups. Child age and sex were relatively evenly distributed across all clusters, suggesting that once socioeconomic and geographic conditions are accounted for, neither variable substantially influences profile membership. Overall, the clustering structure reflects underlying social inequities: poverty, rurality, and limited maternal education consistently define the highest-risk profiles, whereas urban residence, higher household wealth, and greater maternal education characterise the most advantaged groups.

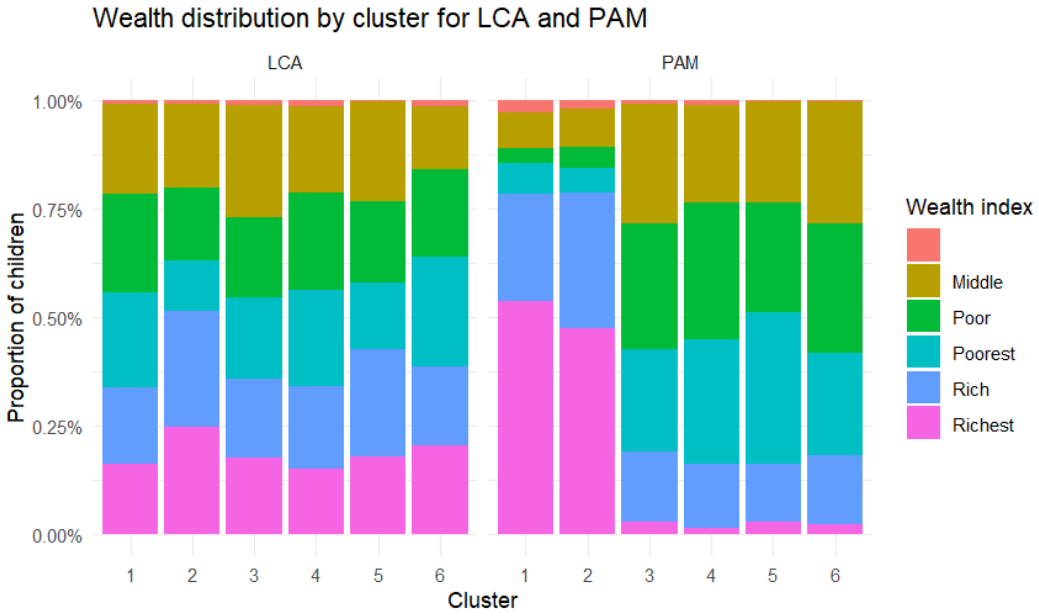

Figure 2. Wealth distribution by cluster for LCA and PAM

4. Discussion

A key finding of this study is the substantial divergence between LCA and PAM cluster assignments at the individual level, as evidenced by near-zero ARI and Kappa statistics. This indicates that the identified clusters are not stable entities for classifying individual children and are highly dependent on the chosen methodological framework. Therefore, the clusters should not be interpreted as definitive group memberships for targeted intervention at the individual level. Instead, the value of this analysis lies in identifying consistent population-level patterns of vulnerability. Both methods highlight the central role of socioeconomic disadvantage, rural residence, and low maternal education as key dimensions associated with increased disability risk. These findings provide important insights for policy by emphasizing structural determinants rather than individual classification.

Beyond their substantive findings, the two clustering frameworks used in this study are grounded in explicit optimisation principles. The LCA model is fitted by maximising the observed-data log-likelihood via the expectation–maximisation (EM) algorithm, which iteratively updates posterior class probabilities and class-specific response parameters until convergence [20, 21]. In contrast, PAM solves a discrete combinatorial optimisation problem by selecting a set of K medoids that minimises the total within-cluster dissimilarity; classical BUILD–SWAP heuristics yield efficient approximations for large datasets [11, 22]. Each medoid serves as an optimisation-derived prototype anchored to an observed data point, enhancing interpretability and robustness to outliers. These complementary optimisation perspectives, likelihood maximisation for LCA and dissimilarity minimisation for PAM, help explain why the methods can diverge at the individual assignment level yet converge on similar population-level vulnerability structures. Framing the analysis through these computational principles underscores that the resulting profiles are not arbitrary groupings but the product of well-defined statistical and optimisation criteria.

By situating these findings within broader global research, the study reinforces established evidence that childhood disability in low-resource settings is closely tied to structural barriers. International studies consistently show that socioeconomic deprivation, geographic marginalisation, and limited access to developmental services shape children's functional outcomes [13, 23, 24]. The similarity between these global patterns and the profiles uncovered here suggests that the mechanisms driving disability risk in the CAR reflect broader determinants documented across LMIC contexts [25, 26, 27]. Importantly, the clustering approach extends beyond confirming

known associations by illustrating how these determinants combine to form coherent groups of children experiencing shared vulnerabilities.

The use of both LCA and PAM can be interpreted as a form of sensitivity analysis across clustering paradigms. The emergence of similar high-level patterns across fundamentally different methods suggests that these patterns are driven by dominant gradients in the data rather than methodological artifacts. However, this consistency does not imply agreement in individual classification.

The identification of a small, acutely disadvantaged cluster has clear implications for policy and programme design. This subgroup likely represents children exposed simultaneously to socioeconomic hardship, low maternal education, and geographic isolation, and who therefore require disproportionately higher levels of support. Prioritising early detection, strengthening referral pathways, and expanding rehabilitative and educational services in underserved rural areas may help mitigate the cumulative disadvantage experienced by this group. More broadly, the cluster structure underscores the importance of integrating disability considerations into poverty alleviation strategies, maternal education initiatives, and rural service delivery systems.

Despite its strengths, this study has several limitations. The cross-sectional nature of the MICS data limits the ability to draw causal inferences. The observed relationships between disability and socioeconomic or demographic factors reflect associations and co-occurrence rather than directional effects. In particular, reverse or bidirectional relationships may exist, for example where childhood disability influences household socioeconomic conditions through caregiving demands. Although both clustering methods revealed comparable population-level patterns, the lack of agreement in individual assignments limits the suitability of the clusters for case-level prediction or clinical decision-making. Several methodological constraints also warrant consideration. Survey-weighted clustering may yield different structures under alternative sampling designs or weighting schemes, which may affect the generalisability of the six-cluster solution to other settings. In addition, LCA depends on a weighted log-likelihood maximised through an expectation–maximisation algorithm that is sensitive to local maxima, meaning that alternative latent solutions cannot be fully excluded even when multiple random starts are used. Likewise, PAM identifies medoids through an optimisation procedure that minimises total within-cluster dissimilarity, and different distance metrics or initialisation strategies may lead to plausible alternative partitions. Together, these factors suggest that the identified clusters represent one reasonable configuration under specific modelling assumptions, and future work should examine robustness across alternative weighting, distance, and optimisation choices, as well as incorporate additional determinants such as nutrition, environmental exposures, and access to specialist services when data permit.

Future research should explore whether incorporating additional variables, such as nutritional status, environmental exposures, or access to healthcare, could improve the stability of clustering solutions. Additionally, alternative clustering approaches or hybrid methods may help generate more reliable and actionable typologies for individual-level targeting.

5. Conclusion

This study integrated model-based and distance-based clustering methods to uncover six distinct vulnerability profiles of childhood disability in the CAR, demonstrating how unsupervised learning can reveal multidimensional patterns that extend beyond traditional regression approaches. By jointly applying LCA and PAM clustering with Gower distance, the analysis highlighted substantial heterogeneity driven primarily by socioeconomic disadvantage, rural residence, and limited maternal education. These structural factors consistently distinguished high-risk groups from their more advantaged counterparts, underscoring the central role of social and geographic inequalities in shaping disability outcomes.

The findings highlight key structural dimensions of vulnerability associated with childhood disability, including poverty, rural residence, and low maternal education. While the clustering approaches provide valuable insights into population-level heterogeneity, the lack of individual-level agreement suggests that these profiles should not be used for direct targeting of individuals. Instead, they offer a framework for understanding broader patterns of inequality and informing population-level policy strategies.

While the two clustering frameworks differed in individual-level assignments, their convergence at the population level reinforces the robustness of the overarching vulnerability patterns. The findings offer valuable insights for designing targeted, equity-focused interventions, particularly for the small but highly disadvantaged subgroup identified across methods. Continued efforts to incorporate richer contextual data, examine longitudinal dynamics, and assess the reproducibility of cluster structures across diverse LMIC settings will strengthen the evidence base for tailoring programmes to the needs of the most vulnerable children.

Acknowledgement

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2026-2877-08."

REFERENCES

1. UNICEF, *Seen, Counted, Included: Using data to shed light on the well-being of children with disabilities*, UNICEF, 2021.
2. B. O. Olusanya, S. M. Wright, M. K. C. Nair, et al., *Global burden of childhood developmental difficulties in low- and middle-income countries*, BMJ Paediatrics Open, vol. 6, e001397, 2022.
3. M. B. Alam, M. Rahman, M. M. Rahman, et al., *Disability and adverse health and nutritional outcomes in South Asia*, Lancet Regional Health – Southeast Asia, vol. 25, 100401, 2024.
4. S. Rotenberg, C. Davey, and E. McFadden, *Disability and healthcare utilisation in 10 African countries*, EClinicalMedicine, vol. 57, 101870, 2023.
5. UNICEF, *Forgotten children of CAR*, UNICEF Press Release, 2024.
6. M. Loeb, C. Cappa, R. Crialesi, and E. de Palma, *Measuring child functioning: The UNICEF/Washington Group Module*, Salud Publica Mex, vol. 59, no. 4, pp. 485–487, 2017.
7. M. M. Rahman, I. M. Alam, M. Mansur, et al., *Functional difficulty among children in Bangladesh*, PLOS ONE, vol. 19, e0300403, 2024.
8. A. L. McCutcheon, *Latent Class Analysis. Quantitative Applications in the Social Sciences, Vol. 64*, SAGE Publications, Newbury Park, CA, 1987.
9. L. M. Collins, and S. T. Lanza, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*, John Wiley & Sons, Hoboken, NJ, 2010.
10. J. C. Gower, *A general coefficient of similarity and some of its properties*, Biometrics, vol. 27, no. 4, pp. 857–871, 1971.
11. L. Kaufman, and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, Hoboken, 1990.
12. ICASEES and UNICEF, *Multiple Indicator Cluster Survey (MICS6) – Central African Republic 2018–2019: Final Report of Survey Results*, ICASEES, Bangui, Central African Republic, 2021.
13. L. M. Banks, H. Kuper, and S. Polack, *Poverty and disability in low- and middle-income countries: A systematic review*, PLOS ONE, vol. 12, no. 12, e0189996, 2017.
14. D. E. Simkiss, C. M. Blackburn, F. O. Mukoro, J. M. Read, and N. J. Spencer, *Childhood disability and socio-economic circumstances in low and middle income countries: A systematic review*, BMC Pediatrics, vol. 11, 119, 2011.
15. H. Akaike, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.
16. P. J. Rousseeuw, *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.
17. L. Hubert, and P. Arabie, *Comparing partitions*, Journal of Classification, vol. 2, pp. 193–218, 1985.
18. J. Cohen, *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
19. E. Emerson, and C. Hatton, *The socio-economic circumstances of children at risk of disability in Britain*, Disability & Society, vol. 22, no. 6, pp. 563–580, 2007.
20. A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–38, 1977.
21. G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.
22. H.-S. Park and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering*, Expert Systems with Applications, vol. 36, no. 2, pp. 3336–3341, 2009.
23. T. Bright, and H. Kuper, *A systematic review of access to rehabilitation for people with disabilities in low- and middle-income countries*, International Journal of Environmental Research and Public Health, vol. 15, no. 10, 2165, 2018.
24. M. Zuurmond, I. Mactaggart, N. Kannuri, G. Murthy, J. E. Oye, and S. Polack, *Barriers and facilitators to accessing health services: A qualitative study amongst people with disabilities in Cameroon and India*, International Journal of Environmental Research and Public Health, vol. 16, no. 7, 1126, 2019.
25. S. Grantham-McGregor, Y. B. Cheung, S. Cueto, P. Glewwe, L. Richter, and B. Strupp, *Developmental potential in the first 5 years for children in developing countries*, The Lancet, vol. 369, no. 9555, pp. 60–70, 2007.
26. S. P. Walker, T. D. Wachs, S. Grantham-McGregor, M. M. Black, C. A. Nelson, S. L. Huffman, et al., *Inequality in early childhood: Risk and protective factors for early child development*, The Lancet, vol. 378, no. 9799, pp. 1325–1338, 2011.

27. World Health Organization, *World Report on Disability*, WHO Press, 2011.