Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2022;8(4):395-409

Original Article

Evaluating the Agreement between K-Median and Latent Class Analysis for Clustering of Psychological Distress Prevalence

Maryam Salari[1,2], Zahra Rahimi[1], Reza Kalantari[3], Jamshid Jamali[1,2*]

[1]Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
[2]Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.
[3]Department of Ergonomics, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran.

ARTICLE INFO

Received 21.08.2022
Revised 18.09.2022
Accepted 23.10.2022
Published 15.12.2022

Key words:
Cross-sectional study;
Psychological distress;
K-6 questionnaire;
Latent class analysis;
K- media;
Mashhad;

ABSTRACT

Introduction: Psychological distress (PD) is one of the most common mental disorders in the general population. Psychological distress is considered a public health priority due to its adverse effects on quality of life, health, performance, and productivity. It can also predict several serious mental illnesses, such as depressive disorder and anxiety. In this study, we intend to identify the behavioral pattern of PD in the population of 18 to 65 years old in Mashhad using two methods, K-median and Latent Class Analysis (LCA), and evaluate the agreement between the two methods.

Methods: This cross-sectional study was performed on 38058 individuals referred to community health care centers in Mashhad of Iran in 2019. The information used in this study was extracted from Sina Electronic Health Record System (SinaEHR) database. A demographic information checklist and a 6-item Kessler psychological distress scale (K-6) were used for data collection. K-median and LCA were used for data analysis.

Results: Out of 38058 participants, 49.3% were women, 86.1% were married, and 63.6% had a diploma and under diploma education. The LCA identified three patterns of PD in answering the items of the K-6 questionnaire, including severe PD (19.7%), low PD (36.7%), and no PD (43.5%). Three clusters were identified by the K-Median method: 1) severe PD (22.0%), 2) low PD (31.1%), and 3) and no PD (46.9%). The agreement between K-Median and LCA was kappa = 0.862.

Conclusion: About 20% of people were classified as having severe PD. Both LCA and k-median methods can reasonably identify the latent pattern of PD with significant entropy, and there was almost complete agreement between the two methods in data clustering. Considering the advantages of the LCA, this method is recommended to identify the latent pattern of PD based on the k-6 questionnaire.

*.Corresponding Author: jamali@mums.ac.ir

Introduction

The current stressful life and exposure to social and economic stresses can increase the prevalence of psychological problems. Mental health disorders are among the top five causes of disabilities in the world. These disorders are strong predictors of death from heart disease, stroke, and cancer.1 According to studies conducted before the COVID-19 pandemic, the prevalence of mental disorders in different countries has varied from 13 to 22%.[1] More than one billion people worldwide were affected by mental disorders in 2016, accounting for 7% of the global burden of disease and 19% of all years of living with a disability.[2] The prevalence of mental disorders in Iran is estimated to be between 10.1% and 50.3[-7] The outbreak of the COVID-19 pandemic poses serious threats to human physical and mental health. Implementation of preventive measures to prevent COVID-19, including quarantines and distancing, has reduced social interactions. These measures also increased social isolation and mental disorders, which have affected many aspects of people's lives.[8-10] Several studies suggest that psychological symptoms have been observed in quarantined individuals, including signs of depression, stress, and anxiety.[11]

The prevalence of depression, anxiety, psychological distress (PD), and insomnia among 66 studies with 221970 participants was estimated to be 31.4%, 31.9%, 41.1%, and 37.9%, respectively.[12] According to the Global Burden of Disease Study 2016 (GBD 2016), Depression and Anxiety from 2005 to 2016 were among the top ten causes of Iranians losing their lives due to disability.[13] The prevalence of anxiety, depression, and PD during the Covid-19 epidemic in the general population was reported to be 38.12%, 40.13%, and 37.54%, respectively.[14] Stress, anxiety, depression, and PD are the most common mental disorders in the general population. These disorders are associated with decreased quality of life. Due to the harmful effects of mental disorders on health, performance, and productivity, attention to them is a public health priority.[1]

PD refers to the non-specific symptoms of minor psychiatric disorders such as stress, anxiety, and depression that are used as indicators of mental health in demographic and epidemiological studies; most of these disorders are affected by PD.[15] High PD levels indicate an impaired mental health disorder.[16, 17] PD is an emotional state characterized by symptoms of depression (e.g., sadness, helplessness, hopelessness, worthless) and anxiety (restless, nervous).[7] PD is used for describing a short but acute period of a specific mental disorder that first presents with features of depression or anxiety. It can be considered a type of abnormality responsible for maladaptive thinking and behavior and requires specialized intervention.[18]

The prevalence of PD in India, Japan, the USA, Canada, Australia, and Egypt is estimated at 20.2%, 6.7%, 3.4%, 12%, 11.1%, and 72.2%, respectively.[17, 19-22] The prevalence of PD in the COVID-19 pandemic among Japanese pregnant women was 16.5%, and among adults living in Nairobi, Keny was 52.8%.[23, 24] This rate was estimated at 26.2% for elderly patients in Vietnam.25 The prevalence of PD among 1468 US adults in April 2020 was 13.6%, while in 2018, it was reported to be 3.9%.[26] A similar study in the United States found that PD increased significantly with the COVID-19 pandemic.[27] After the COVID-19 epidemic,

among the 52730 Chinese, approximately 35% experienced PD, of which more than 5% had severe PD.10 The prevalence of PD among Chinese university students was estimated to be 90.86%.[28] A study in Italy also reported an increase in the percentage of high PD levels compared to European epidemiological statistics.[29] A 2021 meta-analysis study examining the prevalence anxiety disorder symptoms during coronary heart disease; The prevalence of PD was estimated to be 13.29%.[30] To our knowledge, there have been limited studies on psychological distress in Iran. The prevalence of psychological distress in Iran has been reported to vary from 10.10% to 61.5% In two studies.[5, 31]

Epidemiological studies of PD play an essential role in determining the general mental health status of the community and identifying the demographic factors associated with it. The role of such studies in estimating the necessary resources to provide better health services in the country is critical. Community health care centers as the primary level of prevention can play a decisive role in the diagnosis, care, and treatment of high-risk mental health groups in society.

Determining the cut-off points and scoring method in questionnaires that assess mental disorders, including PD, is challenging due to the lack of a gold standard and confuses the researchers. The method of scoring and determining the cut-off point of the questionnaires depends on the culture, characteristics of each region, the statistical population of the study, and the time of the survey. This difference in the selection of cutting points leads to different and sometimes contradictory results. Today, the determination of the cut point by researchers is the most common method for classifying PD levels based on the K-6 questionnaire. Different cut-off points have been presented for the K-6 questionnaire, including points 9, 10, and 13.[15, 20, 32-34]

Clustering is the process of dividing data into groups of similar objects. Each group, known as a cluster, consists of objects that are similar to each other but dissimilar to objects in other clusters. This simplifies the data by reducing the amount of detail, which is similar to lossy data compression. Clustering models data by its clusters and has roots in mathematics, statistics, and numerical analysis. From a machine learning perspective, clusters represent hidden patterns and the search for them is unsupervised learning. Data mining involves working with large databases. Clustering is a common initial step in data mining and analysis, which helps to identify groups of related records that can be used as a foundation for further investigation. This method assists in the creation of population segmentation models, such as customer segmentation based on demographics. Further analysis with standard analytical and data mining techniques can then be used to determine the characteristics of these segments in relation to a desired outcome.[35] which imposes additional computational requirements on clustering analysis. Clustering can be a handy and valid statistical tool if appropriately used and results can be used as a starting point for defining hypotheses and planning future studies such as building more advanced statistical analysis, preferably on an independent data set. Clustering refers to a very general and broad set of methods for finding subgroups or clusters in a dataset, such that points in any one group are more "similar" to each other than to issues in another group. This

may be done to provide a good "summary" of data, look for new insights into the structure of the data, Find homogeneous subgroups among the observations, or to arrange the clusters into a natural hierarchy and Investigate the validity of pre-existing groups.[36]

It is important not to confuse clustering with classification. In classifications which is a supervised learning method we have data for which the groups are known, and we try to learn what differentiates these groups to properly classify future data. In clustering, we look at data for which groups are unknown and undefined and try to learn the groups themselves. Here no response is defined yet and we are just exploring the data. There are many popular clustering methods such as k-median, k-means, and Hierarchical clustering. On the other way, there are some model-based clustering such as Latent Class Analysis (LCA). Similar to the traditional cluster analysis techniques, the objective of LCA is to identify unobserved subgroups comprised of similar individuals. Unlike traditional cluster analysis, however, LCA is a model-based approach to clustering. It identifies subgroups based on posterior membership probabilities rather than somewhat ad hoc dissimilarity measures such as Euclidean distance.[37]

In this study, we use LCA and K-median to diagnose PD based on the pattern of responding to the K6 questionnaire options. These two methods compete with the traditional scoring method. They are based on minimizing the difference of similarity (or homogeneity) observations within clusters and maximizing the difference of heterogeneous observations between clusters. These advanced statistical methods can provide more accurate and valid results concerning latent concepts.

Methods

This cross-sectional study was performed on 38058 individuals referred to community health care centers at Mashhad University of Medical Sciences of Iran in 2019. The information used in this study was extracted from Sina Electronic Health Record System (SinaEHR) database. Sina system has been used since 2016 to electronically record the health records of participants who were referred to community health care centers in Khorasan Razavi, Iran. Age between 18 and 65 years and willingness to participate in the study were involved in the inclusion criteria. Failure to respond to at least four items were included in the exclusion criteria.

Data collection tools in this study were a demographic information checklist and 6-item Kessler psychological distress scale (K-6) questionnaire. The K-6 is a truncated form of k-10, introduced in 2002 by Kessler et al. as a well–known and suitable screening instrument for assessing PD in general populations.[16] The k-6 scale asks participants how frequently they felt sadness, restless, nervous, helpless, hopeless, and worthless over the past month.[16] Items can be scored using a 5-point Likert-type scale, including *none of the time* (0), A *little of the time* (1), *Some of the time* (2), *Most of the time* (3), *and All of the time* (4). A composite scale was calculated, with raw scores ranging from 0 to 24 and greater scores indicating higher levels of PD.[16] The validity and reliability of questionnaire K-6 in Iran (with Cronbach's alpha 0.98, sensitivity 0.73, specificity 0.78, and positive predictive value of 0.52) have been confirmed.[34] In this study, Cronbach's α was 0. 869, which shows acceptable Internal consistency. Previous studies have shown that

both versions of the Kessler PD Scale (K-10 and K-6) have better validity and reliability than the 12-item General Health Questionnaire (GHQ-12).[16, 34]

Confidentiality of identity information, non-disclosure of individuals 'names, lack of prejudice, and interference of the questioners' tendencies have been among the ethical issues considered in this research. This study received approval from the Research Ethics Committee of Mashhad University of Medical Science with the ethics code of IR.MUMS.REC.1399.607.

Qualitative variables were expressed in frequency and percentage, quantitative variables in mean ± standard deviation. Two clustering methods of LCA and the k-median technique were used to identify different patterns of PD. LCA is a multivariate technique that can be used to cluster discrete variables. This approach can be used to find homogeneous subsets of a community according to the pattern of responding to items in a questionnaire.[38]

The k-median method is one of the separation clustering procedures which origins in graph theory.[39] In this method, homogeneous clusters are identified using the minimization of absolute deviations, equal Manhattan distance. In k-median clustering, the total intra-cluster variation (or total within-cluster variation) is minimized and the distance of the points from the center of the cluster is the minimum.[40]

The number of clusters (classes) in these two methods is determined using statistical criteria and the researcher's opinion. Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), entropy, and Lo-Mendell-Rubin test (LMR) are among the most popular and widely used methods for determining the optimal number of classes in model-based clustering such as LCA.[41] In determining the number of clusters, the interpretability of clusters should also be considered.[42] The gap statistic, Silhouette, and Elbow methods have been widely used more than other methods for determining the optimal number of classes in K-median.[43]

The agreement between the clustering results of LCA and K-median was evaluated with Cohen's kappa coefficient. According to convention, a κ-value of 0 to 0.2 indicates slight agreement; 0.2 to 0.4 fair agreement; 0.4 to 0.6 moderate agreement; 0.6 to 0.8 substantial agreement; and 0.8 to 1.0 almost perfect agreement.44 Statistical analysis was conducted using the SPSS version 26 and LatentGold version 5 software. A P-value of less than 0.05 was considered significant.

Results

The majority of participants were women (49.3%), married (86.1%), Diploma and sub-diploma (73%), and employed (40%). The mean Body Mass Index (BMI) was 27.68±4.91, which according to the World Health Organization, 39.6% of people were overweight. Table 1 summarizes the demographic characteristics of the participants.

The mean score of the K-6 questionnaire was 5.19±4.50 of 24. Most of the studied people (over 80%) have experienced none or minor symptoms of helplessness, hopelessness, and worthlessness. Considering the cut-off point of 10 for the k-6 questionnaire based on previous studies 34, the prevalence of PD was estimated at 19.2%.

The optimal number of classes in the LCA model was determined using the good fit criteria (Table 2). The optimal number of clusters in the K-median method was determined using

Table 1. Demographic characteristics of the subjects

		Number	Percent
Gender	Man	19287	50.7
	Woman	18771	49.3
Marital Status	Married	32767	86.1
	Widow	1803	4.7
	Divorced	848	2.2
	Single	419	1.1
Level of Education	Illiterate	4316	11.3
	Diploma and sub-diploma	24215	63.6
	University	4541	11.9
	Seminary education	77	0.2
Job Type	Unemployed	2971	7.8
	Government employee	2807	7.4
	Freelance	12442	32.7
	Other	19838	52.1
BMI	Underweight	557	1.5
	Normal weight	11399	30.0
	Overweight	15069	39.6
	Obesity	10831	28.5
Total		38058	100

Table 2. The Goodness of Fit Criteria for selecting the optimal number of classes in the LCA method

Number of Classes	LL	BIC	AIC	AIC3	CAIC	R^2	Entropy
2 classes	-230011.32	460402.34	460094.65	46013.65	460438.34	0.77	0.77
3 classes	-223028.74	446563.73	446153.48	446201.48	446611.73	0.87	0.85
4 classes	-220677.53	441987.87	441475.06	441535.06	442047.87	0.75	0.76
5 classes	-218628.63	438016.64	437401.26	437473.26	438088.64	0.73	0.76
6 classes	-217054.50	434994.94	434277.00	434361.00	435078.94	0.69	0.74

the Elbow diagram (Figure 1). In selecting the number of classes (clusters), attention was paid to the appropriate interpretability. Finally, the model with three classes (clusters) was selected by considering the statistical criteria and interpretability.

The prevalence and conditional probabilities of each of latent classes based on the indicator variables (items of questionnaire k-6) are presented in Table 3. Examining the structure of the classes formed based on the pattern of answering the items of the K6 questionnaire (Figure 2) shows that class one includes people without PD (42.0%), class two includes people with low PD (37.8%), and class three includes people with severe PD (19.3%). According to the k-median clustering method, the cluster included people without PD 46.9%, with low PD 31.1%, and with severe PD 22.0% (Table 3). The difference in clustering between the two methods is mostly observed in the Low PD classes. The mean score of the K6 in the severe PD class was 12.61±2.81 in the LCA model and 12.20±2.93 in the k-median model. More

Table 3. Prevalence and Conditional Probabilities of latent class based on a pattern of answering K6 questionnaire items

Items	Responses	Method	Without PD classes	Low PD classes	Severe PD classes	Total n (%)
Item 1 Sadness	None of the time	LCA	6183 (37.3 %)	879 (6.3 %)	83 (1.1 %)	7145 (18.8 %)
		k-median	6738 (37.7 %)	303 (2.6 %)	104 (1.2 %)	7146 (18.78%)
	A little of the time	LCA	6903 (41.7 %)	3506 (25.1 %)	291 (3.9 %)	10700 (28.1 %)
		k-median	7469 (41.8 %)	2823 (23.9 %)	408 (4.9 %)	10701 (28.12%)
	Some of the time	LCA	3211 (19.4 %)	7209 (51.5 %)	2210 (29.4 %)	12630 (33.2 %)
		k-median	3401 (19.1 %)	6539 (55.3 %)	2690 (32.1 %)	12630 (33.19%)
	Most of the time	LCA	226 (1.4 %)	2150 (15.4 %)	3861 (51.4 %)	6237 (16.4 %)
		k-median	242 (1.4 %)	1920 (16.2 %)	4075 (48.7 %)	6237 (16.39%)
	All of the time	LCA	36 (0.2 %)	241 (1.7 %)	1067 (14.2 %)	1344 (3.5 %)
		k-median	0 (0.0 %)	249 (2.1 %)	1095 (13.1 %)	1344 (3.53%)
Item 2 Restless	None of the time	LCA	12264 (74.1 %)	2071 (14.8 %)	145 (1.9 %)	14480 (38 %)
		k-median	12834 (71.9 %)	1434 (12.1 %)	212 (2.5 %)	14480 (38.05%)
	A little of the time	LCA	4105 (24.8 %)	6402 (45.8 %)	610 (8.1 %)	11117 (29.2 %)
		k-median	4739 (26.5 %)	5493 (46.4 %)	885 (10.6 %)	11118 (29.21%)
	Some of the time	LCA	190 (1.1 %)	4757 (34 %)	3104 (41.3 %)	8051 (21.2 %)
		k-median	266 (1.5 %)	4211 (35.6 %)	3574 (42.7 %)	8051 (21.16%)
	Most of the time	LCA	0 (0.0 %)	686 (4.9 %)	3049 (40.6 %)	3735 (9.8 %)
		k-median	11 (0.1 %)	638 (5.4 %)	3086 (36.9 %)	3735 (9.81%)
	All of the time	LCA	0 (0.0 %)	69 (0.5 %)	604 (8.0 %)	673 (1.8 %)
		k-median	0 (0.0 %)	58 (0.5 %)	615 (7.3 %)	673 (1.77%)
Item 3 Nervous	None of the time	LCA	13741 (83 %)	3385 (24.2 %)	137 (1.8 %)	17263 (45.4 %)
		k-median	14123 (79.1 %)	2900 (24.5 %)	240 (2.9 %)	17264 (45.36%)
	A little of the time	LCA	2588 (15.6 %)	6845 (48.9 %)	629 (8.4 %)	10062 (26.4 %)
		k-median	3371 (18.9 %)	5693 (48.1 %)	998 (11.9 %)	10063 (26.44%)
	Some of the time	LCA	230 (1.4 %)	3532 (25.3 %)	3660 (48.7 %)	7422 (19.5 %)
		k-median	341 (1.9 %)	3004 (25.4 %)	4077 (48.7 %)	7422 (19.5%)
	Most of the time	LCA	0 (0.0 %)	213 (1.5 %)	2649 (35.3 %)	2862 (7.5 %)
		k-median	15 (0.1 %)	227 (1.9 %)	2620 (31.3 %)	2862 (7.52%)
	All of the time	LCA	0 (0.0 %)	10 (0.1 %)	437 (5.8 %)	447 (1.2 %)
		k-median	0 (0.0 %)	10 (0.1 %)	437 (5.2 %)	447 (1.17%)
Item 4 Helpless	None of the time	LCA	15288 (92.3 %)	6636 (47.5 %)	925 (12.3 %)	22849 (60 %)
		k-median	16644 (93.2 %)	5129 (43.3 %)	1076 (12.9 %)	22851 (60.04%)
	A little of the time	LCA	1152 (7 %)	5472 (39.1 %)	1574 (21 %)	8198 (21.5 %)
		k-median	1114 (6.2 %)	5104 (43.1 %)	1980 (23.7 %)	8198 (21.54%)
	Some of the time	LCA	119 (0.7 %)	1689 (12.1 %)	3393 (45.2 %)	5201 (13.7 %)
		k-median	92 (0.5 %)	1440 (12.2 %)	3669 (43.8 %)	5201 (13.67%)
	Most of the time	LCA	0 (0.0 %)	172 (1.2 %)	1418 (18.9 %)	1590 (4.2 %)
		k-median	0 (0.0 %)	150 (1.3 %)	1440 (17.2 %)	1590 (4.18%)
	All of the time	LCA	0 (0.0 %)	16 (0.1 %)	202 (2.7 %)	218 (0.6 %)
		k-median	0 (0.0 %)	11 (0.1 %)	207 (2.5 %)	218 (0.57%)
Item 5 Hopeless	None of the time	LCA	16202 (97.8 %)	7930 (56.7 %)	731 (9.7 %)	24863 (65.3 %)
		k-median	16041 (89.9 %)	8115 (68.6 %)	707 (8.4 %)	24865 (65.33%)
	A little of the time	LCA	357 (2.2 %)	5016 (35.9 %)	1564 (20.8 %)	6937 (18.2 %)
		k-median	1517 (8.5 %)	3403 (28.8 %)	2017 (24.1 %)	6937 (18.23%)
	Some of the time	LCA	0 (0.0 %)	977 (7 %)	3446 (45.9 %)	4423 (11.6 %)
		k-median	273 (1.5 %)	316 (2.7 %)	3834 (45.8 %)	4423 (11.62%)
	Most of the time	LCA	0 (0.0 %)	51 (0.4 %)	1529 (20.4 %)	1580 (4.2 %)
		k-median	14 (0.1 %)	0 (0.0 %)	1566 (18.7 %)	1580 (4.15%)
	All of the time	LCA	0 (0.0 %)	11 (0.1 %)	242 (3.2 %)	253 (0.7 %)
		k-median	5 (0.0 %)	0 (0.0 %)	248 (3.0 %)	253 (0.66%)

Item 6 Worthless	None of the time	LCA	16449 (99.3 %)	10009 (71.6 %)	1783 (23.7 %)	28241 (74.2 %)
		k-median	16892 (94.6 %)	9274 (78.4 %)	2075 (24.8 %)	28242 (74.21%)
	A little of the time	LCA	110 (0.7 %)	3509 (25.1 %)	2054 (27.3 %)	5673 (14.9 %)
		k-median	826 (4.6 %)	2367 (20.0 %)	2480 (29.6 %)	5673 (14.91%)
	Some of the time	LCA	0 (0.0 %)	428 (3.1 %)	2576 (34.3 %)	3004 (7.9 %)
		k-median	120 (0.7 %)	182 (1.5 %)	2702 (32.3 %)	3004 (7.89%)
	Most of the time	LCA	0 (0.0 %)	34 (0.2 %)	924 (12.3 %)	958 (2.5 %)
		k-median	10 (0.1 %)	9 (0.1 %)	939 (11.2 %)	958 (2.52%)
	All of the time	LCA	0 (0.0 %)	5 (0.0 %)	175 (2.3 %)	180 (0.5 %)
		k-median	2 (0.0 %)	2 (0.0 %)	176 (2.1 %)	180 (0.47%)
Prevalence of classes			42.0 %	38.7 %	19.3 %	
			46.9 %	31.1 %	22.0 %	

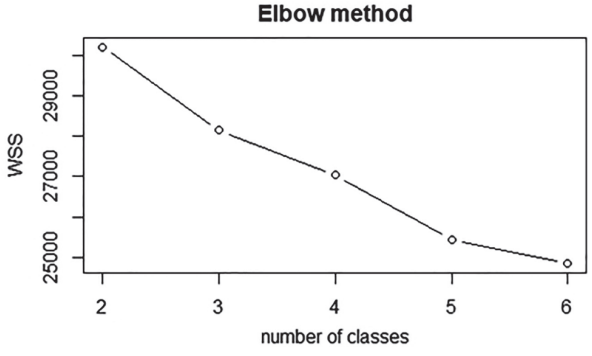

Figure 1. Elbow diagram to determine the number of optimal classes in K-median method

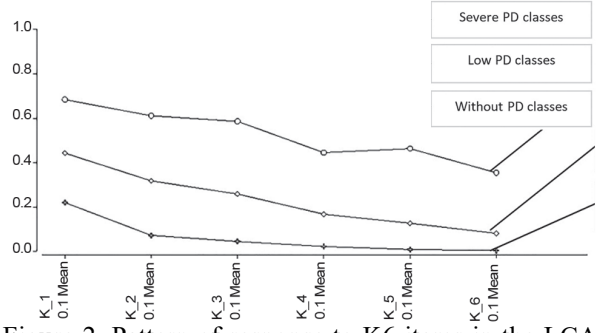

Figure 2. Pattern of response to K6 items in the LCA method

Table 4. Evaluates the agreement between LCA and k-median clustering

Method / Intensity PD	LCA		k-median	
	Count	Mean±SD	Count	Mean±SD
Without PD classes	16559	1.42±1.09	17850	1.62±1.30
Low PD classes	13985	5.67±1.66	11834	5.61±1.50
Severe PD classes	7512	12.61±2.81	8372	12.20±2.93
Total	38056	5.19±4.50	38056	5.19±4.50

Table 5. Evaluates the agreement between LCA and k-median clustering

	k-median			Total (%)
	Without PD classes	Low PD classes	Severe PD classes	
LCA				
Without PD classes	16004 (42.1%)	555 (1.5%)	0 (0.0%)	16559 (43.5%)
Low PD classes	1846 (4.9%)	11236 (29.5%)	903 (2.4%)	13985 (36.7%)
Severe PD classes	0 (0%)	43 (0.1%)	7469 (19.6%)	7512 (19.7%)
Total (%)	17850 (46.9%)	11834 (31.1%)	8372 (22.0)	38056 (100.0)

detailed information about the mean of the total score K6 by two methods is given in Table 4.
Cohen's kappa coefficient showed that the agreement between LCA and the k-median method was (κ=0.862±0.02; p-value<0.001), which is considered a high value (Table 5).

Discussion

Mental disorders are one of the most important diseases today, which has a high percentage of the burden of all diseases in the world. In recent years, PD has been known as a prevalent type of mild psychological disorder. PD is one of the potential mental health problems of the international community that needs to be identified early and treatment interventions initiated. Diagnosis of mental health problems, including PD, is made using questionnaires. However, differences in cut-off points and scoring methods have made it challenging to diagnose psychological problems and analyze questionnaires. Differences in diagnostic tools and the characteristics of the studied populations have led to different results in studies assessing the prevalence of psychological distress.
The K6 questionnaire is one of the most common tools for PD, for which the existence of different cutting points has led to varying results in studies.[11-13] Numerous statistical techniques, including clustering methods to meet the challenge of determining cutting points, have been proposed in the questionnaire. K-median and LCA are two clustering methods without the need for a questionnaire cut-off point and scoring method. Using both of the mentioned methods, three classes (clusters) with different patterns in responding to items of the k-6 questionnaire were discovered. Classes without PD, low PD, and severe PD based on the LCA were 42%, 37.8%, and 19.3%, respectively. According to the k-median method, clusters without PD, low PD, and severe PD had 46.9%, 31.1%, and 22.0% of the population. People in the class (cluster) of severe PD reported most of the symptoms, and in contrast, people in the class (cluster) without PD never experienced these symptoms. In the study of Barragan et al. using the LCA, four patterns were identified that 2.8% of the studied subjects classify in the severe PD class and 13.6% in the moderate PD class.[45]
In considering the cut-off point of 10 for the k-6 questionnaire, 19.2% of the studied people were diagnosed with PD, while both LCA and K-median methods classified more than 50% of people in the low and severe mental distress class. Estimation of the prevalence of PD using K-median and LCA methods was slightly different from each other. There was almost complete agreement between the LCA and k-median models in estimating PD. This small difference can be due to differences in the basic assumptions of the two methods.
LCA is a model-based clustering in which clusters are defined by parametric probability distributions. In LCA assumed that the whole population consists of several subpopulations or clusters in which, in each class, the variables have a different multivariate probability density function, while the entire data set is modeled with a mixture of these distributions (finite mixture density). LCA is a model-based method, so it can estimate and test parameters. Furthermore, the number of the classes can be determined based on good fit criteria and likelihood tests. Local independence in LCA means that the presence or absence of PD symptoms in a class is not related to the presence or absence of other PD symptoms. The LCA can be used for items

that include metric and non-metric variables.[39] There are generalizations to the LCA that show it can also be used to analyze longitudinal data and estimate trajectories over time.[39]

The k-median clustering method is a distance-based clustering method that classifies clusters as data subsets that have small intra-cluster distances and large inter-cluster distances from other clusters. This model attempts to find clusters that cluster similar observations without making assumptions about their distribution or attempting to fit the distribution of the mixture.[46] The k-median clustering is a non-parametric method and does not use any parametric assumption. However, due to the limitations of the k-median method based on scale dependence and random initial value, clusters must be linearly separable. There are extensions of the k-median approach suitable for asymmetric and rectangular dissimilarity matrices.[39]

Based on the results of a similar study, the LCA had a better fit than the K-mean, which is another type of distance-based clustering method.[47-51] In a study comparing three models of LCA, k-median, and k-mean, assuming the number of clusters is known, all three methods can cluster the data well. In this case, the k-median clusters the data structure better than the k-mean. When the number of clusters was 2 or 3, the LCA performed better than the mean k-median. When the number of clusters was four or more, the k-median performed better than LCA.[39] Also, when the number of clusters is unknown, AIC3 can determine the number of clusters in the LCA better than other goodness fit criteria.[39] Previous studies have shown that model-based approaches such as LCA for clustering perform well and are sometimes better than distance-based approaches. When the number of variables is large, or the number of classes is low, or the sample size is large or the prevalence of classes is non-uniform, The LCA can work well.[52] The k-median model performs better than the LCA when applied to dichotomous data.[39] From the point of view of fitting the LCA and k-median methods in statistical software, there is an LCA approach in more software than k-median, including latent variable model software such as Mplus and Latent Gold.

In this study, some of the limitations of previous studies were removed, but this study also had some limitations. In this study, people who were voluntarily referred to community health care centers were examined; some people with psychological disorders may not go to these centers; this underestimates the prevalence for the general public in this study. In Iranian community healthcare centers, healthcare providers complete the electronic health record system of individuals. There is a possibility that individuals may not be honest in answering the items of the K6. If individuals had completed the questionnaire themselves, they might have been more honest in answering the K6.

Conclusion

Most of the people were classified in low or no PD clusters. About 20% of the people were classified in the Severe PD cluster, which indicates the low prevalence of severe PD among the clients of Mashhad community health care centers. Since PD can reliably predict severe mental illnesses, it is necessary to pay attention to PD in different societies. Community health care centers can prevent serious mental disorders by using appropriate health care and services and are well-planned

to provide mental health services, especially for high-risk groups of PD.

In this study, we tried to increase the validity and accuracy of the results by using clustering methods. Most studies in this field are limited to descriptive methods of data, and differences in scoring and cutting points have caused significant differences in the results of investigations. Our findings showed that the results of fitting the two models of LCA and k-median are in good agreement, and the performance of these two models in data clustering is appropriate. Considering the mentioned advantages of LCA over K-median and the traditional scoring method, we recommend LCA for the classification of PD using the k-6 questionnaire.

Conflicts of interest

The authors declared no conflicts of interest.

Abbreviations

PD, Psychiatric Distress; K-6, Six Item Kessler Psychological Distress Scale; LCR, Latent class regression

Acknowledgments

We thank Dr. Ehsan Mussa Farkhani for providing the data. This study was financially supported by Mashhad University of Medical Sciences as a part of a MSc thesis (Second Author).

References

1. Alizade Z, Rejali M, Feizi A, Afshar H, Hassanzade Kashtali A, Adibi P. Investigation of psychological disorders profile (anxiety, depression and psychological distress) in adult population of Isfahan province. J Torbat Heydariyeh Uni Med Sci. 2016;3(4):42-8. doi.

2. Rehm J, Shield KD. Global Burden of Disease and the Impact of Mental and Addictive Disorders. Curr Psychiatry Rep. 2019;21(2):10. doi: 10.1007/s11920-019-0997-0.

3. Noorbala AA, Bagheri Yazdi SA, Yasamy MT, Mohammad K. Mental health survey of the adult population in Iran. Br J Psychiatry. 2004;184:70-3. doi: 10.1192/bjp.184.1.70.

4. Pouretemad HR, Naghavi HR, Malekafzali H, Noorbala AA, Davidian H, Ghanizadeh A, et al. Prevalence of Mood Disorders in Iran. Iranian Journal of Psychiatry. 2006;1(2). doi.

5. Rahimi Z, Esmaily H, Taghipour A, Mosa Farkhani E, Jamali J. Evaluation of the Prevalence of Psychological Distress and Its Related Demographic Factors in 18-65 Year-Old Population of Khorasan Razavi: A Cross-Sectional Study. Iran J Epidemiol. 2021;16(4):316-24. doi.

6. Emami H, Ghazinour M, Rezaeishiraz H, Richter J. Mental health of adolescents in Tehran, Iran. J Adolesc Health. 2007;41(6):571-6. doi: 10.1016/j.jadohealth.2007.06.005.

7. Jafari N, Loghmani A, Montazeri A. Mental health of Medical Students in Different Levels of Training. Int J Prev Med. 2012;3(Suppl 1):S107-12. doi.

8. Wang C, Tee M, Roy AE, Fardin MA, Srichokchatchawan W, Habib HA, et al. The impact of COVID-19 pandemic on physical and mental health of Asians: A study of seven middle-income countries in Asia. PLoS One. 2021;16(2):e0246824. doi: 10.1371/journal.pone.0246824.

9. Javed B, Sarwer A, Soto EB, Mashwani Z-U-R. The coronavirus (COVID-19) pandemic's impact on mental health. The International journal of health planning and management. 2020;35(5):993-6. doi: 10.1002/hpm.3008.

10. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. Gen Psychiatr. 2020;33(2):e100213. doi: 10.1136/gpsych-2020-100213.

11. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet. 2020;395(10227):912-20. doi: 10.1016/s0140-6736(20)30460-8.

12. Wu T, Jia X, Shi H, Niu J, Yin X, Xie J, et al. Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis. J Affect Disord. 2021;281:91-8. doi: 10.1016/j.jad.2020.11.117.

13. Feigin VL, Roth GA, Naghavi M, Parmar P, Krishnamurthi R, Chugh S, et al. Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. 2016;15(9):913-24. doi.

14. Necho M, Tsehay M, Birkie M, Biset G, Tadesse E. Prevalence of anxiety, depression, and psychological distress among the general population during the COVID-19 pandemic: A systematic review and meta-analysis. Int J Soc Psychiatry. 2021;67(7):892-906. doi: 10.1177/00207640211003121.

15. Firdaus G. Increasing Rate of Psychological Distress in Urban Households: How Does Income Matter? Community Ment Health J. 2018;54(5):641-8. doi: 10.1007/s10597-017-0193-9.

16. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med. 2002;32(6):959-76. doi.

17. Murugan N, Mishra AK, Chauhan RC, Velavan A. Psychological distress among adult urban population of Puducherry. Int J Community Med Public Health. 2018;5(8):3265-9. doi: 10.18203/2394-6040.ijcmph20182984.

18. Vaziri S, Shaydayi Aghdam S, No0bakht L, Khalili M, Vaziri Y, Masumi R, et al. Comparison of physical symptoms in people with low and high psychological distress Patients Disease. Thoughts and Behavior in Clinical Psychology. 2018;13(49):57-66. doi.

19. Enticott JC, Lin E, Shawyer F, Russell G, Inder B, Patten S, et al.

Prevalence of psychological distress: How do Australia and Canada compare? Aust N Z J Psychiatry. 2018;52(3):227-38. doi: 10.1177/0004867417708612.

20. Weissman JF, Pratt LA, Miller EA, Parker JD. Serious Psychological Distress Among Adults: United States, 2009-2013. NCHS Data Brief. 2015(203):1-8. doi.

21. Tanji F, Tomata Y, Zhang S, Otsuka T, Tsuji I. Psychological distress and completed suicide in Japan: A comparison of the impact of moderate and severe psychological distress. Prev Med. 2018;116:99-103. doi: 10.1016/j.ypmed.2018.09.007.

22. Farrag NS, El-Gilany AH, Abdelsalam SA. Prevalence and predictors of psychological distress among primary healthcare service users in Mansoura district, Egypt. Health Soc Care Community. 2019;27(6):1451-7. doi: 10.1111/hsc.12816.

23. Takeda T, Yoshimi K, Kai S, Inoue F. Association Between Serious Psychological Distress and Loneliness During the COVID-19 Pandemic: A Cross-Sectional Study with Pregnant Japanese Women. Int J Womens Health. 2021;13:1087-93. doi: 10.2147/ijwh.S338596.

24. Tippens JA, Hatton-Bowers H, Honomichl R, Wheeler LA, Miamidian HM, Bash KL, et al. Psychological distress prevalence and associated stressors and supports among urban-displaced Congolese adults in Kenya. Am J Orthopsychiatry. 2021;91(5):626-34. doi: 10.1037/ort0000564.

25. Nguyen LH, Vu HM, Vu GT, Tran TH, Pham KTH, Nguyen BT, et al. Prevalence and Factors Associated with Psychological Distress among Older Adults Admitted to Hospitals After Fall Injuries in Vietnam. Int J Environ Res Public Health. 2019;16(22). doi: 10.3390/ijerph16224518.

26. McGinty EE, Presskreischer R, Han H, Barry CL. Psychological Distress and Loneliness Reported by US Adults in 2018 and April 2020. Jama. 2020;324(1):93-4. doi: 10.1001/jama.2020.9740.

27. Daly M, Robinson E. Psychological distress and adaptation to the COVID-19 crisis in the United States. J Psychiatr Res. 2021;136:603-9. doi: 10.1016/j.jpsychires.2020.10.035.

28. Zhang M, Zhang J, Zhang F, Zhang L, Feng D. Prevalence of psychological distress and the effects of resilience and perceived social support among Chinese college students: Does gender make a difference? Psychiatry Res. 2018;267:409-13. doi: 10.1016/j.psychres.2018.06.038.

29. Mazza C, Ricci E, Biondi S, Colasanti M, Ferracuti S, Napoli C, et al. A Nationwide Survey of Psychological Distress among Italian People during the COVID-19 Pandemic: Immediate Psychological Responses and Associated Factors. Int J Environ Res Public Health. 2020;17(9). doi: 10.3390/ijerph17093165.

30. Cénat JM, Blais-Rochette C, Kokou-Kpolou CK, Noorishad PG, Mukunzi JN, McIntee SE, et al. Prevalence of symptoms of

depression, anxiety, insomnia, posttraumatic stress disorder, and psychological distress among populations affected by the COVID-19 pandemic: A systematic review and meta-analysis. Psychiatry Res. 2021;295:113599. doi: 10.1016/j.psychres.2020.113599.

31. Monfared A, Akhondzadeh L, Soleimani R, Maroufizadeh S, Pouy S, Asgari F. Psychological Distress and Coping Strategies Among Clinicians and Medical Students During the COVID-19 Pandemic: A Cross-sectional Study in Guilan, Iran. Shiraz E-Med J. 2021;22(5):e109764. doi: 10.5812/semj.109764.

32. Tran TT, Adams-Bedford J, Yiengprugsawan V, Seubsman SA, Sleigh A. Psychological Distress following Injury in a Large Cohort of Thai Adults. PLoS One. 2016;11(10):e0164767. doi: 10.1371/journal.pone.0164767.

33. Nahidi S, Blignault I, Hayen A, Razee H. Psychological Distress in Iranian International Students at an Australian University. J Immigr Minor Health. 2018;20(3):651-7. doi: 10.1007/s10903-017-0590-8.

34. Hajebi A, Motevalian A, Amin-Esmaeili M, Rahimi-Movaghar A, Sharifi V, Hoseini L, et al. Adaptation and validation of short scales for assessment of psychological distress in Iran: The Persian K10 and K6. Int J Methods Psychiatr Res. 2018;27(3):e1726. doi: 10.1002/mpr.1726.

35. Rai P, Singh S. A survey of clustering techniques. Int J Comput Appl. 2010;7(12):1-5. doi.

36. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol. 2001;63(2):411-23. doi: 10.1111/1467-9868.00293.

37. Magidson J, Vermunt JK. Latent class models for clustering: A comparison with K-means. Canadian Journal of Marketing Research. 2002;20(1):36-43. doi.

38. Jamali J, Ayatollahi SM, Jafari P. A New Measurement Equivalence Technique Based on Latent Class Regression as Compared with Multiple Indicators Multiple Causes. Acta Inform Med. 2016;24(3):168-71. doi: 10.5455/aim.2016.24.168-171.

39. Brusco MJ, Shireman E, Steinley D. A comparison of latent class, K-means, and K-median methods for clustering dichotomous data. Psychol Methods. 2017;22(3):563-80. doi: 10.1037/met0000095.

40. Everitt B. An R and S-PLUS companion to multivariate analysis. London: Springer Science & Business Media; 2005.

41. Tein JY, Coxe S, Cham H. Statistical Power to Detect the Correct Number of Classes in Latent Profile Analysis. Struct Equ Modeling. 2013;20(4):640-57. doi: 10.1080/10705511.2013.824781.

42. Jamali J, Roustaei N, Ayatollahi SMT, Sadeghi E. Factors affecting minor psychiatric disorders in Southern Iranian nurses: A latent class regression analysis. Nurs Midwifery Stud. 2015 4(2):e28017. doi: 10.17795/

nmsjournal28017.

43. Vergara VM, Salman M, Abrol A, Espinoza FA, Calhoun VD. Determining the number of states in dynamic functional connectivity using cluster validity indexes. J Neurosci Methods. 2020;337:108651. doi: 10.1016/j.jneumeth.2020.108651.

44. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. doi.

45. Barragán A, Yamada AM, Gilreath TD. Psychological Distress Behavioral Patterns Among Latinos: We Don't See Ourselves as Worthless. Community Ment Health J. 2019;55(2):232-40. doi: 10.1007/s10597-018-0273-5.

46. Papachristou N, Barnaghi P, Cooper BA, Hu X, Maguire R, Apostolidis K, et al. Congruence Between Latent Class and K-Modes Analyses in the Identification of Oncology Patients With Distinct Symptom Experiences. J Pain Symptom Manage. 2018;55(2):318-33.e4. doi: 10.1016/j.jpainsymman.2017.08.020.

47. Jamali J, Ayatollahi SMT. Classification of Iranian Nurses According to their Mental Health Outcomes Using GHQ-12 Questionnaire: a Comparison Between Latent Class Analysis and K-means Clustering with Traditional Scoring Method. Mater Sociomed. 2015;27(5):337-41. doi: 10.5455/msm.2015.27.337-341.

48. Magidson J, Vermunt JJCjomr. Latent class models for clustering: A comparison with K-means. Canadian Journal of Marketing Research. 2002;20(1):36-43. doi.

49. Fahey MT, Thane CW, Bramwell GD, Coward WA. Conditional Gaussian mixture modelling for dietary pattern analysis. J R Statist Soc. 2007;170(1):149-66. doi: https://doi.org/10.1111/j.1467-985X.2006.00452.x.

50. Flann N, Xu B, Recker M, Qi X, Ye LJR. Clustering Educational Digital Library Usage Data: A Comparison of Latent Class Analysis and K-means Algorithms. 2013;5(2). doi.

51. Wagstaff K, Cardie C, Rogers S, Schrödl S, editors. Constrained k-means clustering with background knowledge. Icml; 2001.

52. Anderlucci L, Hennig C. The Clustering of Categorical Data: A Comparison of a Model-based and a Distance-based Approach. Communications in Statistics - Theory and Methods. 2014;43(4):704-21. doi: 10.1080/03610926.2013.806665.