American Journal of Epidemiology

Vol. 177, No. 7
DOI: 10.1093/aje/kws289
Advance Access publication:
February 27, 2013

Practice of Epidemiology

A Framework for Multiple Imputation in Cluster Analysis

Xavier Basagaña*, Jose Barrera-Gómez, Marta Benet, Josep M. Antó, and Judith Garcia-Aymerich

* Correspondence to Dr. Xavier Basagaña, Centre for Research in Environmental Epidemiology, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain (e-mail xbasagana@creal.cat).

Initially submitted November 23, 2011; accepted for publication June 7, 2012.

Multiple imputation is a common technique for dealing with missing values and is mostly applied in regression settings. Its application in cluster analysis problems, where the main objective is to classify individuals into homogenous groups, involves several difficulties which are not well characterized in the current literature. In this paper, we propose a framework for applying multiple imputation to cluster analysis when the original data contain missing values. The proposed framework incorporates the selection of the final number of clusters and a variable reduction procedure, which may be needed in data sets where the ratio of the number of persons to the number of variables is small. We suggest some ways to report how the uncertainty due to multiple imputation of missing data affects the cluster analysis outcomes—namely the final number of clusters, the results of a variable selection procedure (if applied), and the assignment of individuals to clusters. The proposed framework is illustrated with data from the Phenotype and Course of Chronic Obstructive Pulmonary Disease (PAC-COPD) Study (Spain, 2004–2008), which aimed to classify patients with chronic obstructive pulmonary disease into different disease subtypes.

classification; cluster analysis; imputation; missing data

Abbreviations: COPD, chronic obstructive pulmonary disease; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease.

Missing data is an important and common problem in epidemiologic studies. The most common strategy for dealing with missing data is still a *complete case analysis*, where participants with missing data on any variable are excluded from the analyses (1). Complete case analysis can introduce selection bias that may lead to incorrect inferences, because the analysis sample no longer retains the properties of the original population (2). In addition, this type of analysis leads to a loss of information and therefore of efficiency, which comes from 2 sources: 1) not having observed the value of a particular variable in a given subject (i.e., truly missing values) and 2) dropping observed values on some variables for a subject with missing values on other variables. The latter issue is of special concern when dealing with many variables, as is often the case in cluster analysis, where the aim is to categorize participants into homogeneous groups based on several characteristics. For example, in a study with 50 variables, if each variable had 5% of its values missing independently of the other variables, a complete case analysis would include only 8% of the subjects in the original sample.

Several statistical techniques exist for the proper treatment of missing data under certain assumptions. Multiple imputation is one such technique and, although it is still not widely used (1), it is becoming increasingly popular, in part because most statistical packages provide implementations that are relatively easy and because of the appearance of several papers addressing practical questions on how to implement this technique (3–9). Multiple imputation is especially well-suited for situations where the main interest is in a population parameter, such as a regression coefficient. However, its use in cluster analysis, where the main aim is not to estimate a population parameter but to predict

cluster membership for each participant, entails several difficulties which are not well characterized in the current literature.

Our aim in this paper is to provide a framework for data analysts on how to integrate multiple imputation of missing values into cluster analysis. Cluster analysis involves many analytical decisions, such as deciding what algorithm and metric to use, how to choose the number of clusters, or whether dimensionality reduction is needed and, if so, how it should be performed. For illustration purposes, we will make a choice on every one of those items, but readers should be cautioned that by no means do we claim that they are the optimal ones.

OVERVIEW OF MULTIPLE IMPUTATION

Multiple imputation involves imputing every missing datum several times, say r times, resulting in r completed data sets—that is, r data sets with the same number of variables and participants as the original one but with all missing values filled in by imputation. Without going into technical detail, which can be found elsewhere (10, 11), the imputations are randomly drawn from a distribution conveniently derived from the data, taking into account the relationship between variables and the relationship of each variable with the missing patterns in the remaining ones. Since imputations are random and not deterministic, a missing value may be replaced with a different value in each of the r completed data sets, and therefore the r data sets are not equal.

Once the r completed data sets have been obtained, the data analysis is straightforward. Suppose we are interested in a regression parameter β. Then, we would fit the regression model separately in each of the r data sets, to obtain r estimations of β ($\hat{\beta}_1, \ldots, \hat{\beta}_r$) and r estimations of the standard error. The r estimations are then combined using Rubin's formulas (12), and the final standard error is composed of a within-imputation part, which is the average of the r standard errors, and a between-imputation part, which reflects the variance that exists in the r estimations $\hat{\beta}_1, \ldots, \hat{\beta}_r$. If no missing data existed in the original data set, the r data sets would be equal and there would be no between-imputation variance. However, as the percentage of missing data increases and the imputation model becomes less predictive, the differences in the estimations of β increase and so does the between-imputation variance. Thus, the between-imputation variance reflects the uncertainty we have in the estimation of β because of the occurrence of missing data. If only a single imputation were performed, the missing data would incorrectly be treated as known, the uncertainty associated with the imputation process would not be accounted for, and the final standard errors would be too small (2). The multiple imputation procedure provides correct standard errors by accounting for the uncertainty associated with missing data imputation.

When inference is done on a population parameter, such as a regression coefficient, the estimations are always presented with a measure of its variability, such as the standard error, and the method of multiple imputation easily modifies that to incorporate the variability due to the imputation process. The situation is different in the context of cluster analysis. In that case, the objective is to classify each subject into a certain number of homogeneous groups according to the values of the variables of interest, so in this regard, inference is aimed at the individual and not at a population measure. In addition, it is often the case that a measure of uncertainty for cluster assignment is not presented in practice, and hence it is not clear how to incorporate the uncertainty due to imputations into the final results.

Multiple imputation provides valid results when the imputation model is correct and the missing-at-random hypothesis holds—that is, when the probability of data being missing does not depend on the unobserved data, conditional on the observed data (10). This will be assumed throughout this paper. Several methods for implementing multiple imputation exist. Here, we used the Stata implementation (StataCorp LP, College Station, Texas) of the method of chained equations to perform the imputations (5, 8, 11). This technique is very flexible, since every variable can have a different distribution (e.g., Gaussian, logistic, Poisson), and other conditions such as bounds on variables can be incorporated. Some authors have discussed the difficulties that arise when applying this method to large data sets and have provided guidance on how implement it (3, 7, 9).

OVERVIEW OF CLUSTER ANALYSIS

Cluster analysis is the process whereby data elements are classified into homogeneous groups (13). It is different from discriminant analysis, where the groups are known and the investigator only wants to determine what characteristics define these known groups. Numerous clustering methods exist which fall into different families, such as hierarchical, partitional, or model-based algorithms (13–15). Here we used the k-means algorithm, which is probably the most widely used partitional technique (16).

Ideally, one could find the optimal clustering of individuals by performing an exhaustive search of all possible partitions, but this is not computationally feasible. Thus, one needs to rely on heuristic algorithms, such as k-means, that reduce the search but are not guaranteed to reach a global solution. The k-means method requires the number of clusters, k, as an input. The algorithm starts with k initial clusters, which may be chosen randomly, using the output of a hierarchical clustering technique or using other criteria. Once the initial clusters are defined, the centroid of each cluster is calculated. In the next iteration, the distance of each individual to the cluster centroids is calculated, and each individual is moved to the cluster whose centroid is closest. Then, cluster centroids are recalculated, and the same process is repeated until no persons can be moved between clusters. The final solution will depend on the initial centroids, and for that reason it is suggested to run the algorithm several times using different starting values and choose the best solution according to some criterion (16).

To find the optimal number of clusters, k, several methods have been suggested. They usually involve repeating the clustering algorithm using different values of k and comparing the results using some criterion (16). The

comparison of the fit of 2 classifications with different numbers of groups is not straightforward, since some penalization for the number of clusters needs to be used to prevent choosing as many clusters as observations (17).

The number of variables included in the cluster analysis is a relevant issue. A problem arises when applying *k*-means or any other clustering algorithm to high-dimensional data. It has been shown that adding more variables to an analysis may degrade the final classification if the number of persons is small relative to the number of variables (18, 19). This is due to a bias/variance trade-off by which the variance associated with using many variables defeats the possible classification benefit derived from these variables. This leads to the suggestion of choosing a small number of salient variables to create the clusters. This problem is usually defined as *feature selection* (17–19). There is general agreement that it is good practice to stipulate that the number of subjects be at least 10 times the number of variables (18). Like the problem of choosing *k*, the problem of choosing the best subset of variables is not trivial. The main difficulty is that most of the existing criteria for evaluating clustering depend on distances, and distances are altered by changing the dimensionality (17). Therefore, it is difficult to compare 2 clustering classifications based on different numbers of variables.

Recently, a new criterion, called CritCF and defined in Web Appendix 1 (available at http://aje.oxfordjournals.org/), was developed. Higher values for CritCF are preferred. This criterion can rank partitions based on different numbers of clusters and different numbers of variables (17). With this property, one can design a search strategy to find both the optimal number of clusters and the final variables that need to be included in the analysis. As before, this search strategy will not be exhaustive, and achievement of the global optimum is not guaranteed. One possible search strategy is a backward sequential selection algorithm, which would work as follows (17). For the number of clusters $k = 2, \ldots, k_{max}$, where the maximum number of clusters is fixed a priori, the algorithm starts with a set of *selected variables* containing all variables and an empty set of *removed variables*. The *k*-means algorithm is run excluding one of the variables in the selected set at a time. The variable that, when excluded, provides a higher value of CritCF is removed from the set of selected variables and added to the set of removed variables. The process keeps iterating until none of the selected variables brings any improvement in CritCF when excluded. The combination of *k* and the set of variables that gives a higher CritCF is chosen as the final model.

INTEGRATION OF MULTIPLE IMPUTATION INTO CLUSTER ANALYSIS

In this section, we illustrate how multiple imputation can be integrated into a cluster analysis and suggest some ways to present the results to reflect how the uncertainty associated with the imputed values affects the final results—namely the number of clusters, the set of variables included in the final analysis, and the assignment of individuals to clusters. The process is illustrated using a subset of data from the Phenotype and Course of Chronic Obstructive Pulmonary Disease (PAC-COPD) Study (20). Briefly, this Spanish study (2004–2008) used clustering techniques to identify clinically and epidemiologically meaningful subtypes of chronic obstructive pulmonary disease (COPD) derived from a comprehensive list of variables on clinical, functional, and biological characteristics in a cohort of patients with COPD. The subset used here included 342 participants and 85 variables. The range of missing data in the different variables went from 0% to 47.7%, with 68.2% of the variables having less than 5% of the data missing. Only 13.7% of the participants had complete data on all 85 variables. Overall, 5.9% of the information was missing; that is, 5.9% of the total of 29,070 cells ($85 \times 342 = 29{,}070$) had missing data.

The steps of the proposed framework are described in Table 1. We started by applying a multiple imputation technique to obtain *r* completed data sets. In our example, *r* was set to 100 (in Web Appendix 2, we discuss how to choose a value for *r*; for example, using a figure like Web Figure 1). Next, we applied the backward search algorithm to each of the 100 data sets, with the maximum number of clusters explored, k_{max}, fixed arbitrarily to 5. As a result, we obtained 100 values for the optimal number of clusters, corresponding to the best solutions, according to CritCF, in each data set. The final number of clusters, k_{fin}, can be chosen as the one appearing with the highest frequency in the 100 data sets. In addition, we propose to show the distribution of the final number of clusters over the 100 data sets in order to illustrate the impact of the imputations on deciding the final *k*. The higher the frequency of k_{fin}, the lower the influence of the missing data on the optimal number of clusters. In the example, the optimal result was 2 clusters in 99 of the data sets and 3 in the remaining one. Therefore, 2 was selected as the final number of clusters (k_{fin}), and the missing data seemed to introduce little uncertainty into this decision. Another possible way to select the number of clusters would be to examine the distribution of CritCF over the 100 data sets by the number of clusters, as shown in Figure 1. Then one could choose, for example, the number of clusters with the highest median CritCF, which in this case would also be $k = 2$. The plot shown in Figure 1 gives a better sense of how much better a specific *k* performs when compared with others. However, direct information on the percentage of times one would end up choosing a value of *k* versus another is harder to extract from this figure. Often, but not necessarily, a high degree of overlap between 2 box plots implies that both values of *k* would beat each other in a significant number of data sets.

In the third step, we kept only the results for $k = 2$ and explored the set of variables that the backward search algorithm (described above) retained for each of the 100 imputed data sets. The final subset of variables is likely to be different for each data set. Since having different clustering rules based on different sets of variables for each data set impairs interpretation of the meaning of the clusters, we propose to begin by choosing a common set of variables that will subsequently be used to define the final partition in all data sets. As in the process of selecting the number of clusters, we also want to reflect the uncertainty that the imputations

Table 1. Proposed Framework for the Application of Multiple Imputation in Cluster Analysis

1. Multiple imputation to obtain r completed data sets.
2. Cluster analysis with variable selection algorithm for $k = 2, \ldots, k_{max}$ in each of the 1, . . ., r imputed data sets.
   2a. Decide the optimal number of clusters (k_{fin}).
      → For example, the one selected in most data sets according to the goodness-of-fit criteria.
   2b. Describe uncertainty in k_{fin} associated with missing data.
      → For example, show frequency of selection of each k or box plots of the goodness-of-fit criteria by k (Figure 1).
3. Retain only the results for the r data sets with $k = k_{fin}$.
   3a. Decide the variables to be included in the final analysis (we call this set of variables $\{var\}_{fin}$).
      → For example, keep the variables that are selected in at least 50% of the data sets; or keep the s most frequent variables, where s is determined by the median number of variables selected in the r data sets.
   3b. Describe the uncertainty in the variable selection process associated with missing data.
      → For example, show frequency of selection of each variable (Figure 2).
4. Refit the cluster analysis with $k = k_{fin}$ in r data sets containing only the variables in $\{var\}_{fin}$.
5. Relabel the clusters so that they all have the same meaning in the r data sets.
6. Allocation of subjects to clusters.
   6a. Assign each subject to a cluster.
      → For example, assign subjects to the cluster they were assigned to in most data sets.
   6b. Describe the uncertainty in subject allocation to clusters.
      → For example, describe the distribution of the frequency of assignment to each cluster (Table 2).
7. Description of clusters.
      → Description can be done by restricting each subject to be assigned to only 1 cluster, or taking into account that a subject may be assigned to different clusters in different data sets (Table 3).
8. Future analyses using the cluster membership of each subject.
      → For example, regression models including cluster membership as a dependent or independent variable. r different data sets with potentially different cluster membership for each subject should be used according to the usual multiple imputation procedure for regression models in order to account for the uncertainty of cluster assignment.

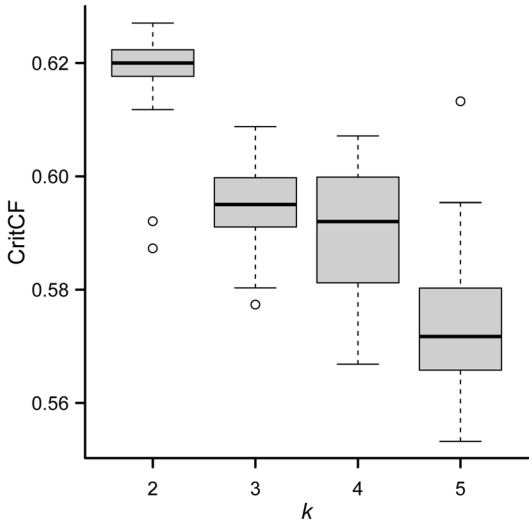

Figure 1. Box plots of the between-imputation distribution of CritCF by number of clusters (k), PAC-COPD Study, Spain, 2004–2008. Each box plot is based on 100 values. PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease.

introduce when making this decision. We propose presenting a table or figure with all variables that are selected in at least 1 of the 100 data sets, along with their frequency of appearance, as in Figure 2. In the example, 48 of the original 85 variables appeared in at least 1 data set.

Based on Figure 2, the investigator needs to make a decision on which variables to include in the final analysis. One possibility would be to include all those that appear at least once, even though many of them appear in a very small percentage of the imputed data sets. Retaining the variables that were selected in more than a given percentage of the data sets (e.g., 50%) is another alternative. An additional possibility would be to select the top s variables, where s is determined by some summary statistic of the number of variables selected in the r data sets. In the example, the median number of selected variables in the different data sets was 16, while the third quartile was 18 (Figure 2). In the example, choosing 16 variables (the median) was equivalent to choosing those that appeared in more than 50% of the data sets. Figure 2 shows a point where there is a substantial jump in the percentage of appearance (from 71% to 46%). This would be another way to make the choice, similarly to what is often done in factor analysis to decide the number of factors (21) or in the context of variable selection for prognostic models (22). In the example, this last criterion would lead to choosing the first 15 variables to perform the final

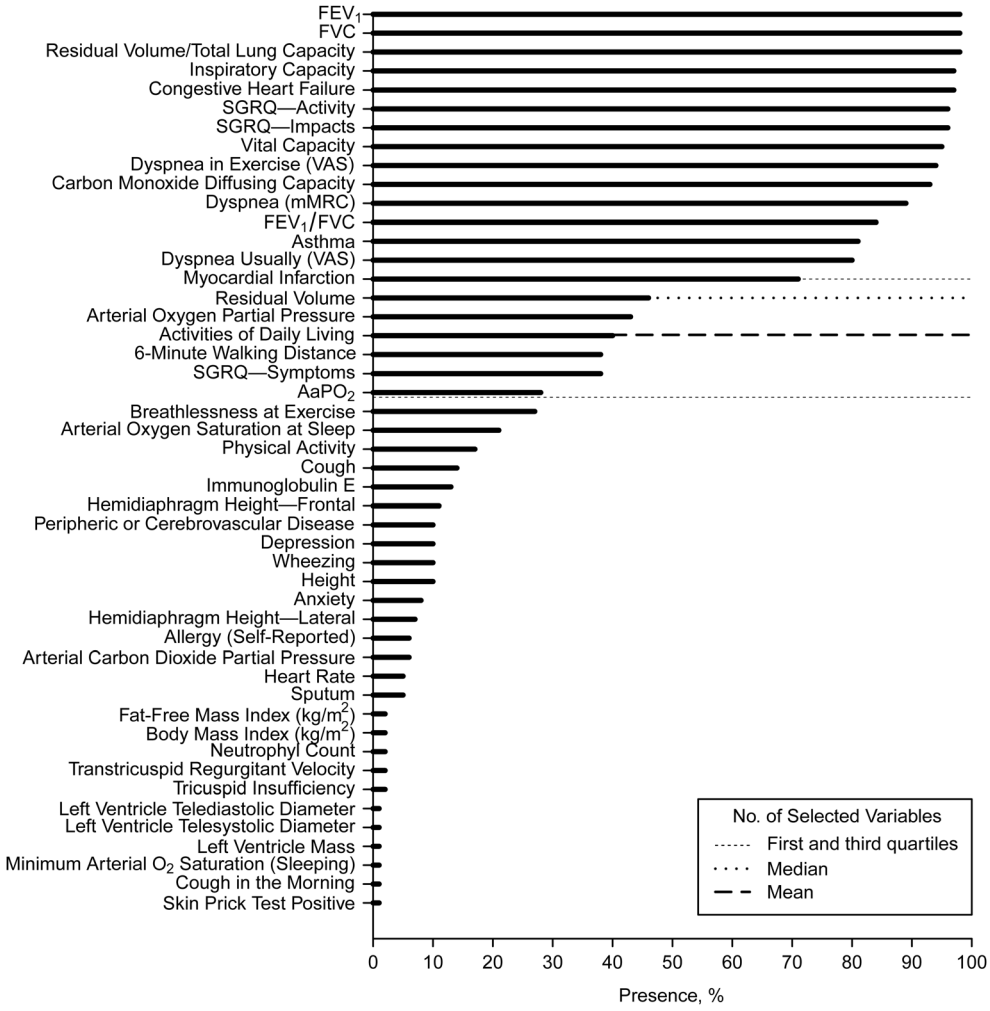

Figure 2. Variables that remained in the final set of variables in at least 1 data set after the variable selection algorithm was applied for $k = 2$, by percentage of appearance, PAC-COPD Study, Spain, 2004–2008. $AaPO_2$, oxygen alveolar-arterial partial pressure difference; FEV_1, [postbronchodilator] forced expiratory volume in 1 second; FVC, [postbronchodilator] forced vital capacity; mMRC, modified Medical Research Council scale; O_2, oxygen; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease; SGRQ, Saint George Respiratory Questionnaire; VAS, Visual Analogue Scale.

Table 2. Distribution of the Frequencies of Assignment (Proportions) for Each Cluster (for $k = 2$ and $k = 3$), PAC-COPD Study, Spain, 2004–2008

	Minimum	First Quartile	Median	Third Quartile	Maximum
$k = 2$					
Cluster 1	0.64	1	1	1	1
Cluster 2	0.51	1	1	1	1
$k = 3$					
Cluster 1	0.43	0.65	0.67	0.67	0.67
Cluster 2	0.41	0.49	0.64	0.67	0.68
Cluster 3	0.38	0.47	0.49	0.56	0.68

Abbreviation: PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease.

Table 3. Description of the Variables (Mean Values) by Cluster, PAC-COPD Study, Spain, 2004–2008[a]

Variable	% of Subjects With Missing Values	Raw Data			Imputed Data		
		Cluster 1 ($n = 186$)	Cluster 2 ($n = 156$)	*F* Value[b]	Cluster 1 ($n = 188$[c])	Cluster 2 ($n = 154$[d])	*F* Value[b]
FEV_1, % predicted	0.0	62.0	41.0	192.03	61.7	40.7	242.66
FVC, % predicted	0.0	79.6	64.4	66.50	79.3	64.4	93.82
Residual volume/total lung capacity, %	7.9	50.6	61.8	131.72	50.7	61.8	138.57
Inspiratory capacity, L	5.6	71.8	51.5	112.10	71.3	51.7	136.49
Congestive heart failure[e]	0.9	4.3	8.4	1.49	4.9	8.3	2.36
SGRQ—activity (range,[f] 0–100)	1.2	33.1	64.0	183.71	33.9	64.3	218.34
SGRQ—impacts (range, 0–100)	1.2	16.9	37.7	125.72	17.4	38.1	153.15
Prebronchodilator vital capacity, % predicted	5.6	76.9	59.8	105.23	76.6	60.0	137.30
Dyspnea in exercise (VAS) (range, 0–10)	1.2	4.1	6.4	67.79	4.1	6.5	77.14
Carbon monoxide diffusing capacity, % predicted	13.5	73.1	55.6	46.78	72.1	55.5	63.53
Dyspnea mMRC (range, 0–5)	1.2	2.0	3.3	94.28	2.0	3.3	111.02
FEV_1:FVC ratio, %	0.0	58.1	47.8	72.00	58.1	47.5	78.01
Asthma[g]	1.2	7.1	11.0	1.37	7.5	11.3	1.55
Dyspnea usually (VAS) (range, 0–10)	1.2	1.9	4.2	63.97	1.9	4.2	82.92
Myocardial infarction[e]	0.9	8.7	12.9	0.85	9.4	12.7	1.57
Residual volume, L	7.9	140.6	172.6	32.83	140.3	172.4	38.06

Abbreviations: FEV_1, [postbronchodilator] forced expiratory volume in 1 second; FVC, [postbronchodilator] forced vital capacity; mMRC, modified Medical Research Council scale; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease; SGRQ, Saint George Respiratory Questionnaire; VAS, Visual Analogue Scale.

[a] The columns under "Raw Data" show the results obtained when the final cluster assignment was decided by majority vote and the missing values of the variables were excluded from the calculation of the means. The columns under "Imputed Data" show the results obtained when using the 100 data sets with imputed missing values and variable cluster assignment.

[b] *F* values correspond to the ratio of the variance of the group means (between-group variance) to the overall variance of the variable, with higher values representing higher relevance of the variable for separating cluster groups. *F* values were obtained by means of linear regression models using each variable as the outcome and the cluster group as the exposure.

[c] Median over the 100 data sets. Minimum and maximum values were 176 and 215, respectively.

[d] Median over the 100 data sets. Minimum and maximum values were 127 and 166, respectively.

[e] Physician-diagnosed.

[f] Range of possible scores.

[g] Self-reported.

analysis. We chose to continue the analysis with the 16 most frequent variables, a number that fits well in the 10-subjects-per-variable rule.

In the fourth step, the cluster technique should be applied again to each of the 100 data sets using only the common set of selected variables and the selected number of clusters. Before proceeding to assign each subject to a cluster, a previous process (fifth step) needs to be performed to prevent a potential problem caused by the automatic labeling of the clusters by the statistical package. For example, in an analysis with 4 clusters, the group labeled "cluster 1" in one data set may be labeled "cluster 2" in another data set, precluding a correct interpretation. This problem can be solved owing to the fact that the same variables are used in all data sets; a strategy is described in Web Appendix 3.

After the fourth and fifth steps, each subject is not necessarily assigned to the same cluster in each of the 100 data sets. A final decision on the cluster to which each individual belongs (sixth step) can be taken by majority vote—that is, by assigning the individual to the cluster to which he/she was assigned with the highest frequency in the 100 data sets. The frequency of assignment is of interest by itself, since it can be interpreted as the person's probability of belonging to that cluster over the missing data distribution, and it provides a measure of missing-data-driven uncertainty in the classification of each individual. To provide a measure of uncertainty for the entire sample, one can show the distribution of these group-membership probabilities in every cluster group. For example, the minimum, the maximum, and the 3 quartiles can be of interest. As shown in Table 2, the missing data did not influence cluster assignment in the analysis with 2 clusters, since all subjects were assigned to the same cluster in almost all of the imputed data sets. For illustration only, we also show the results for 3 clusters. In that case, the frequencies of assignment were substantially lower. For example, persons who were finally assigned to cluster 3 were assigned with a median frequency of 50%—that is, in 50 out of 100 data sets, they were classified into cluster 2, but in the remaining 50% they were assigned to other clusters.

As a final step, we need to describe and interpret the clusters. The description can be based on either 1) the final cluster assignment (majority vote) and the original data set (with missing values) or 2) the multiply imputed data sets and the multiple assignments to clusters. The former implicitly assumes that the cluster membership is known, while the latter takes into account the fact that participants were not always assigned to the same cluster. The latter resembles the technique called *fuzzy k-means*, an extension of *k*-means where individuals have some probability of arising from different clusters (17). Results from the example using both approaches are shown in Table 3. Cluster 2 was characterized by a worse status in several disease domains: more airflow limitation, higher hyperinflation, more dyspnea, and worse quality of life, while subjects in cluster 1 exhibited less impairment in all disease domains.

Our framework includes a final step for cases where the variable indicating cluster membership is to be used in posterior analyses, for example, as part of a regression model. In that case, the usual multiple imputation framework for regression can be used. That is, one would fit the regression model to each of the 100 data sets separately, which will have different cluster assignments for some persons, and then pool the results using multiple imputation rules.

SIMULATION STUDY

We performed a small simulation study, described in Web Appendix 4. Briefly, we used a subset of 10 variables from our example, and kept only complete cases. This was used as the full data set in our simulation study. Then we artificially created different percentages of missing values. The resulting data sets were analyzed using our framework. The results were compared with the results obtained in the full data set (without missing values) and with a complete case analysis. In brief, the analyses with our framework detected the same number of clusters as the analyses with the full data set, and the agreement of the individual classification was very high (average $\kappa > 0.91$ in all scenarios). The variables selected were the same ones as those selected in the full data set, although in some simulations fewer variables were selected. Depending on the scenario, the complete case analysis ended up choosing 2 or 3 as the optimal number of clusters. When we forced $k = 2$, the same variables as in the full data set were selected. The classification agreement could not be examined for persons with missing values, because they are excluded in a complete case analysis. The classification agreement with the full data set for persons without missing values was poorer than with multiple imputation, even reaching very small values in some simulations (Web Table 1). Additional details are provided in Web Appendix 4.

DISCUSSION

We have provided herein a framework for using multiple imputation in cluster analysis when the data set contains missing values. We have stressed the importance of providing measures on how the missing data influence the different results of a cluster analysis and have suggested ways to characterize this uncertainty. Our framework is very flexible and allows using other methods for cluster analysis, measures of fit, and search strategies than the ones used here. This is important, since some methods may work better for some research questions or may even be superior in some data sets but not in others, and there is no clear guidance on which ones to use. However, the use of multiple imputation would be more helpful for non-model-based clustering techniques, since model-based methods such as latent class analysis can directly handle missing values in the estimation process (23).

Using multiple imputation in cluster analysis problems has important advantages over other commonly used ways of treating missing data, such as restricting the analysis to complete cases (24) or replacing missing values with a single value (25). In the first case, bias can be introduced and efficiency is lost, especially in high-dimensional data sets where a complete case analysis can lead to dropping a high percentage of the subjects. In the second case, although bias can be avoided by using an appropriate method of imputation, one does not take into account the uncertainty associated with imputation and how it affects the final results.

Throughout this paper, we have investigated only the uncertainty due to missing data. For example, we can interpret the percentage of times an individual is assigned to a cluster as his/her probability of belonging to a cluster *over the missing data distribution*, but the reader should note that this is not his/her probability of belonging to the cluster. There are other sources of uncertainty that are not taken into account when computing this probability, such as what would happen if the model were fitted with a different sample of individuals. Larsen (26) suggested a method for computing uncertainties about cluster membership based on the bootstrapping method and also showed how to integrate multiple imputation of missing values in that process. Similarly, uncertainty in variable selection due to sampling variation could also be incorporated via bootstrap (22).

We used a subset of the data from the PAC-COPD Study to illustrate our method (20). The PAC-COPD Study had a different research objective and used a different method than the one described here. The results presented here are only used for illustrative purposes, and readers interested in the interpretation of the results in the COPD context should refer to the original publication (20).

In conclusion, multiple imputation can be incorporated into cluster analysis. The proposed framework makes good use of all available information by using models to replace missing values, while at the same time providing an honest picture of how the uncertainty in the imputations affects each of the different outputs of a cluster analysis. An R package (R Foundation for Statistical Computing, Vienna, Austria) implementing this framework can be downloaded from http://www.creal.cat/xbasagana/software.html.

ACKNOWLEDGMENTS

Author affiliations: Centre for Research in Environmental Epidemiology, Barcelona, Spain (Xavier Basagaña, Jose

Barrera-Gómez, Marta Benet, Josep M. Antó, Judith Garcia-Aymerich); Hospital del Mar Research Institute, Barcelona, Spain (Xavier Basagaña, Jose Barrera-Gómez, Marta Benet, Josep M. Antó, Judith Garcia-Aymerich); Consorcio de Investigación Biomédica en Red especializado en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain (Xavier Basagaña, Jose Barrera-Gómez, Marta Benet, Josep M. Antó, Judith Garcia-Aymerich); and Department of Experimental and Health Sciences, Faculty of Health and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain (Josep M. Antó, Judith Garcia-Aymerich).

Jose Barrera-Gómez and Marta Benet contributed equally to this work.

The PAC-COPD Study was supported by grants from the Fondo de Investigación Sanitaria (grants PI020541, PI052486, PI052302, and PI060684), Ministry of Health, Madrid, Spain; the Agència d'Avaluació de Tecnologia i Recerca Mèdiques (grant 035/20/02), Catalonia Government, Barcelona, Spain; the Spanish Society of Pneumology and Thoracic Surgery (grant 2002/137); the Catalan Foundation of Pneumology (grant 2003 Beca Marià Ravà); the Red Respira (grant C03/11); the Red de Centros de Investigación Cooperativa en Epidemiología y Salud Pública (grant C03/09); the Fundació La Marató de TV3 (grant 041110); and Novartis Farmacèutica, Barcelona, Spain. The CIBERESP is funded by the Instituto de Salud Carlos III, Ministry of Health, Madrid, Spain. Dr. Judith Garcia-Aymerich holds a research contract from the Instituto de Salud Carlos III (grant CP05/00118).

We thank Drs. Juan Ramón González and Alejandro Cáceres for their help with compilation of the R package.

The funding sources had no involvement in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the article for publication. The researchers were independent of the funders.

Conflict of interest: none declared.

REFERENCES

1. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008;168(4):355–357.
2. Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087–1091.
3. He Y, Zaslavsky A, Landrum M, et al. Multiple imputation in a large-scale complex survey: a practical guide. *Stat Methods Med Res*. 2010;19(6):653–670.
4. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol*. 2010;171(5):624–632.
5. Royston P. Multiple imputation of missing values. *Stata J*. 2004;4(3):227–241.
6. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol*. 2010;172(4):478–487.
7. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol*. 2009;169(9):1133–1139.
8. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681–694.
9. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377–399.
10. Schafer JL. *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall, Inc; 1997.
11. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simul*. 2006;76(12):1049–1064.
12. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
13. Milligan GW. Cluster analysis. In: Kotz S, Campbell BR, Balakrishnan N, et al, eds. *Encyclopedia of Statistical Sciences*. New York, NY: John Wiley & Sons, Inc; 1998:120–125.
14. Jain AK, Topchy A, Law MHC, et al. Landscape of clustering algorithms. In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*. Washington, DC: IEEE Computer Society Press; 2004:I-260–I-263.
15. Zhong S, Ghosh J. A unified framework for model-based clustering. *J Mach Learn Res*. 2003;4:1001–1037.
16. Steinley D. K-means clustering: a half-century synthesis. *Br J Math Stat Psychol*. 2006;59(1):1–34.
17. Breaban M, Luchian H. A unifying criterion for unsupervised clustering and feature selection. *Pattern Recognit*. 2011;44(4):854–865.
18. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell*. 2000;22(1):4–37.
19. Wang Y, Miller DJ, Clarke R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *Br J Cancer*. 2008;98(6):1023–1028.
20. Garcia-Aymerich J, Gomez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax*. 2011;66(5):430–437.
21. Cattell RB. The screen test for the number of factors. *Multivariate Behav Res*. 1966;1(2):245–276.
22. Heymans MW, van Buuren S, Knol DL, et al. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med Res Methodol*. 2007;7:33.
23. Lanza ST, Collins LM, Lemmon DR, et al. PROC LCA: a SAS procedure for latent class analysis. *Struct Equ Modeling*. 2007;14(4):671–694.
24. Sugar CA, James GM, Lenert LA, et al. Discrete state analysis for interpretation of data from clinical trials. *Med Care*. 2004;42(2):183–196.
25. Sandborgh M, Lindberg P, Denison E. Pain Belief Screening Instrument: development and preliminary validation of a screening instrument for disabling persistent pain. *J Rehabil Med*. 2007;39(6):461–466.
26. Larsen MD. Multiple imputation for cluster analysis [abstract]. In: *Abstracts From the Joint Conference of the Classification Society of North America and Interface Foundation of North America, Washington University in St. Louis, June 8–12, 2005*. Paris, France: International Federation of Classification Societies; 2005:9. (http://www.classification-society.org/meetings/csna05/abstracts2005csna.pdf). (Accessed February 18, 2013).