CME Article

Biostatistics 304. Cluster analysis

Y H Chan

In Cluster analysis, we seek to identify the "natural" structure of groups based on a multivariate profile, if it exists, which both minimises the within-group variation and maximises the between-group variation. The objective is to perform data reduction into manageable bite-sizes which could be used in further analysis or developing hypothesis concerning the nature of the data. It is exploratory, descriptive and non-inferential.

This technique will always create clusters, be it right or wrong. The solutions are not unique since they are dependent on the variables used and how cluster membership is being defined. There are no essential assumptions required for its use except that there must be some regard to theoretical/conceptual rationale upon which the variables are selected.

For simplicity, we shall use 10 subjects to demonstrate how cluster analysis works. We are interested to group these 10 subjects into compliance-on-medication-taking (for example) subgroups basing on four biomarkers, and later to use the clusters to do further analysis – say, to profile compliant vs non-compliant subjects. The descriptives are given in Table I, with higher values being indicative of better compliance.

Table I. Descriptive statistics of the biomarkers.

	Descriptive statistics				
	N	Minimum	Maximum	Mean	Standard deviation
x1	10	79.2	87.3	83.870	2.6961
x2	10	73.2	83.3	77.850	3.6567
x3	10	61.8	81.1	72.080	6.4173
x4	10	44.5	51.5	48.950	2.5761
Valid N (listwise)					

Faculty of Medicine
National University of Singapore
Block MD11
Clinical Research Centre #02-02
10 Medical Drive
Singapore 117597

Y H Chan, PhD
Head
Biostatistics Unit

Correspondence to:
Dr Y H Chan
Tel: (65) 6874 3698
Fax: (65) 6778 5743
Email: medcyh@nus.edu.sg

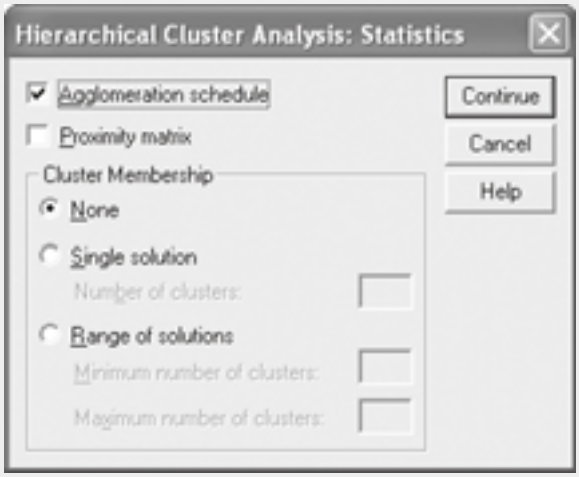
SPSS offers three separate approaches to Cluster analysis, namely: TwoStep, K-Means and Hierarchical. We shall discuss the Hierarchical approach first. This is chosen when we have little idea of the data structure. There are two basic hierarchical clustering procedures – agglomerative or divisive. Agglomerative starts with each object as a cluster and new clusters are combined until eventually all individuals are grouped into one large cluster. Divisive proceeds in the opposite direction to agglomerative methods. For n cases, there will be one-cluster to n-1 cluster solutions.

In SPSS, go to Analyse, Classify, Hierarchical Cluster to get Template I

Template I. Hierarchical cluster analysis.

Put the four biomarkers into the Variable(s) option. If there is a string variable which labels the cases (here "subno' contains the labels A to J), put "subno" in the Label Cases by option - otherwise, leave it empty. Presently, as we want to cluster the Cases, leave the bullet for Cases checked. Leave the Display for the Statistics and Plots checked. Click on the Statistics folder to get Template II.

Template II. Statistics folder.



Leave the Agglomeration schedule checked. The Proximity matrix gives the distances or similarities between items (this could be very messy if n is large) – leave it unchecked.

Clicked on the Plots folder in Template I to get Template III

Template III. Plots folder.

Check the Dendrogram box. For Icicle - check none (as we do not need this plot)

Click on the Method folder in Template I. In Template IV, we need to address two basic questions in forming clusters.

1. How to measure Similarity between objects? Since all the four biomarkers are quantitative variables, use the Interval Measure option. Choose Squared Euclidean distance which gives the straight line distance between two objects – click on the Help button to see the definitions of the other Interval Measure options. The other two options for data-type are Counts (study with Likert scales) and Binary (study with yes/no scales).

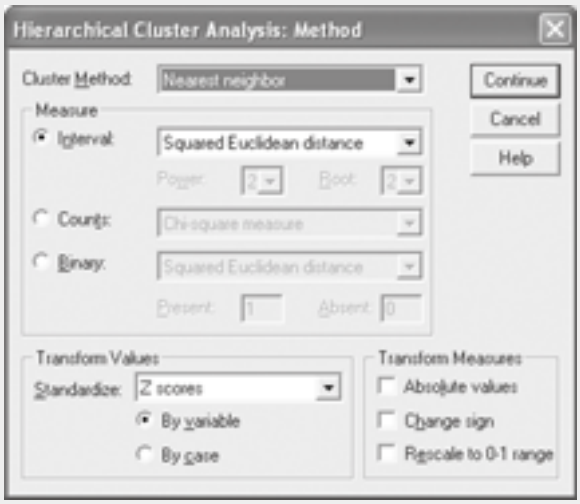
   From Table I, the means (sd) of the four variables are quite different. Variables with a bigger variation have more impact on the distance measure. Thus, it may be necessary to standardise the variables (use the Z-scores, By Variable option). This will eliminate the effects due to scale differences. On the other hand, if one believes that there is a "natural" pattern being reflected in the present-scales of the variables, then standardisation may not be useful.

   Standardising by case allows us to remove response-style effects from respondents. We do not want the clusters to just reflect that there are groups where one feels everything is "okay", another feels everything "sucks" and the last, a "so-so" group. Standardising by case allows us to see the relative importance of one variable to another – by standardising each question to the each respondent's mean score, for example attitudinal studies.

2. How are the clusters being formed? In the Cluster Method, choose the Nearest Neighbor option. This technique is also known as Single Linkage which uses the minimum distance between two objects to do the clustering and has the potential disadvantage of forming long snake-like chains.

   The Furthest Neighbour (also known as Complete linkage) option, that uses the maximum distance between two objects, may help to eliminate the snaking problem. The Between-groups and Within-groups are Average linkage methods which use the average distance of all individuals in one cluster to another. These are not affected by extreme values as do single/complete linkage and tend to combine clusters with approximately the same variance. The Centroid and Median methods are least affected by outliers. In the Ward's method, there is a bias towards forming clusters of equal sizes.

Template IV. Method folder.



The output of the above analysis will only have one table (Table II) and one figure (Fig. 1)

Table II shows the Agglomeration schedule, using Squared Euclidean distance (standardised) measure and Nearest-Neighbor (Single linkage) cluster. This displays the cases or clusters combined at each stage, the distances between the cases or clusters being combined, and the last cluster level at which a case joined the cluster.

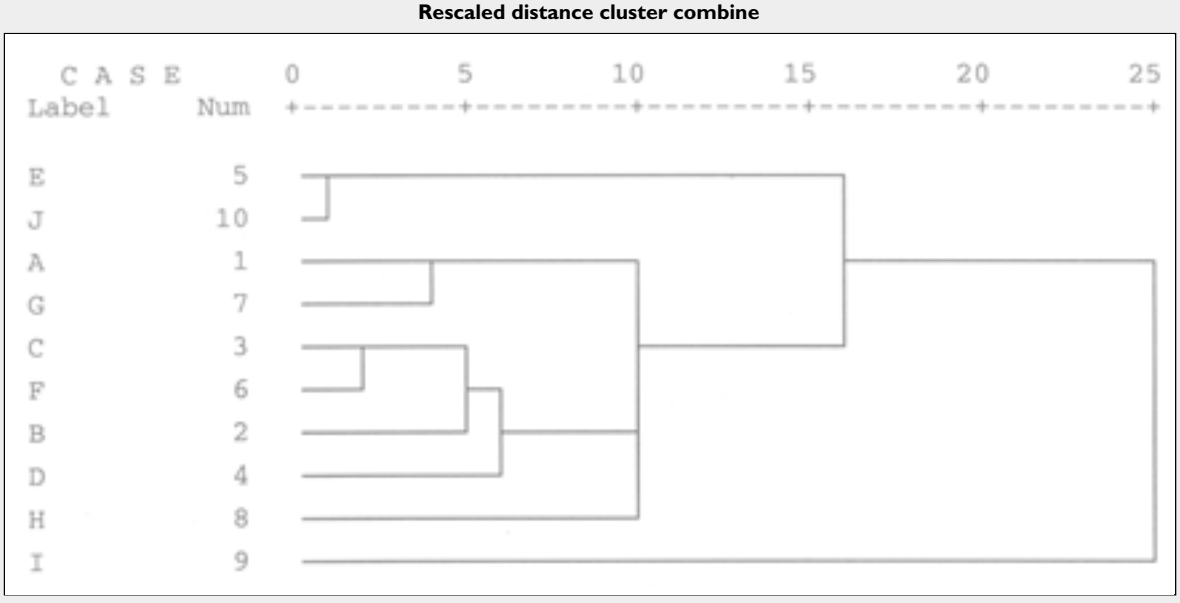
Under the Cluster Combined columns, the first two subjects to be clustered are 5 & 10, then 3 & 6, then 1 & 7, then (3 & 6) with subject 2, etc. The Coefficients column shows the distance where the clusters were being formed. The information given in the Stage Clusters First Appears columns just indicates when an object is joining an existing cluster or when two existing clusters are being combined. This table shows the numerical illustration of the clustering.

The dendrogram (Fig. 1) is the graphical equivalent of the Agglomeration schedule.

Table II. Agglomeration schedule, nearest neighbor (single linkage) and squared euclidean distance (standardised).

Agglomeration schedule						
	Cluster combined			Stage cluster first appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next stage
1	5	10	.689	0	0	8
2	3	6	1.046	0	0	4
3	1	7	1.724	0	0	7
4	2	3	1.920	0	2	5
5	2	4	2.150	4	0	6
6	2	8	3.338	5	0	7
7	1	2	3.376	3	6	8
8	1	5	4.818	7	1	9
9	1	9	7.451	8	0	0

Fig. 1 Dendrogram using single linkage.



A picture tells a thousand words! The dendrogram shows that we have an outlier (subject 9, label I) – known as a Runt or Entropy group, which is completely on its own.

The third and final question in Cluster Analysis is: how many clusters are to be formed?

A possible three cluster solution is subject 9 (label I) with two clusters, [B,C,F,D,A,G,H] and [E,J]. Another possibility is four clusters with [I], [A,G] , [E,J] and [B,C,F,D,H]. We can save the cluster membership by clicking on the Save folder in Template I – see Template V. If we decide to have four clusters after studying the dendrogram, choose the Single solution and Number of clusters = 4. SPSS will create a new variable CLU4_1 in the dataset. Perhaps, we want to do further exploration with different cluster solutions; we could use the Range of solutions option: Minimum = 2 and Maximum = 4, for example. SPSS will create three new variables CLU4_1, CLU3_1 and CLU2_1 for the 4, 3 & 2 cluster memberships, respectively.

Are we done? Things are not that simple (that's life!) What happens if we decide to use another procedure to form the clusters, say, Furthest Neighbour (also known as Complete Linkage) and stick to the Squared Euclidean distance measure? We will only show the dendrogram solution (Fig. 2)

A completely different set of solutions!

It is probably clear by now that the selection of the final cluster solution is dependent on the researcher's judgment, given that the relevant variables have been used! One could try the other possibilities of similarity measures and cluster-linkage options to explore the data.

Cluster analysis could also be performed on the variables instead of objects. In Template I, choose the Cluster Variables option. Fig. 3 shows the dendrogram with Standardised Eluclidean distance and Nearest Neighbour for the clustering of the variables.

Template V. Saving the cluster membership.

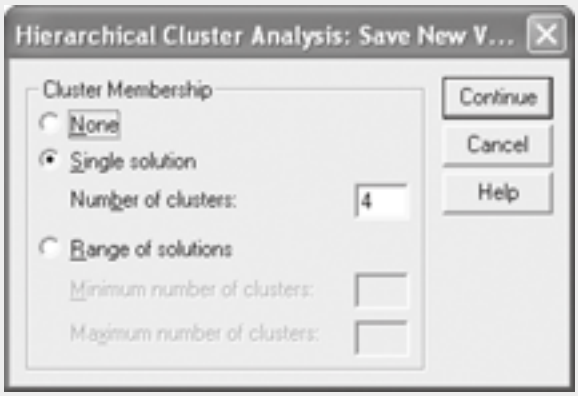


Fig. 2 Dendrogram using complete linkage.

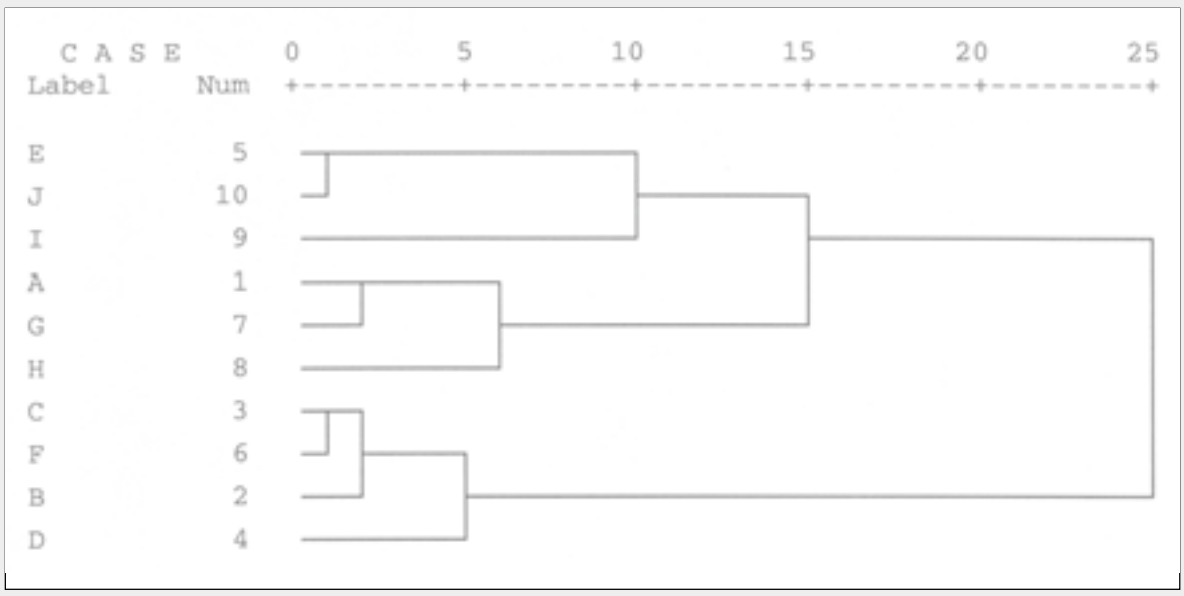
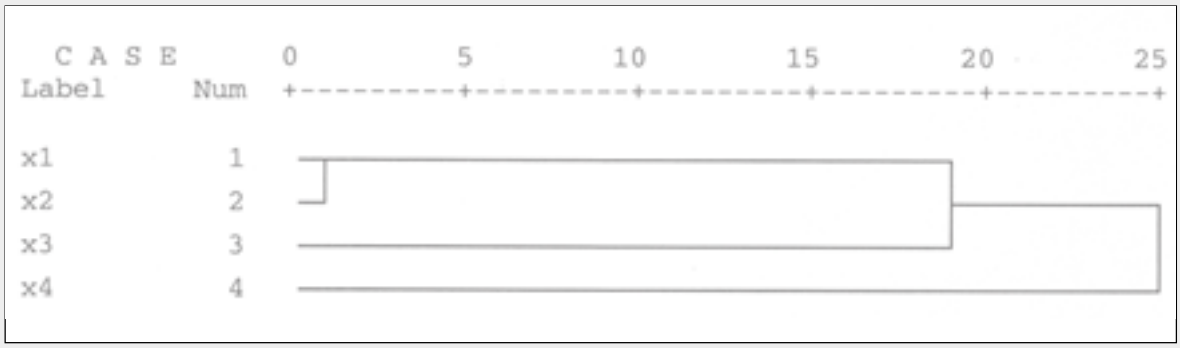


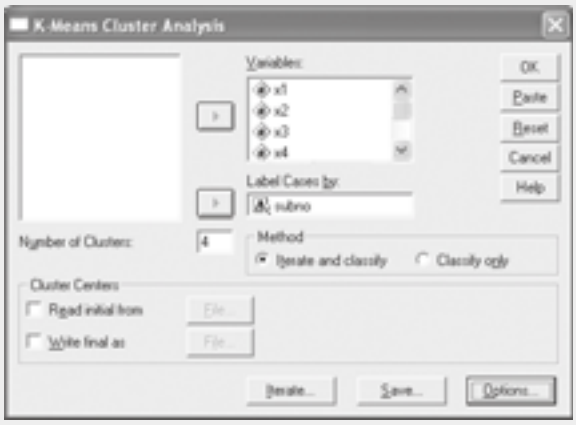
Fig. 3 Dendrogram for cluster of variables – single linkage, standardised squared euclidean distance.



The K-Means (also known as Quick Cluster) analysis could be used if we know the number of clusters to be obtained. This technique is non-hierarchical which does not involve the dendrogram-type of construction. Each cluster has an initial centre and objects within a pre-specified distance are included in the resulting cluster. Clusters' centres are updated, objects may be reassigned, and the process continues until all objects are duly classified to a cluster.

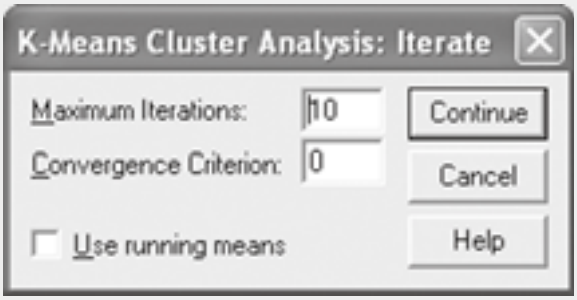
Template VI shows the options for a K-Means clustering. Number of clusters = 4 (say), choose Iterate and classify Method to allow for the objects to be "reclassified" during the clustering process. Leave the "Cluster Centres: Read initial from" unchecked – this will let the program to choose its own random cluster initial centre. Different results could be obtained when different cluster initial centres are being used! Note that K-Means do not standardise the variables for us. We will have to do it on our own using Analyze, Descriptive Statistics, Descriptive – save standardised values as variables option.

Template VI. K-Means cluster analysis.



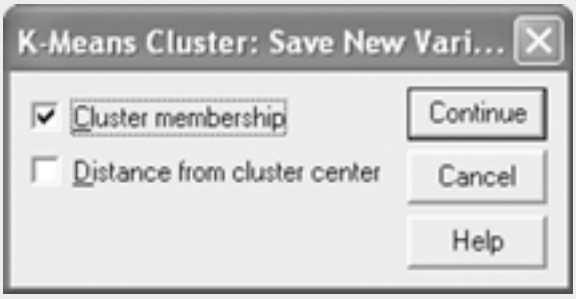
Click on Iterate folder to specify the number of iterations required (Template VII).

Template VII. Maximum number of iterations declared.



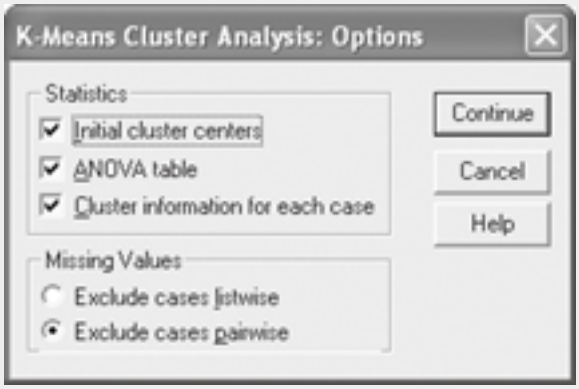
Click on the Save folder for Cluster membership (Template VIII).

Template VIII. K-Means: saving cluster membership.



Click on the Option folder in Template VI. Check all the boxes for Statistics, and Exclude cases pairwise for Missing Values (this will make use of all available non-missing data), see Template IX.

Template IX. K-Means options.



Tables IIIa-g show the outputs in a K-Means analysis. Table IIIa shows the starting cluster initial centres with Table IIIb showing that the iteration completed at the second run (where all the numbers are small). When we have more cases to be clustered, the iteration process may take longer and we have to change the maximum of number of iterations in Template VII to a higher number. We could also save the last unconverged cluster centres in a file to be served as the initial cluster centres for the next run of the process - this is done by checking the "Cluster centres: Write final as" button in Template VI.

The following results (Tables IIIc-g) could only be used when the iteration process converges. Table IIIc shows the cluster membership with Table IIId specifying the final cluster centres. Table IIIe shows the Squared Euclidean distances (the only option available in K-Means) between the clusters. Table IIIf shows which variables are significantly different amongst the clusters and Table IIIg gives the number of objects in each cluster.

Table IIIa. Initial cluster centres.

Initial cluster centres

	Cluster			
	1	2	3	4
x1	82.5	79.2	86.9	86.1
x2	76.1	73.2	80.3	83.3
x3	61.8	72.3	71.5	81.1
x4	51.0	44.5	49.0	51.0

Table IIIb. Iteration history.

Iteration history

	Change in cluster centres			
Iteration	1	2	3	4
1	3.294	3.903	2.032	4.055
2	.000	.000	.000	.000

Table IIIc. K-Means: cluster membership.

Cluster membership

Case number	Subno	Cluster	Distance
1	A	1	3.323
2	B	3	2.032
3	C	3	2.032
4	D	4	4.055
5	E	4	6.005
6	F	4	4.127
7	G	1	3.294
8	H	1	3.976
9	I	2	3.903
10	J	2	3.903

Table IIId. Final cluster centres.

Final cluster centres

	Cluster			
	1	2	3	4
x1	84.2	80.2	85.5	85.0
x2	76.5	73.4	80.9	80.2
x3	63.9	73.7	72.4	79.0
x4	49.2	48.0	48.0	50.0

Table IIIe. Distances between final cluster centres.

Distances between final cluster centres

	Cluster			
Cluster	1	2	3	4
x1		11.064	9.632	15.525
x2	11.064		9.254	10.069
x3	9.632	9.254		6.960
x4	15.525	10.069	6.960	

Table IIIf. ANOVA table for each variable by clusters.

ANOVA

	Cluster		Error			
	Mean square	df	Mean square	df	F	Sig.
x1	11.934	3	4.936	6	2.418	.165
x2	26.845	3	6.635	6	4.046	.069
x3	115.593	3	3.976	6	29.070	.001
x4	2.353	3	8.778	6	.268	.846

Table IIIg. Number of cases in each cluster.

Number of cases in each cluster		
Cluster	1	3.000
	2	2.000
	3	2.000
	4	3.000
Valid		10.000
Missing		.000

The Two-Step Cluster (see Template X) analysis allows us to have a combination of continuous and categorical variables which both hierarchical and K-means procedures do not cater for. It also allows us to specify the number of clusters required or to let the program to decide the optimal number of clusters.

When all the variables are continuous, the Euclidean Distance Measure is used. These variables will be standardised during the analysis. When a combination of continuous and categorical variables are used, the Log-likelihood distance measure have to be used. This likelihood distance measure assumes that variables in the cluster model are independent with all continuous variables assumed to have a normal distribution and all categorical variables to have a multinomial distribution. Fortunately the Two-Step procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions.

We will not generate any output results for this procedure. Those who are interested could click on the Help button in Template X to see a complete illustration of cluster analysis using the Two-Step procedure.

Template X. Two-step cluster analysis.

In conclusion, we have to bear in mind that Cluster analysis is an exploratory technique where we hope to find distinct groups based on a multivariate profile. It is an art rather than a science. However, it can be an invaluable tool to identify latent patterns in a huge dataset that could not be discerned by any other multivariate statistical method.

SINGAPORE MEDICAL COUNCIL CATEGORY 3B CME PROGRAMME

Multiple Choice Questions (Code SMJ 200504A)

	True	False
Question 1. Which cluster-linkage method in the hierarchical technique has a potential to produce snakelike clusters?		
(a) The Single linkage.	❑	❑
(b) The Complete linkage.	❑	❑
(c) The Ward's linkage.	❑	❑
(d) The Centroid linkage.	❑	❑
Question 2. Which cluster-linkage method in the hierarchical technique is not affected by outliers?		
(a) The Single linkage.	❑	❑
(b) The Complete linkage.	❑	❑
(c) The Average linkage.	❑	❑
(d) The Centroid linkage.	❑	❑
Question 3. Which technique provides a dendrogram?		
(a) The Hierarchical technique.	❑	❑
(b) The K-means technique.	❑	❑
(c) The Two-Step technique.	❑	❑
(d) All of the above.	❑	❑
Question 4. Which technique could determine the optimal number of clusters for us automatically?		
(a) The Hierarchical technique.	❑	❑
(b) The K-means technique.	❑	❑
(c) The Two-Step technique.	❑	❑
(d) All of the above.	❑	❑
Question 5. Which of the following statements are true?		
(a) K-Means is a non-hierarchical technique.	❑	❑
(b) Results from Cluster analysis are unique.	❑	❑
(c) Continuous variables must be standardised before clustering.	❑	❑
(d) Clusters will always be created.	❑	❑

Doctor's particulars:

Name in full: _____

MCR number: _____ Specialty: _____

Email address: _____

Submission instructions:

A. Using this answer form

1. Photocopy this answer form.
2. Indicate your responses by marking the "True" or "False" box ☑
3. Fill in your professional particulars.
4. Either post the answer form to the SMJ at 2 College Road, Singapore 169850 OR fax to SMJ at (65) 6224 7827.

B. Electronic submission

1. Log on at the SMJ website: URL http://www.sma.org.sg/cme/smj
2. Either download the answer form and submit to smj.cme@sma.org.sg OR download and print out the answer form for this article and follow steps A. 2-4 (above) OR complete and submit the answer form online.

Deadline for submission: (April 2005 SMJ 3B CME programme): 12 noon, 25 May 2005

Results:

1. Answers will be published in the SMJ June 2005 issue.
2. The MCR numbers of successful candidates will be posted online at http://www.sma.org.sg/cme/smj by 20 June 2005.
3. Passing mark is 60%. No mark will be deducted for incorrect answers.
4. The SMJ editorial office will submit the list of successful candidates to the Singapore Medical Council.