Sequential Analysis of Partial Least Squares Structural Equation Modeling and Self-Organizing Maps Using a Mediation Approach and Latent Variables for Alzheimer's Disease Analysis

Wahyuni[1] **Bambang Widjanarko Otok**[1]* **Jerry Dwi Trijoyo Purnomo**[1]

[1]*Department of Statistics, Institut Teknologi Sepuluh Nopember, Indonesia*
Corresponding author's Email: bambang_wo@its.ac.id
(Received: January 3, 2025. Accepted: February 25, 2026. Published: February 28, 2026.)

Abstract

Alzheimer's disease heterogeneity challenges clinical management. This study introduces an integrative framework combining Partial Least Squares Structural Equation Modeling (PLS-SEM) and Self-Organizing Maps (SOM) for patient stratification using National Alzheimer's Coordinating Center data. PLS-SEM validated four latent constructs, identifying functional disability as the principal predictor of clinical severity, with indirect effects mediated by cognitive and social activities. Comparing raw indicators against validated latent scores, SOM outperformed K-Means and hierarchical clustering, exhibiting enhanced robustness and clustering quality. Utilizing latent scores, SOM delineated three statistically distinct clinical phenotypes: a mild phenotype comprising the vast majority, a severe phenotype, and a small high-risk phenotype. This integrated approach offers a clinically relevant framework for precise patient segmentation, facilitating targeted therapeutic interventions.

Keywords: Alzheimer's disease, PLS-SEM, Clustering, Self-organizing maps, Latent variable scores, Patient segmentation.

1. Introduction

Alzheimer's disease (AD) is a complex, progressive neurodegenerative disorder characterized by cognitive decline and behavioral changes. Research on AD is often conducted using simple statistical methods such as descriptive analysis, t-tests, ANOVA, Kruskal-Wallis, and Wilcoxon tests to compare clinical or demographic factors. For example, non-parametric tests are frequently used to compare levels of Alzheimer's-related anxiety based on demographic variables, while ANOVA is applied to test the effect of amyloid burden in pre-symptomatic individuals. However, these methods have limitations in handling complex multimodal data, as they tend to overlook simultaneous interactions among indicators and cannot effectively address data heterogeneity [1, 2]. Conventional clinical assessment methods have several drawbacks, including limitations in analyzing interacting variables, assumptions of data normality that are often unmet in clinical data, and the requirement for large sample sizes to achieve valid results [3].

To address these issues of complexity and heterogeneity, this study employs the Structural Equation Modeling (SEM) approach. SEM is a multivariate statistical technique used to explain causal relationships between latent variables, account for measurement error, and analyze complex interrelationships among various variables simultaneously [4]. SEM has two main approaches: Covariance-Based SEM (CB-SEM) and Partial Least Squares SEM (PLS-SEM) [4]. CB-SEM is designed to test or confirm established theories by meeting certain goodness-of-fit criteria [5]. This method requires strict statistical assumptions, including multivariate normality and large sample sizes to ensure unbiased parameter estimates [6]. In contrast, PLS-SEM employs a variance-based approach that provides more flexibility regarding normality assumptions and is highly effective for small sample sizes [7]. The fundamental difference lies in purpose; CB-SEM is intended for theory validation, while PLS-SEM is more suitable for model exploration and prediction in complex research contexts [8]. Therefore, this study employs PLS-SEM due to its strong capability in handling non-normal data and limited sample sizes [9].

Previous studies have shown that PLS-SEM is effective in integrating various clinical indicators and biomarkers to comprehensively model the condition of Alzheimer's disease. Analyzing Alzheimer's using PLS-SEM allows for an understanding of the causal relationships among complex and heterogeneous variables, thus providing a more comprehensive picture of the patient's condition compared to conventional analysis [3, 10, 11]. Although PLS-SEM offers advantages in modeling causal relationships, it has a

key limitation in that it is not designed to segment or cluster subjects automatically based on similar characteristics [12]. In addition, PLS-SEM assumes population homogeneity, which makes it less effective in identifying hidden heterogeneity, even though Alzheimer's patients often have highly variable clinical profiles [5].

To address these limitations, this study integrates Self-Organizing Maps (SOM) as a complementary method. SOM is an unsupervised neural network capable of mapping high-dimensional data into a two-dimensional representation in the form of a topological map [13]. The main advantage of SOM lies in its ability to preserve the neighborhood structure of the original feature space, enabling patterns of proximity and similarity among subjects to be visualized more clearly [14]. Additionally, SOM can identify clusters or subgroups of patients based on similar characteristics, which cannot be achieved through PLS-SEM alone [15]. Therefore, this study applies an integrated approach by combining PLS-SEM and SOM to provide a more comprehensive analysis of complex relationships and hidden patterns in the data.

The integration of PLS-SEM and SOM is carried out sequentially. First, PLS-SEM is used to validate the measurement model and produce accurate latent variable scores from Alzheimer's clinical indicators. These validated latent scores are then used as input for the SOM algorithm to perform topological mapping and patient clustering. This integrated approach enables patient segmentation based on statistically validated constructs, rather than on raw data that might contain noise and measurement errors. The main contributions of this study include:

1. Factor analysis influencing the progression of Alzheimer's disease using PLS-SEM, including identification of significant indicators and analysis of mediation effects.
2. Segmentation of Alzheimer's patients using SOM, which is proposed as the primary clustering method based on unique clinical profiles.
3. Comparative evaluation of SOM performance against baseline clustering methods, such as K-Means and Hierarchical Clustering, to validate its effectiveness in identifying distinct clinical phenotypes.

The remainder of this paper is structured as follows: Section 2 presents a literature review on PLS-SEM and clustering; Section 3 presents materials and methods, including dataset description and the proposed integrated approach; Section 4 reports experimental results, structural model validation, and SOM optimization; and Section 5 provides discussion on clinical implications, conclusions, and suggestions for future research.

2. Literature review

2.1 Structural equation modeling-partial least squares (SEM-PLS)

PLS-SEM is a causal-predictive approach to SEM that emphasizes prediction in estimating statistical models, whose structures are designed to provide causal explanations [9, 16]. The technique thereby overcomes the apparent dichotomy between explanation as typically emphasized in academic research and prediction, which is the basis for developing managerial implications [5].

2.2 Self organizing maps (SOM)

Self-Organizing Maps (SOM) represent a particular type of neural network designed for unsupervised learning, which projects high-dimensional data into lower-dimensional, topologically preserving maps. SOM excels in clustering, visualization, and classification tasks, enabling intuitive interpretation of complex data structures [17, 18]. Modern adaptations include visualization enhancements that integrate data topology and cluster boundaries to yield deeper insights into the underlying data [19]. Hardware implementations of SOM and variants such as Concurrent Self-Organizing Maps (CSOM) enhance computational efficiency and classification accuracy, promoting their use in real-time data mining applications [20, 21].

2.3 Clustering

Clustering methods identify natural, unlabeled groupings within datasets. For instance, K-Means divides data into a fixed number of clusters, whereas hierarchical clustering creates nested groups based on similarity. Widely used in bioinformatics, finance, and healthcare, these techniques uncover hidden patterns to optimize resources and precision medicine. However, big data's volume, velocity, and variety pose significant computational challenges for traditional algorithms. To overcome this, modern parallel clustering methods leverage distributed frameworks like Apache Spark and Map Reduce to efficiently process massive datasets [22]. Different approaches offer distinct advantages: partitioning methods are

fast but require precise setup; hierarchical models offer flexibility at a higher computational cost; and density- or grid-based techniques excel at managing noise and arbitrary shapes [23, 24]. Furthermore, extended fuzzy clustering utilizes sampling and incremental learning to process datasets exceeding standard memory capacities [25]. In engineering, clustering organizes data to support design optimization, quality control, and process planning, highlighting the need for scalable and efficient methods [26, 27]. Ultimately, ongoing research into scalable, parallel clustering algorithms continues to broaden their application, driving better data-driven decision-making across various disciplines [28, 29, 30].

2.4 Clinical clustering alzheimer's

In clinical, particularly Alzheimer's disease research, the use of integrated PLS-SEM and SOM for clustering clinical data offers an advanced methodological framework for patient stratification and pathway analysis. SOM aids in identifying distinct clinical clusters or phenotypes based on cognitive, pathological, and disability indicators, while PLS-SEM interprets causal mechanisms among biological and clinical variables within these clusters [31]. Such integration enhances understanding of disease heterogeneity, supports robust mediation analyses, and aids in developing precision diagnostics and targeted interventions. This approach addresses challenges with real-world health data, including sparsity and imbalance, by combining the data-driven strengths of unsupervised clustering and the theoretically informed causal modeling offered by PLS-SEM [31, 32].

3. Research methodology

3.1 Data source

Data for this retrospective cohort study were obtained from the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS), specifically utilizing the clinical data freeze collected between January 2024 and June 2025. To ensure analytical reliability and eliminate measurement noise in the subsequent PLS-SEM and SOM pipeline, the initial dataset underwent a rigorous inclusion and exclusion process. Listwise deletion was applied to handle missing values and uninformative clinical codes. Specifically, patient records were strictly excluded if they contained: (1) code "-4" (Not available/Not assessed), (2) code "9" (Unknown) across any cognitive, social, or functional variables, and (3) placeholder values for Body Mass Index (BMI) designated as "888.8". Following this rigorous preprocessing, the final analytical cohort consisted of n = 1,389 patients.

3.2 Research variables

In the context of this research, the variables suspected to influence the occurrence of AD are presented in Table 1.

3.3 Structural equation modeling (PLS-SEM)

The structural model, which describes the relationship between exogenous and endogenous latent variables, is formulated as follows:

$$\eta = \beta\eta + \gamma\xi + \zeta \tag{1}$$

The formula for exogenous latent variables with reflecting indicators is as follows:

$$X = \lambda\xi + \delta \tag{2}$$

And reflective indicators:

$$Y = \lambda\eta + \varepsilon \tag{3}$$

Both reflective and formative indicator models are evaluated using measurement models, also known as outer models. The validity and reliability of the outer model with reflected indicators are evaluated. Examining the indicators' cross-loading against their latent variables is how the evaluation is done. Convergent validity, on the other hand, stresses that there should be a strong correlation between indicators within a single latent variable [20]. The loading factor, or the connection between the item or component score and the latent variable score it creates, is used to evaluate convergent validity in PLS with reflective indicators. The loading factor should be larger than 0.50, preferably greater than 0.70, the Average Variance Extracted (AVE) should be greater than 0.50, and the communality should be greater than 0.70 [34]. The following formula can be used to determine the AVE value:

$$AVE = \frac{\sum_{i=1}^{I} \lambda_i^2}{\sum_{i=1}^{I} \lambda_i^2 + \sum_{i=1}^{I} var(\varepsilon_i)} \tag{4}$$

The significance of weight values and multicollinearity are the criteria used to evaluate the measurement model using formative indicators [35].

Table 1. Data dictionary of clinical indicators

Variable Laten	Indicator	Range	Skala
Risk Factors (RF)	Hypertension (X11)	0: None	Nominal
		1:Active	
		2:Inactive	
	Diabetes(X12)	0: None	Nominal
		1:Active	
		2:Inactive	
	Stroke(X13)	0: None	Nominal
		1:Active	
		2:Inactive	
	Obesity(X14)	10.0 – 100.0	Ratio
	Smoking (X15)	1:No	Nominal
		2:Yes	
Functional and Social Disability (FSD)	Depression Score (X21)	0 – 15	Ordinal
	Education Level (X22)	12:High School	Ordinal
		16: Bachelor	
		18: Master	
		20: PhD	
	Residence Status (X23)	1: Alone	Nominal
		2: Partner	
		3: Friend	
		4: Group	
		5: Other	
	Leisure Activities (X24)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
	Game (X25)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
Variabel Laten	Indikator	Range	Skala
Functional and Social Disability (FSD)	Game (X25)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
	Social Engagement (X26)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
Cognitive & Social Activities (CS)	Ability to Manage Finances (Y11)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
	Ability to Remember Important Dates (Y12)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
	Ability to Shop Independently (Y13)	0: Normal	Ordinal
		1: Difficult	
		2: Assistance	
		3: Dependent	
Clinical Dementia Rating (CDR)	Dementia Severity (Y21)	0.0: Normal	Ordinal
		0.5:Questionable	
		1.0: Mild	
		2.0: Moderate	
		3.0: Severe	

The statistic used to evaluate the inner model is the coefficient of determination R^2. The formula used is Eq. (5):

$$R^2 = \frac{\sum_{d=1}^{D}(\hat{Y}_d - \bar{Y})^2}{\sum_{d=1}^{D}(Y_d - \bar{Y})^2} \quad (5)$$

According to [36] the measurement model and the structural model's combined performance are validated using a single metric called Goodness of Fit (GoF). The following equation displays the formula to determine the GoF value:

$$GoF = \sqrt{\overline{AVE} \times \overline{R^2}} \quad (6)$$

The model's predictive power is verified using $Q-Square$ predictive relevance. According to the interpretation of the Q square predictive relevance result, the structural model isconsidered to fit the data or have pertinent predictions if the Q^2 value is close to 1 [4]. The following equation displays the Q-square formula:

$$Q^2 = 1 - (1 - R_1^2)(1 - R_2^2) \dots \left(1 - R_j^2\right) \quad (7)$$

Hypothesis testing was conducted using the Non-Parametric Bootstrapping method with 5,000 resamples. The goal was to obtain stable t-statistics and values to test the significance of each pathway. Mediation analysis was performed by calculating the Variance Accounted For (VAF) value to assess the strength of mediation of Cognitive & Social variables on the clinical severity level of Alzheimer's. VAF measures the extent of the indirect effect compared to the total effect.

$$VAF = \frac{a \times b}{(a \times b) + c'} \quad (8)$$

3.4 Clustering using self-organizing maps (SOM)

After the latent variables were constructed and validated through the PLS-SEM procedure in the previous stage, this study proceeded with cluster analysis using the Self-Organizing Maps (SOM) algorithm. The SOM algorithm was applied to map the multidimensional data of the latent variables into a lower-dimensional space in order to identify similar patterns or clinical phenotypes of Alzheimer's. The input used in the SOM algorithm consisted of latent variable scores generated from the PLS-SEM model in section 3.1, which included the Risk Factor Score, Cognitive and Social Activity Score, Functional and Social Disability Score, and Clinical Severity Score. These score data were normalized using z-score transformation to ensure each variable had a uniform range before being entered into the SOM network. The SOM architecture consists of two layers: an input layer and a competitive output layer. Each neuron j in the grid is associated with a weight vector w_j that has the same dimension as the input vector x. The learning process follows a competitive algorithm comprising three phases:

1. Competition**:** For each input vector x, the algorithm computes the Euclidean distance to all neurons. The neuron with the smallest distance is declared the winner or Best Matching Unit (BMU), denoted as c:

$$c = \arg\min_i \|x - \boldsymbol{m}_i\| \quad (9)$$

2. Cooperation: The BMU determines its topological neighbors within a specific radius defined by a neighborhood function.
3. Adaptation: The weights of the BMU and its neighbors are updated to move closer to the input vector.

3.5 Baseline clustering methods

To validate the superiority of the SOM framework in identifying distinct Alzheimer's phenotypes, this study employed several conventional clustering algorithms as baselines:

1. K-Means Clustering: A well-established partitioning algorithm that divides the dataset into a pre-defined number of clusters k by minimizing the within-cluster variance. The objective function, known as the Within-Cluster Sum of Squares (WCSS), is defined as:

$$J = \sum_{j=1}^{k} \sum_{i \epsilon S_j} \left\| x_i - \mu_j \right\|^2 \quad (10)$$

2. Hierarchical Clustering (Ward's Method): An agglomerative hierarchical approach that sequentially merges clusters to minimize the total intra-cluster variance. The increase in Error Sum of Squares (ESS) when merging clusters A and B is calculated as:
3.

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|m_A - m_B\|^2 \quad (11)$$

3.6 Cluster evaluation metrics

To objectively evaluate and compare the performance of the clustering algorithms, the

following internal validity and stability metrics were utilized:

1. Silhouette Coefficient: This metric assesses clinical separability by comparing intra-cluster cohesion with inter-cluster separation.

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \quad (12)$$

2. Davies-Bouldin (DB) Index: Evaluates clustering quality by measuring the ratio of within-cluster dispersion to between-cluster separation. Lower DB Index values signify better clustering, indicating compact and widely separated clusters:
3.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i+s_j}{d_{ij}}\right) \quad (13)$$

4. Cluster Stability (Jaccard Index): To ensure the derived phenotypes are robust against data perturbation and not mere statistical artifacts, clustering stability was measured across multiple resamples. The similarity between two cluster sets (C_1 and C_2) is quantified using the Jaccard Index:

$$J(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \quad (14)$$

3.7 Integrated PLS-SEM SOM Framework

The integration of PLS-SEM and SOM techniques enables researchers to capitalize on strengths from both approaches PLS-SEM's capability to model complex causal relationships and SOM's power in unsupervised data visualization and clustering. This combination facilitates enriched data exploration and validation, especially when handling heterogeneous populations or multidimensional constructs where traditional SEM might face challenges [38]. The integrated approach supports segmenting data into meaningful clusters identified via SOM, followed by causal relationship modeling and hypothesis testing through PLS-SEM, thus bridging exploratory and confirmatory analyses in a unified framework [39].

4. Results And discussion

The interaction between latent variables is illustrated by the structural equation model. By focusing on investigating the structural paths that connect latent constructs, hypothesis testing in this study is carried out through model specification, measurement, parameter estimation, and model evaluation.

4.1 Partial least squares structural equation modeling

4.1.1 Parameter estimation of the measurement and structural models in PLS-SEM

1. The exogenous latent variables, namely risk factors and functional and social disabilities, have coefficients as follows: Hypertension (0.190) ,diabetes (0.515) stroke (-0.650) , Obesity (0.417) , smoking (0.087) , depression score (-0.258) , education (0.383) , residential status (0.227) , leisure activities (0.820), gaming activities (0.548), and involvement in activities (0.655). Leisure activities (0.820) and diabetes(0.515)are the strongest indicators of the exogenous latent variables, while other indicators such as hypertension and depression score have low or negative contributions, indicating that not all indicators significantly reflect the latent variables.

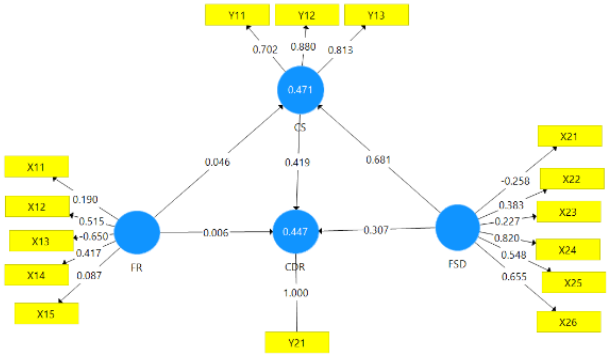

Figure. 1 Parameter estimation of the structural equation model

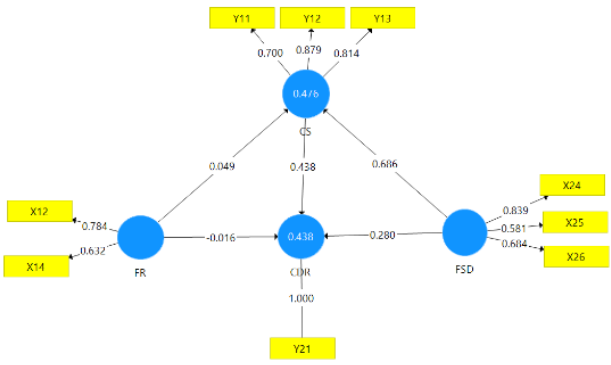

Figure. 2 Convergent validity and factor loadings of clinical indicators.

 doi: 10.22266/inassexpress.2026.003

2. The following are the coefficients for the indicators of the endogenous latent variables of cognitive and social factors, and clinical severity: Ability to Manage Finances (0.702) ,Ability to Remember Important Dates (0.880), Ability to Shop Independently (0.813) , and CDR (1.000) . According to these estimation results, these factors have a considerable impact on the, cognitive and social factors, as well as clinical severity.

4.1.2 Evaluation of the measurement model (Outer model)

1. Validitas Konvergen

Factor loadings for the remaining variables became significant after the non-significant indicators were eliminated from the PLS-SEM model.

All latent variable indicators

1. The exogenous latent variables, namely risk factors and functional and social disabilities, have coefficients as follows: diabetes (0.784) Obesity (0.632) ,leisure activities (0.839) , gaming activities (0.581), and involvement in activities (0.684).
2. The following are the coefficients for the indicators of the endogenous latent variables of cognitive and social factors, and clinical severity: Ability to Manage Finances (0.700), Ability to Remember Important Dates (0.879), Ability to Shop Independently (0.814) , and CDR (1.000).

2. Reliabilitas Konstrak dan AVE

Table 3 below presents a clear overview of the AVE values obtained for the latent variables when the self-concept indicators are excluded from the model.

Reliability testing is conducted to assess the internal consistency of a construct (latent variable), and it is more recommended to use Composite Reliability. In general, a construct is considered to have adequate reliability if the Composite Reliability value is ≥ 0.70 [3].

Table 2. Latent variable evaluation and construct reliability

Laten Variabel	AVE	*Composite Reliability*
Risk Factors	0.508	0.671
Functional and Social Disability	0.503	0.748
Cognitive & Social	0.642	0.842
Clinical Severity	1.000	1.000

Table 3. Discriminant validity using cross-loading criteria

	ξ_1	ξ_2	η_1	η_2
ξ_1	0.712			
ξ_2	0.043	0.709		
η_1	0.079	0.688	0.801	
η_2	0.031	0.581	0.630	1.000

Table 4. Inner model predictive relevance and explanatory power

Latent Variable	R^2	Q^2
Cognitive & Social	0.476	0.298
Clinical Severity	0.438	0.433

3. Validitas Diskriminan

Based on the cross-loading criteria, each indicator exhibits a higher correlation with its own latent variable than with other latent variables, thereby confirming that discriminant validity has been satisfied.

Discriminant Validity Cross-loading is used to assess discriminant validity testing on reflective indicators [16]. Discriminant validity is considered good if the correlation of each indicator with its own latent variable is higher than its correlation with other latent variables.

4.1.3 Evaluation of the structural model (Inner model)

Several metrics are evaluated, such as predictive significance determined by the coefficient of determination for endogenous latent variables. In Table 5, nilai $Q^2 > 0$ values for all endogenous constructs are greater than zero, indicating adequate predictive relevance of the model. The R^2 values

Table 5. Path coefficients and hypothesis testing results

Hyp	Path	β	SE	T	Result
H1	RF → CSA	0.049	0.021	2.356	Accept
H2	RF →CDR	-0.016	0.021	0.759	Reject
H3	FSD → CSA	0.686	0.026	26.104	Accept
H4	FSD → CDR	0.280	0.044	6.431	Accept
H5	CSA →CDR	0.438	0.038	11.588	Accept

Table 6. Confidence intervals for structural paths

Path	β	2.5%	97.5%
RF⟶ CSA	0.049	0.009	0.090
RF ⟶CDR	-0.016	-0.051	0.029
FSD ⟶ CSA	0.686	0.636	0.739
FSD ⟶ CDR	0.280	0.194	0.365
CSA ⟶CDR	0.438	0.365	0.511

indicate moderate explanatory power for both Cognitive & and Clinical Severity, suggesting that a substantial proportion of variance in these constructs is explained by the model.

4.1.4 Path coefficients and hypothesis testing

Hypothesis testing was conducted using a bootstrapping procedure with 5.000 sub-samples to assess the significance of the structural paths. The significance of the relationships was determined based on T-statistics and P-values, where a T-statistic greater than 1.96 indicates a supported hypothesis. The proposed hypotheses for this study are as follows:

1. H1: Risk Factors significantly influence Cognitive and Social Activities.
2. H2: Risk Factors significantly influence Clinical Severity.
3. H3: Functional and Social Disability significantly influences Cognitive and Social Activities.
4. H4: Functional and Social Disability significantly influences Clinical Severity.
5. H5: Cognitive and Social Activities significantly influence Clinical Severity.

The results show that Risk Factors (RF) influence Cognitive-Social Activity (H1) but do not have a direct impact on the Level of Clinical Severity (H2), indicating the presence of an indirect pathway of influence. Conversely, Functional Disability (FSD) emerges as the dominant predictor that significantly increases clinical severity (H4) and triggers a decline in social activity (H3). These findings are reinforced by the significant effect of Cognitive-Social Activity on clinical severity (H5), confirming that a decline in social interaction markedly worsens patients' symptoms.

The 95% Confidence Interval (CI) results further support the hypothesis testing. For the supported paths (H1, H3, H4, and H5), the interval between 2.5% and 97.5% does not include zero, confirming that the effects are statistically significant. Conversely, the path from RF to CDR (H2) includes zero within its interval [-0.051, 0.029], indicating a non-significant relationship.

Table 7. Mediation analysis of cognitive and social factors

Mediation	Paths	Original Path
Direct	RF⟶ CSA	0.049
	RF ⟶CDR	0.006
	FSD ⟶ CSA	0.686
	FSD ⟶ CDR	0.581
	CSA ⟶CDR	0.438
Indirect	RF⟶ CSA⟶ CDR	0.022
	FSD ⟶ CSA ⟶CDR	0.301
Total Indirect	RF ⟶CDR	0.022
	FSD ⟶ CDR	0.301

4.1.5 Mediation analysis

This analysis examines the mediating roles of Cognitive and social factors. Based on the path analysis, it was found that the direct effect of Risk Factors (RF) on CDR is positive, at 0.006. Although this direct effect is quite small, the indirect effect through the CSA mediator is recorded at 0.022. The VAF calculation result for the RF pathway reaches 78.6%, which falls within the range of partial mediation but is close to the threshold for full mediation (>80%). This indicates that most of the effect of risk factors (RF) on the severity of dementia (CDR) does not occur directly, but rather through a decline in cognitive and social function (CSA). Meanwhile, the variable Functional and Social Disability (FSD) has a total effect that is positive and significant on CDR. This indicates that the higher the level of functional and social difficulties experienced by an individual (scores approaching 3), the higher the difficulty scores on the CSA and CDR factors. The VAF calculation

Table 8. Clustering performance comparison across algorithms

Data Type	Clustering	Silhouette	DB index	Stability
Raw Indicators	K Means	0.3603	1.7355	0.754
	Hierarchical Ward method	0.3066	2.0241	0.892
	SOM	0.3889	1.4100	0.558
Latent Variable Scores (PLS-SEM)	K Means	0.5910	1.053	0.504
	Hierarchical Ward method	0.6051	0.971	0.706
	SOM	0.6181	0.892	0.825

 doi: 10.22266/inassexpress.2026.003

result for the FSD variable is 34.1%, which falls within the 20%-80% range, confirming the occurrence of Partial Mediation. This means that functional disability (FSD) can worsen the condition of dementia (CDR) through two pathways: directly due to physical limitations, and indirectly through a decline in cognitive and social abilities.

4.2 Clustering results and validation

Following the validation of the structural model PLS-SEM, the latent variable scores for each respondent were extracted to serve as the primary input for the clustering analysis. This sequential approach ensures that the identified phenotypes are derived from statistically validated constructs, effectively filtering out clinical noise and measurement errors inherent in raw indicators. To establish methodological superiority and ensure the robustness of the findings, a two-fold comparative analysis was conducted. First, clustering performance was evaluated using raw clinical indicators as a baseline. Second, the same algorithms were applied to the latent variable scores to demonstrate the efficacy of the integrated PLS-SEM and SOM framework in enhancing patient segmentation quality.

As presented in Table 9, the comparative analysis demonstrates that clustering algorithms applied directly to raw clinical indicators show suboptimal performance marked by vulnerability to noise and high dimensionality. This is evidenced by low Silhouette coefficients (all below 0.40) and elevated Davies-Bouldin (DB) indices (all exceeding 1.40), indicating poor cluster separation and cohesion. Although the Hierarchical Ward method showed high stability (0.892) on raw data, it had the lowest Silhouette value (0.3066), implying consistent replication of clusters with poor separation and high overlap, thus clinically less meaningful.

In contrast, employing latent variable scores extracted via PLS-SEM as input features markedly improved clustering quality in all baseline algorithms. These latent scores, derived from statistically validated constructs, effectively reduced clinical noise and measurement errors from raw indicators. The Self-Organizing Maps (SOM) algorithm, when applied to the latent variable scores, outperformed both K-Means and Hierarchical Ward's method across all evaluation indices. Specifically, SOM recorded the highest Silhouette coefficient (0.6181), signifying precise clinical phenotype formation with clear separation boundaries, and the lowest DB index (0.8920), indicating optimal intra-cluster cohesion and inter-cluster separation. Additionally, SOM showed the strongest reproducibility with a Jaccard Stability Index of 0.825, outperforming K-Means (0.504) and Hierarchical Ward (0.706) methods under the latent score condition.

These findings collectively validate that integrating PLS-SEM to extract latent variables prior to clustering effectively enhances patient segmentation quality by filtering measurement noise and modeling complex non-linear heterogeneity. Moreover, the SOM framework provides methodological superiority in clustering clinical phenotypes, achieving improved cluster separation, cohesion, and stability over conventional clustering algorithms applied either on raw data or latent scores. This validates the application of the combined PLS-SEM and SOM approach as a robust and reliable method for patient phenotyping in heterogeneous diseases such as Alzheimer's disease [40, 41].

4.2.1 SOM training and visualization

The U-Matrix (Unified Distance Matrix) visualization illustrates the topological distance between adjacent neurons on the SOM grid. As depicted in Figure 3, the vast majority of the map is dominated by green nodes, which represent areas of low inter-neuron distance and high data homogeneity. This extensive "valley" accurately reflects Cluster 1 (Mild Phenotype), encompassing 85.8% of the patient cohort, indicating that the majority of patients share highly similar, low-severity clinical profiles.

In stark contrast, the top edge of the map features isolated neurons transitioning into yellow, orange, and white hues, representing exceptionally

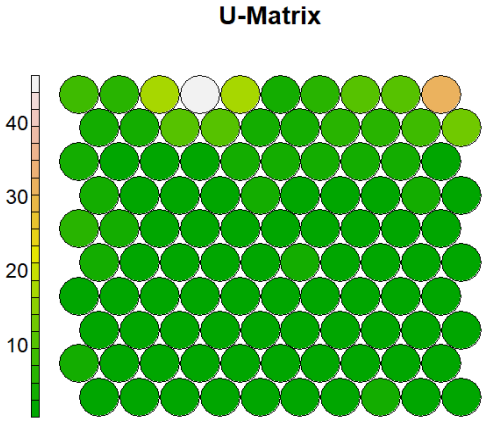

Figure. 3 Topological distance visualization using SOM U-Matrix

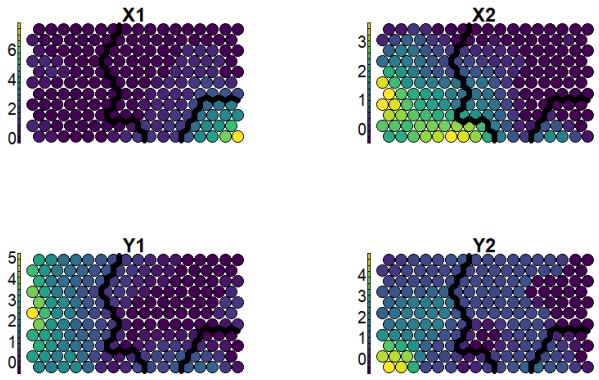

Figure. 4 Component planes analysis across clinical latent constructs

high distance values (exceeding 40). These high-value nodes act as topological boundaries or "walls" that distinctly separate the extreme outlier cases namely Cluster 2 (Severe Phenotype) and Cluster 3 (High-Risk Phenotype) from the general population. This massive distance gradient visually validates the sharp clinical escalation present in the severe and high-risk phenotypes, demonstrating the SOM's efficacy in detecting and isolating minority subgroups with highly divergent disease manifestations.

4.2.2 Component planes analysis

The component planes analysis reveals distinct spatial distribution patterns across the SOM grid for the four latent constructs: Risk Factors (X1), Functional and Social Disability (X2), Cognitive & Social Activities (Y1), and Clinical Severity (Y2). The superimposed black boundary lines clearly delineate the three identified clusters.

The largest topological region (Cluster 1), covering the upper and right sections of the map, displays uniformly dark purple hues across all planes, visually confirming the low disease burden of the Mild Phenotype. In the bottom-left region (Cluster 2), there is a strong co-localization of high values (indicated by yellow and light green hues) across the X2, Y1, and Y2 planes. This spatial convergence visually confirms the PLS-SEM structural findings, demonstrating that patients with severe functional impairments concurrently exhibit high cognitive decline and clinical severity. Conversely, the bottom-right region (Cluster 3) displays an isolated peak of high values strictly within the X1 (Risk Factors) plane, while remaining dark in the clinical severity planes. This configuration perfectly isolates the high-risk demographic subgroup that possesses severe underlying risk factors but has not yet manifested severe clinical dementia.

4.2.3 Cluster identification and validation

Based on the optimization of SOM mapping and latent score extraction, the 1,389 Alzheimer's patients in this study were classified into three clinically distinct phenotypes:

1. Cluster 1 (Mild Phenotype, n = 1,193): This cluster encompasses the vast majority of the cohort (85.8%). Patients in this group exhibit the lowest average age (70.50 years) and display negative mean Z-scores across all clinical latent variables. This indicates a very low disease burden regarding risk factors, functional impairment, cognitive decline, and clinical severity. These findings suggest a milder disease manifestation with a favorable prognosis, manageable with standard care protocols.
2. Cluster 2 (Severe Phenotype, n = 157): Consisting of 11.3% of the cohort, this group represents patients with the oldest average age (73.90 years). This phenotype is characterized by a sharp and critical escalation in disease burden, demonstrating exceptionally high scores in Functional and Social Disability (2.27), Cognitive-Social decline (1.92), and Clinical Severity (1.64). The massive effect sizes in these domains confirm a clinically distinct, high-severity cohort requiring intensive medical intervention.
3. Cluster 3 (High-Risk Phenotype, n = 39): Representing a small minority (2.8%) with an average age of 70.10 years, this subgroup is uniquely distinguished by an extraordinarily high Risk Factor score (4.43). Despite this extreme risk profile, their functional impairment (0.174), cognitive decline (0.284), and clinical severity (0.055) remain only slightly above the

Table 9. ANOVA results and effect sizes for cluster validation

Path	F Palue	P Value	η^2
Age	10.06	$4.5e-05$	0.0143
Risk Factors	917.7	$<2e-16$	0.5698
Functional and Social Disability	1355	$<2e-16$	0.6617
Cognitive & Social	631.9	$<2e-16$	0.4770
Clinical Severity	366.7	$<2e-16$	0.3460

cohort average. This suggests a transitional or high-risk phenotype where underlying risk factors (such as severe hypertension or obesity) are highly elevated, but extensive clinical dementia has not yet fully manifested.

4.2.4 Statistical Validation of Clusters

The ANOVA results indicate that all variables exhibit highly significant differences across clusters ($p < 0.001$), substantiating the clusters' statistical legitimacy. However, given the large sample size ($N = 1389$), reliance on $p - values$ alone can be misleading, as even negligible differences may attain statistical significance. To address this limitation, effect sizes were quantified using Eta Squared (η^2), which expresses the proportion of total variance in each variable explained by cluster membership [42, 43].

Following Cohen's seminal benchmarks [42] effect sizes interpreted via Eta Squared (η^2) broadly correspond to the following thresholds: Small effect ($\eta^2 \approx 0.01$), Medium effect ($\eta^2 \approx 0.06$), and Large effect ($\eta^2 \approx 0.14$). These thresholds serve as practical guidelines widely adopted in clustering validation to assess the substantive significance of group differences beyond mere statistical significance [43]. Applying these criteria to the clinical results reveals critical insights:

1. Age ($\eta^2 = 0.0143$) : Despite achieving statistical significance, Age's effect size is decidedly small, accounting for approximately 1.4% of the variance between clusters. This minimal proportion suggests that Age exerts a negligible influence on cluster differentiation and is not a principal factor driving the formation of the phenotypes.
2. Risk Factors ($\eta^2 = 0.5698$) and Functional & Social Disability ($\eta^2 = 0.6617$): Both variables reveal extraordinarily large effect sizes, explaining over 56% and 66% of the variance, respectively. These values far exceed conventional large-effect benchmarks, marking RF and FSD as the dominant clinical dimensions underpinning the clustering structure. Such massive effect sizes imply robust discriminative capacity and reinforce their centrality in defining the disease phenotypes.
3. Cognitive & Social Activities ($\eta^2 = 0.4770$): CSA demonstrates a large effect size consistent with strong between cluster variability, elucidating nearly 48% of the variance. This underscores its significant contribution to internal cluster separability.
4. Clinical Severity ($\eta^2 = 0.3460$) : CDR's medium to large effect size indicates meaningful, albeit comparatively lower, cluster discrimination, accounting for roughly 35% of the variance.

While η^2 provides a useful measure of explained variance, it is essential to acknowledge potential limitations, including upward bias in small samples [44]. Although omega squared (ω^2) is often considered a more conservative estimate, η^2 remains highly robust and acceptable in one-way ANOVA for effect size reporting, particularly in large cohorts like this study [45]. The conjoint presence of highly significant ANOVA results and substantial effect sizes for the four clinical variables confers strong empirical support for the internal validity of the clustering solution. The data clearly indicate that cluster separation is principally driven by the latent disease constructs (RF and FSD), with complementary contributions from cognitive decline (CSA) and severity (CDR). Importantly, the fact that Age does not materially differentiate the clusters highlights the biological and pathological—rather than merely demographic—nature of the cluster distinctions. Furthermore, these quantitative validations align with the topographic precision observed in the SOM U-Matrix, confirming that the clusters represent meaningful phenotypic subtypes within Alzheimer's disease progression, not mere artifacts of statistical noise.

5. Conclusion

This study demonstrates the advantages of an integrated analytical approach by combining PLS-SEM and SOM to map the clinical phenotypes of Alzheimer's patients. The main findings of this study highlight that using latent variable scores as clustering inputs is superior to using raw clinical indicators. The proposed approach successfully improves segmentation quality significantly, as evidenced by an increased Silhouette Score of 0.6181 and the achievement of very high topographic precision with a Quantization Error (QE) of 0.0865. Through this integration, the model is able to filter out clinical noise and capture complex non-linear heterogeneity, resulting in a more stable and statistically valid three-cluster classification (Mild, Severe, and High Risk Phenotypes). The results of ANOVA and effect size (η^2) analyses confirm that the resulting phenotypes have profound clinical relevance, wherein functional disability is identified as a primary driver operating through cognitive-social pathways. Overall, this

methodology offers a robust and promising framework in patient segmentation.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization; Wahyuni and Bambang Widjanarko Otok, methodology; Wahyuni and Jerry Dwi Trijoyo Purnomo; software, formal analysis, data curation, writing original draft preparation, Wahyuni; writing review and editing, Bambang Widjanarko Otok and Jerry Dwi Trijoyo Purnomo; supervision, Bambang Widjanarko Otok and Jerry Dwi Trijoyo Purnomo.

Acknowledgments

The authors would like to thank the National Alzheimer's Coordinating Center (NACC) for providing the clinical dataset used in this research.

References

[1] J. D. Grill *et al.*, "Disclosing risk factors to individuals without cognitive impairment," *Pract. Neurol.*, vol. 18, pp. 38-41, 2019.

[2] M. Nemoto *et al.*, "Combining multimodal behavioral data of gait, speech, and drawing for classification of Alzheimer's disease and mild cognitive impairment," *J. Alzheimers Dis.*, vol. 84, no. 1, pp. 315-327, 2021.

[3] S. Kono and M. Sato, "The potentials of partial least squares structural equation modeling (PLS-SEM) in leisure research," *J. Leisure Res.*, vol. 54, no. 3, pp. 371-395, 2022.

[4] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed. Thousand Oaks, CA, USA: Sage, 2017.

[5] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," *Eur. Bus. Rev.*, vol. 31, no. 1, pp. 2-24, 2019.

[6] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 4th ed. New York, NY, USA: Guilford Press, 2016.

[7] W. J. Reinartz, M. Haenlein, and J. Henseler, "An empirical comparison of the efficacy of covariance-based and variance-based SEM," *Int. J. Res. Mark.*, vol. 26, no. 4, pp. 332-344, 2009.

[8] N. F. Richter, R. R. Sinkovics, C. M. Ringle, and C. Schlägel, "A critical look at the use of SEM in international business research," *Int. Mark. Rev.*, vol. 33, no. 3, pp. 376-404, 2016.

[9] M. Sarstedt, C. M. Ringle, and J. F. Hair, "Partial least squares structural equation modeling," in *Handbook of Market Research*, C. Homburg, M. Klarmann, and A. Vomberg, Eds. Cham, Switzerland: Springer, 2017, pp. 1-40.

[10] A. E. Legate, J. F. Hair, J. L. Chretien, and J. J. Risher, "PLS-SEM: Prediction-oriented solutions for HRD researchers," *Hum. Resour. Dev. Q.*, vol. 34, no. 1, pp. 91-110, 2021.

[11] Y. Haji-Othman and M. S. S. Yusuff, "Assessing reliability and validity of attitude construct using partial least squares structural equation modeling (PLS-SEM)," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 12, no. 5, pp. 378-385, 2022.

[12] M. Sarstedt, C. M. Ringle, D. Smith, R. Reams, and J. F. Hair, "Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers," *J. Fam. Bus. Strategy*, vol. 5, no. 1, pp. 105-115, 2014.

[13] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin, Germany: Springer, 2001.

[14] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586-600, 2000.

[15] N. Liu, J. Wang, and Y. Gong, "Deep Self-Organizing Map for visual classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2015, pp. 1-6.

[16] H. Wold, "Soft modeling: The basic design and some extensions," in *Systems under Indirect Observation: Causality, Structure, Prediction*, K. G. Jöreskog and H. Wold, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 1-54.

[17] D. Miljkovic, "Brief review of self-organizing maps," in *Proc. 40th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. (MIPRO)*, 2017, pp. 1061-1066.

[18] P. Stefanovič and O. Kurasova, "Visual analysis of self-organizing maps," *Nonlinear Anal. Model. Control*, vol. 16, no. 4, pp. 488-504, 2011.

[19] K. Tasdemir and E. Merényi, "Exploiting data topology in visualization and clustering of self-organizing maps," *IEEE Trans. Neural Netw.*, vol. 20, no. 4, pp. 549-562, 2009.

[20] Z. Wang *et al.*, "Hardware implementations of concurrent self-organizing maps," *IEEE Access*, vol. 10, pp. 12345-12356, 2022.

[21] V. E. Neagoe and A. D. Ropot, "Concurrent self-organizing maps for pattern classification," in *Proc. 1st Int. IEEE Symp. Intell. Syst.*, 2002, pp. 304-309.

[22] Z. Dafir, Y. Lamari, and S. C. Slaoui, "A survey on parallel clustering algorithms for Big Data," *Artif. Intell. Rev.*, vol. 54, no. 4, pp. 2411–2443, 2020.

[23] V. W. Ajin and L. D. Kumar, "Big data and clustering algorithms," in *Proc. IEEE Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, 2016, pp. 1–5.

[24] T. Alasali and Y. Ortakci, "Clustering techniques in data mining: A survey of methods, challenges, and applications," *Comput. Sci.*, 2024.

[25] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means algorithms for very large data," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1130–1146, 2012.

[26] M. A. Mahdi, K. M. Hosny, and I. Elhenawy, "Scalable clustering algorithms for big data: A review," *IEEE Access*, vol. 9, pp. 80015–80027, 2021.

[27] D. T. Pham and A. A. Afify, "Clustering techniques and their applications in engineering," *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 221, no. 11, pp. 1445–1459, 2007.

[28] R. He *et al.*, "Clustering enabled wireless channel modeling using big data algorithms," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 177–183, 2018.

[29] S. Pitafi, T. Anwar, and Z. Sharif, "A taxonomy of machine learning clustering algorithms, challenges, and future realms," *Appl. Sci.*, vol. 13, no. 6, p. 3529, 2023.

[30] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big data clustering: A review," in *Int. Conf. Comput. Sci. Appl.*, Springer, 2014, pp. 707–720.

[31] F. Mandreoli *et al.*, "Clustering and interpretation of clinical data using PLS-SEM and SOM," *J. Biomed. Inform.*, vol. 125, p. 103964, 2022.

[32] N. K. Avkiran, "An in-depth discussion and illustration of partial least squares structural equation modeling in health care," *Health Care Manage. Sci.*, vol. 21, pp. 401-408, 2017.

[33] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *J. Mark. Res.*, vol. 18, no. 1, pp. 39-50, 1981.

[34] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Multivariate Data Analysis*, 6th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2006.

[35] M. Mehmetoglu and S. Venturini, *Structural Equation Modeling with Partial Least Squares Using Stata and R*. Boca Raton, FL, USA: CRC Press, 2021.

[36] M. Rodliyah, "Estimasi Score Factor dengan Partial Least Square (PLS) pada Measurement Model," M.S. thesis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, 2016.

[37] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *J. Acad. Mark. Sci.*, vol. 43, no. 1, pp. 115-135, 2015.

[38] E. E. Rigdon, C. M. Ringle, and M. Sarstedt, "Structural modeling of heterogeneous data with partial least squares," *Rev. Ind. Organ.*, vol. 38, pp. 255-296, 2010.

[39] O. Sohaib *et al.*, "Integrating PLS-SEM and neural networks for predictive modeling," *Expert Syst. Appl.*, vol. 120, pp. 1-15, 2019.

[40] L. Gonçalves *et al.*, "SOM clustering of clinical data," *Artif. Intell. Med.*, vol. 42, no. 1, pp. 51-67, 2008.

[41] P. Kassomenos *et al.*, "Clustering of health data using self-organizing maps," *Environ. Res.*, vol. 109, pp. 222-230, 2009.

[42] H. Y. Kim, "Statistical notes for clinical researchers: Effect size," *Restor. Dent. Endod.*, vol. 41, no. 4, pp. 312-314, 2016.

[43] R. Norouzian and L. Plonsky, "Eta- and partial eta-squared in L2 research: A cautionary review and guide to more precise reporting," *Second Lang. Res.*, vol. 34, no. 2, pp. 257-271, 2017.

[44] P. D. Bliese and R. R. Halverson, "Group size and measures of group-level properties: An examination of eta-squared and ICC values," *J. Manage.*, vol. 24, no. 2, pp. 157-172, 1998.

[45] K. Okada, "Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA," *Behaviormetrika*, vol. 40, no. 2, pp. 129-147, 2013.