Eur Respir J 2010; 36: 472–474
DOI: 10.1183/09031936.00035210

EDITORIAL

Use of cluster analysis to define COPD phenotypes

M. Weatherall*,#, P. Shirtcliffe#,¶, J. Travers¶ and R. Beasley#,¶

The current classification of airways disorders is imprecise, with an overlap of phenotypes (*e.g.* asthma, chronic bronchitis and emphysema), resulting in difficulties in differentiating the disorders from each other. This has led to considerable diagnostic, management and prognostic uncertainty. The traditional approach has been to present this phenotypic overlap in the Venn diagram format [1]; however, this results in $\geqslant$15 phenotypes, whose pathogenesis or response to treatment have not been clearly defined [2, 3]. More recent work [4–8], including that of BURGEL *et al.* [8] published in the current issue of the *European Respiratory Journal*, has used cluster analysis to characterise different types of airways disorders. But what is cluster analysis, is it a reasonable approach to take, and how valid are the conclusions?

Cluster analysis is a collection of methods for defining groups of individuals based on measured characteristics, so that they are grouped based on their differences (or similarities), into clusters [9–11]. The groupings are constructed such that the degree of association is strong between members of the same cluster and weak between members of different clusters [4].

Cluster analysis is distinct from other ways of trying to understand multivariate data, which include principal component and factor analysis, discriminant analysis and multivariate regression. Principal component (as used by BURGEL *et al.* [8]) and factor analysis produce linear combinations of measured variables, in the sense that new derived variables are produced by multiplying each of the original variables by a scaling parameter and adding the resulting numbers. Discriminant analysis (as also used by MOORE *et al.* [5]) starts with known groups and finds scaled combinations of the measured variables that best distinguish those known groups. Multivariate regression can have a set of response variables predicted by a set of explanatory variables.

There are three major considerations in designing a cluster analysis. The first relates to selection of the individuals. If the individuals are, in fact, too similar, then finding clusters within a relatively homogenous group may be misleading. For example, if individuals with airflow obstruction are selected from a tertiary referral centre, then cluster analysis may simply identify phenotypes that represent referral patterns; for example, a lack of response to inhaled corticosteroids. If individuals are too disparate, then this may result in outlying groups being put in very small clusters that do not reflect a meaningful underlying disease process. A random population survey can overcome these selection effects but is likely to include fewer individuals with severe disease.

The second consideration is selection of variables for measurement. Variables should reflect putative mechanisms and clinical characteristics of different phenotypes. Obviously, one wants to choose variables that have the largest chance of being discriminatory between clusters. We acknowledge that this is to some extent a chicken and egg problem (one is performing a cluster analysis to find the groups distinguished by the variables one chooses). One should also avoid variables that are close to measuring the same thing, as the extra noise generated may obscure the clusters. If the variables represent epiphenomena, then clusters may represent these rather than underlying pathogenic or clinical features. Other considerations are whether treatments modify the values of variables chosen (*e.g.* inhaled corticosteroids affect those related to variable airflow obstruction) or if the disease process modifies the values of the variables that define the disease (*e.g.* variable airflow obstruction due to airways inflammation may lead to irreversible airflow obstruction due to remodelling).

A third consideration is how many variables to choose to enter into a cluster analysis. The key here is, once you have put subjects into clusters, you need some way of looking back to the original variables to describe the clusters. If too many variables are used it will be difficult to describe the clusters in a meaningful way. One of the purposes of seeking phenotypes of obstructive airways disease, often unstated, is to generate an allocation rule so future patients can be classified. If a very large number of variables are entered into a cluster analysis, then this means the underlying relationships of the variables to different phenotypes are not well understood. We suggest that around 10 variables may be a useful number, but we acknowledge that determining the optimal number should be a subject of future research. Dimension-reduction techniques such as principal components analysis (as used in BURGEL *et al.* [8]), where many variables could be related to different phenotypes, may offer a way of reducing the number of variables entering into an analysis. In our view, the clinical meaning of these derived variables is uncertain and places the clusters at some distance from the clinical variables from which they are derived.

Cluster analysis is not usually based on a probability distribution for the underlying groups. In general this means that it is not usual to perform statistical tests on a cluster structure for any particular data set and method. Cluster analysis can always find clusters in data, even if data sets are completely unstructured. It

*University of Otago Wellington, #Capital and Coast District Health Board, and ¶Medical Research Institute of New Zealand, Wellington, New Zealand.

CORRESPONDENCE: R. Beasley, Medical Research Institute of New Zealand, Private Bag 7902, Wellington 6242, New Zealand. E-mail: Richard.Beasley@mrinz.ac.nz

has been suggested that this lack of a basis in formal statistical modelling means that cluster analysis is probably best seen as hypothesis-generating rather than -solving [4].

There are a large number of ways of actually carrying out cluster analysis [9–11]. There can be pre-processing of the actual measured variables; for example, by performing a principal components analysis of the measured variables to find a smaller subset of derived variables that capture the measured information in a smaller number of dimensions. There are a number of measurements of distance (such as the Euclidean distance and Gower's distance), depending on whether variables are continuous, ordinal or binary (or a mixture of these). There are also a large number of methods of creating clusters from these distance measurements. Two broad classes of doing this are hierarchical and nonhierarchical methods. In hierarchical methods, individuals and clusters are, most commonly, merged (agglomerative) or, less commonly, divided (divisive). For these hierarchical methods there are, in turn, a large number of ways of determining the proximity of clusters. Nonhierarchical methods also exist; for example, the k-means approach (used by HALDAR *et al.* [7]), which relies on defining some values to tentatively identify clusters and building clusters around these. Another method assumes a mixture of multivariate, normally distributed clusters is present, and based on some assumptions about the shape of the clusters, uses information criteria to determine the optimal number of clusters [9, 10].

Once individuals are placed into clusters, relevant meaning must be given to these clusters. For example, can the clusters be described in a way that does, in fact, reflect the underlying aetiology and the clinical, physiological and immunological features that are assessed in practice? Importantly, can the clusters give guidance for allocation of other individuals to the phenotypic groups represented by the clusters? Although cluster analysis is dependent on the choice of individuals, variables and methodology, it is more data-driven than other methods of defining phenotypes and may therefore be less susceptible to bias by historical and *a priori* assumptions.

The main conclusion from BURGEL *et al.* (as they state in their discussion [8]), is that chronic obstructive pulmonary disease (COPD) patients with similar airflow obstruction belong to different phenotypes, and have different symptoms (dyspnoea), outcomes (exacerbation numbers and predicted mortality) and differ in terms of age and comorbidities. It is interesting to compare this paper with the other two papers that have used cluster analysis to characterise COPD as summarised in table 1 [4, 6]. The theme that emerges from these analyses (which all differ in terms of the source of research participants, variables chosen, cluster method and subsequent clusters) is that there is a real need for a multidimensional assessment of COPD. At a more specific level, it is worthy of note that both WARDLAW *et al.* [4] and WEATHERALL *et al.* [6] identified a cluster characterised by severe and markedly variable airflow obstruction with features of atopic asthma, chronic bronchitis and emphysema. Patients in this phenotypic group would be unlikely to meet the inclusion criteria of the major randomised, controlled trials of either asthma [12] or COPD [13]. As a result, there is not a strong evidence base for the management of this important group of patients with the most severe disease and morbidity [3, 4, 13]. BURGEL *et al.* [8] also comment on the implications of cluster analysis for clinical trials.

TABLE 1. Summary of three papers using cluster analysis to identify chronic obstructive pulmonary disease (COPD) phenotypes

First author [ref.]	Source of research participants	Variables	Cluster method	Summary of clusters
BURGEL [8]	Respiratory clinic patients with COPD, excluding asthma	Age, BMI, dyspnoea score, exacerbation rate, FEV_1 % pred, HAD scale, pack-yrs, SGRQ score	Principal components analysis on variables, Euclidean distance, Ward's method on derived variables	Young, severe disease, underweight Young, moderate disease Old, mild disease, overweight Old, moderate disease, overweight
WARDLAW [4]	Not stated, well-characterised patients with asthma or COPD	Age, bronchodilator reversibility, FEV_1/FVC ratio, FEV_1 % pred, pack-yrs, serum IgE, sex, % sputum eosinophils	Z-score distance measurement, Ward's method	Classical COPD Asthma/COPD overlap Asthma with high sputum eosinophils and high IgE Asthma with low sputum eosinophils and low IgE
WEATHERALL [6]	Community-based random sample with FEV_1/FVC <70% or a history of current wheeze	Bronchodilator reversibility, $D_{L,CO}$/VA % pred, F_{ENO}, FEV_1/FVC ratio, FEV_1 % pred, FRC % pred, pack-yrs, serum IgE, sputum production	Gower's distance, average distance method	Emphysema Asthma/COPD overlap Atopic asthma with increased F_{ENO} Airflow obstruction and sputum production without increased F_{ENO} Airflow obstruction without other features

FEV_1: forced expiratory volume in 1 s; FVC: forced vital capacity; BMI: body mass index; % pred: % predicted; HAD: hospital anxiety and depression; SGRQ: St George's Respiratory Questionnaire; Ig: immunoglobulin; $D_{L,CO}$: diffusing capacity of the lung for carbon monoxide; VA: alveolar volume; F_{ENO}: fraction expired nitric oxide; FRC: functional residual capacity.

Where to from here? We agree with WARDLAW *et al.* [4] that these techniques seem particularly suited to the study of diseases that express considerable diversity and as such are ideally placed to address the multidimensional complexity apparent in airways disorders. Further cluster analyses, both population-based and clinic-based, will contribute to a greater understanding of the true patterns of airways disorders. The clinical application of cluster analysis will depend on developing diagnostic criteria to allow new individuals to be allocated to groups based on the identified clusters, as illustrated by MOORE *et al.* [5]. Ultimately, whether different treatment strategies provide different outcomes for these groups will provide confirmation, or otherwise, of the clinical value of cluster analysis. This knowledge could lead to different pharmacological treatments and other interventions directed at specific phenotypic groups [14]. We consider that achieving this goal is worthy of the research endeavour required.

STATEMENT OF INTEREST

None declared.

REFERENCES

1 American Thoracic Society. Standards for diagnosis and care of patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1995; 152: s77–s121.
2 Marsh SE, Travers J, Weatherall M, *et al.* Proportional classifications of COPD phenotypes. *Thorax* 2008; 63: 761–767.
3 Gibson PG, Simpson JL. The overlap syndrome of asthma and COPD: what are its features and how important is it? *Thorax* 2009; 64: 728–735.
4 Wardlaw A, Silverman M, Siva R, *et al.* Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 2005; 35: 1254–1262.
5 Moore WC, Meyers DA, Wenzel SE, *et al.* Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med* 2010; 181: 315–323.
6 Weatherall M, Travers J, Shirtcliffe PM, *et al.* Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009; 34: 812–818.
7 Haldar P, Pavord ID, Shaw DE, *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; 178: 218–224.
8 Burgel P-R, Paillasseur J-L, Caillaud D, *et al.* Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010; 36: 531–539.
9 Everitt R. An R and S-Plus Companion to Multivariate Analysis. London, Springer-Verlag, 2005.
10 Khattree R, Naik DN. Multivariate Data Reduction and Discrimination with SAS software. Cary, SAS Institute, 2000.
11 McLachlan GJ. Cluster analysis and related techniques in medical research. *Stat Meth Med Res* 1992; 1: 27–48.
12 Travers J, Marsh S, Williams M, *et al.* External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax* 2007; 62: 219–223.
13 Travers J, Marsh S, Caldwell B, *et al.* External validity of randomized controlled trials in COPD. *Respir Med* 2007; 101: 1313–1320.
14 Beasley R, Weatherall M, Travers J, *et al.* Time to define the disorders that make up the syndrome of COPD. *Lancet* 2009; 374: 670–672.