Review

Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping

Elsie Horne, BSc, MSc; Holly Tibble, BSc, MPhil(Cantab); Aziz Sheikh, BSc, MSc, MBBS, MD; Athanasios Tsanas, BSc, BEng, MSc, DPhil(Oxon)

Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Elsie Horne, BSc, MSc
Usher Institute, Edinburgh Medical School
University of Edinburgh
Nine Edinburgh Bio Quarter
9 Little France Road
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 1316517887
Fax: 44 1316517887
Email: Elsie.Horne@ed.ac.uk

Abstract

Background: In the current era of personalized medicine, there is increasing interest in understanding the heterogeneity in disease populations. Cluster analysis is a method commonly used to identify subtypes in heterogeneous disease populations. The clinical data used in such applications are typically multimodal, which can make the application of traditional cluster analysis methods challenging.

Objective: This study aimed to review the research literature on the application of clustering multimodal clinical data to identify asthma subtypes. We assessed common problems and shortcomings in the application of cluster analysis methods in determining asthma subtypes, such that they can be brought to the attention of the research community and avoided in future studies.

Methods: We searched PubMed and Scopus bibliographic databases with terms related to cluster analysis and asthma to identify studies that applied dissimilarity-based cluster analysis methods. We recorded the analytic methods used in each study at each step of the cluster analysis process.

Results: Our literature search identified 63 studies that applied cluster analysis to multimodal clinical data to identify asthma subtypes. The features fed into the cluster algorithms were of a mixed type in 47 (75%) studies and continuous in 12 (19%), and the feature type was unclear in the remaining 4 (6%) studies. A total of 23 (37%) studies used hierarchical clustering with Ward linkage, and 22 (35%) studies used k-means clustering. Of these 45 studies, 39 had mixed-type features, but only 5 specified dissimilarity measures that could handle mixed-type features. A further 9 (14%) studies used a preclustering step to create small clusters to feed on a hierarchical method. The original sample sizes in these 9 studies ranged from 84 to 349. The remaining studies used hierarchical clustering with other linkages (n=3), medoid-based methods (n=3), spectral clustering (n=1), and multiple kernel k-means clustering (n=1), and in 1 study, the methods were unclear. Of 63 studies, 54 (86%) explained the methods used to determine the number of clusters, 24 (38%) studies tested the quality of their cluster solution, and 11 (17%) studies tested the stability of their solution. Reporting of the cluster analysis was generally poor in terms of the methods employed and their justification.

Conclusions: This review highlights common issues in the application of cluster analysis to multimodal clinical data to identify asthma subtypes. Some of these issues were related to the multimodal nature of the data, but many were more general issues in the application of cluster analysis. Although cluster analysis may be a useful tool for investigating disease subtypes, we recommend that future studies carefully consider the implications of clustering multimodal data, the cluster analysis process itself, and the reporting of methods to facilitate replication and interpretation of findings.

(JMIR Med Inform 2020;8(5):e16452) doi: 10.2196/16452

KEYWORDS

asthma; cluster analysis; data mining; machine learning; unsupervised machine learning

Introduction

Background

There is mounting evidence to suggest that some disease labels are in fact *umbrella terms*, which encompass distinct disease subtypes with different underlying mechanisms and clinical symptom manifestations [1-3]. This has encouraged the investigation into heterogeneity within disease populations, which has received considerable interest across diverse domains of medicine [4-6]. There are numerous motivations for better understanding heterogeneity within disease populations, from the development of targeted therapeutics [6] to the delivery of more personalized care in clinical practice [7].

It is now understood that asthma is one such umbrella term used to encompass multiple diverse underlying disease symptoms and pathophysiology [7]. Asthma is a common chronic condition characterized by reversible airway obstruction. The Global Burden of Disease Study 2017 estimated the global prevalence of asthma (both symptomatic and asymptomatic) to be 273 million [8]. This study estimated that in 2017, there were 43 million new cases of asthma and 495,000 deaths attributed to asthma [9]. Attempts to categorize asthma into distinct disease subtypes date back to the 1940s [10] and are ongoing. However, the methods for discovering these underlying categories have shifted from observing clinical patterns to using data-driven approaches such as *cluster analysis* [11].

Cluster analysis is a statistical technique used to identify subgroups in data based on multiple variables (for convenience, herein, we have used the term *features*). It is an *unsupervised* statistical learning method, and the correct number of underlying clusters is typically unknown *a priori* [12]. The technique has found increasing use in recent years because of the practical unmet clinical need to identify subtypes of disease and stratify patients to improve health care delivery. This has been made feasible by the increasing availability of clinical datasets and the development of statistical software packages facilitating the application of algorithmic methods.

Clinical datasets are often *multimodal*; for the purposes of this paper, we defined a multimodal dataset as a dataset that includes features from different sources, measured on different scales. For completeness and to avoid ambiguity, we clarified that the term multimodal has a different meaning in statistical literature (ie, features with multiple modes in terms of its distribution); the use of the term in this study is aligned with clinical literature (having features from different sources). Popular methods of cluster analysis such as k-means and hierarchical clustering with the Ward method have been developed for continuous features measured on a common scale. In practice, however, many of these techniques are frequently applied to multimodal clinical datasets comprising different feature types measured on different scales, conditions that violate some of the underlying principles and assumptions made by algorithmic methods [13]. Although steps can be taken to prepare multimodal clinical data for cluster analysis [13], the results of a previous review suggest that these steps are rarely taken in practice [11]. This previous review focused on the clinical findings of the studies and touched only briefly on the challenges of clustering multimodal data specifically.

Objectives

This review aimed to comprehensively explore whether studies applying cluster analysis to multimodal clinical data to subtype asthma are using appropriate clustering methodologies. The contribution of this study is to make recommendations for the robust application of cluster analysis to multimodal clinical data. We believed this would be of interest to the ever-growing number of asthma researchers engaging or planning to engage in disease subtyping, as well as to the wider community of researchers applying cluster techniques for the purpose of disease subtyping.

Methods

Eligibility Criteria and Search Strategy

This review is reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Multimedia Appendix 1 shows the completed PRISMA checklist.

We sought to identify studies that applied cluster analysis to multimodal clinical data with the aim of identifying subtypes of asthma. One researcher (EH) searched PubMed and Scopus databases (search queries are provided in Textbox 1) to retrieve studies focusing on patients diagnosed with asthma, which included the term *cluster analysis* or *clustering*. Our search was restricted to studies published between January 1, 2008, and May 23, 2019, as Haldar et al's study [14] is widely acknowledged to be the first to apply cluster analysis to identify subtypes of asthma. Our search excluded comment articles, editorials, letters, reviews, and meta-analyses. We excluded articles that were not written in English.

We excluded nonrelevant studies by first screening the abstracts, then referring to the full text where necessary. We excluded studies in which (1) none of the aims or objectives were to identify subtypes of asthma (studies looking exclusively at, eg, childhood wheeze were excluded); (2) the data were not multimodal (ie, were measured from a common source and on a common scale); and (3) none of the features were considered clinical (eg, studies concerned only with -omics data). Finally, we excluded studies that used latent class analysis or mixture models to group their data to narrow the scope of this review to methods that cluster samples based on pairwise dissimilarities. The use of latent class analysis to distinguish asthma phenotypes has been reviewed previously by Howard et al [15].

Textbox 1. Search query to identify studies to include in this review.

- The following query was inserted in PubMed on May 23, 2019:

  English[Language] AND ("2008/01/01"[Date - Publication] : "2019/05/23"[Date - Publication]) AND ("cluster analysis"[Text Word] OR "clustering"[Text Word]) AND "asthma*"[Text Word] NOT (comment[Publication Type] OR editorial[Publication Type] OR letter[Publication Type] OR review[Publication Type] OR meta-analysis[Publication Type])*

- The following query was inserted in Scopus on May 23, 2019:

  PUBYEAR > 2007 AND (TITLE-ABS-KEY ("cluster analysis") OR TITLE-ABS-KEY("clustering")) AND TITLE-ABS-KEY ("asthma*") AND SRCTYPE ("j") AND DOCTYPE ("ar") AND LANGUAGE ("English")*

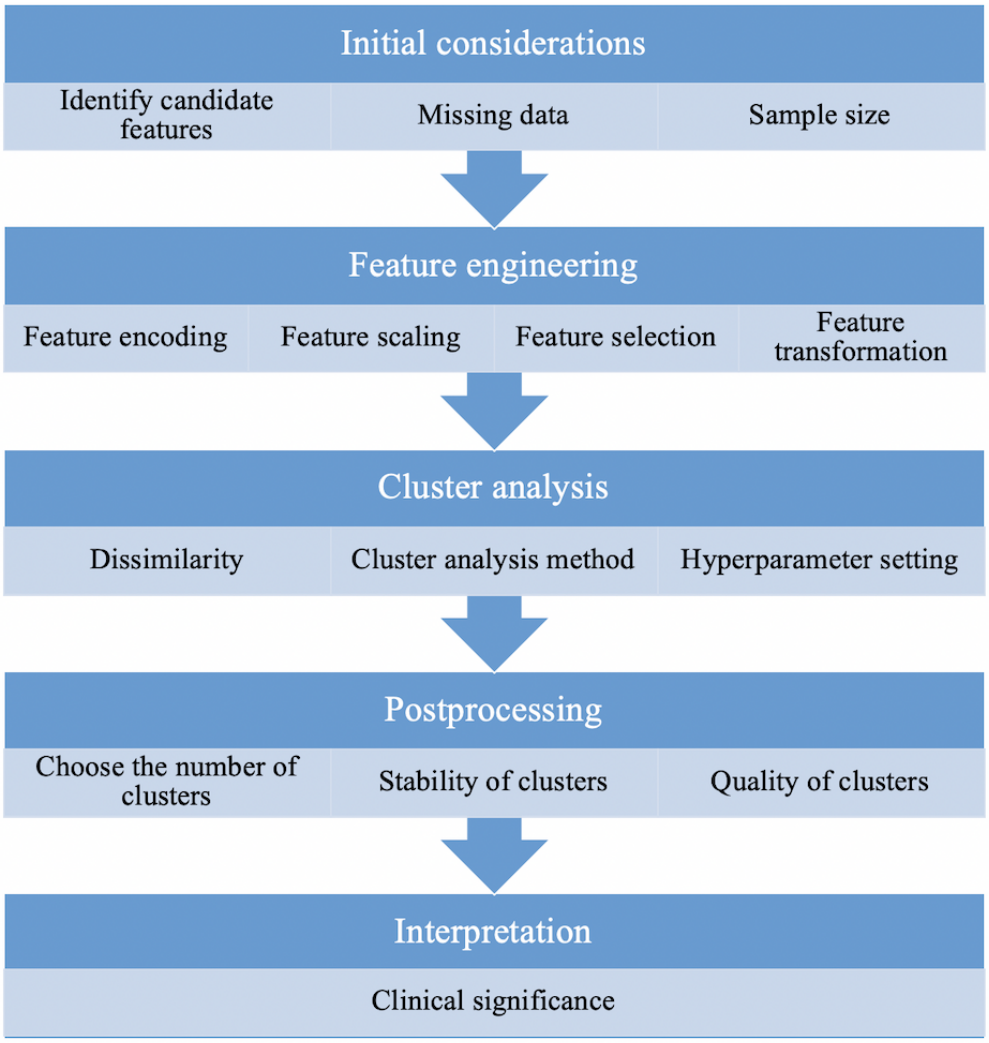
Data Extraction

In total, 2 researchers (EH and HT) independently extracted information from the full text and supplementary material of each study. Information was extracted following the steps outlined in the following *Cluster Analysis Steps* section. The data dictionary, which provides details of all items extracted, is presented in Multimedia Appendix 2.

Cluster Analysis Steps

To provide context for this review, we outlined the key steps in the application of cluster analysis to multimodal clinical data. Figure 1 summarizes the steps in the order in which they generally occur, but as with most analytic processes, this depends on the context, and the process may be somewhat iterative.

Figure 1. Schematic of the typical cluster analysis steps.



Initial Considerations

Identify Candidate Features

The first step is to identify the set of features of interest, which we referred to as *candidate features*. These may be identified based on previous studies or clinical input using domain expertise. In some cases, all the candidate features may be used in the cluster analysis (we referred to the features used in cluster analysis as *cluster features*). In other cases, formal feature selection processes may be applied to the candidate features to identify the cluster features, as covered in the *Feature Selection* section.

Missing Data

Most common cluster analysis methods use *complete case analysis* (ie, the cluster features have no missing entries, which, in practice, might be achieved by removing samples for which any cluster feature entry is missing). However, it may be more data efficient to develop a strategy to work around missing entries instead of discarding samples. Missing values may be handled through the calculation of dissimilarities, as described by Hastie et al [16]. Alternatively, missing data could be imputed, or for categorical features, a missing category could be introduced.

Sample Size

Despite the widespread use of cluster analysis, at present, there is no consensus regarding the minimum sample size required to ensure stable and meaningful clustering. Dolnicar et al [17] suggested that 70 samples per cluster feature is adequate, based on the findings of their simulation study. Small sample sizes may obscure the true clustering by causing the user to pick the wrong number of clusters (see the *Choosing the Number of Clusters* section) or by producing solutions that are neither reproducible nor stable (see the *Stability* and *Quality* subsections).

Feature Engineering

Feature Types

The features that we may want to use in a clustering algorithm often come from multimodal clinical data. Hence, they may be of different types (eg, continuous, nominal, ordinal, binary, etc) and are likely to be measured on different scales (eg, kilogram for mass, years for age). Most dissimilarity measures and clustering algorithms assume that the features are of the same type and are measured on a common scale. These requirements can be addressed using *feature encoding* and *feature scaling*.

Feature Encoding

When dealing with categorical features, it is vital to consider how these are encoded (nominal, ordinal, or binary), as this determines how they are treated in the calculation of dissimilarities and in the clustering algorithm. A common approach is to encode ordinal features as integers and to encode nominal features as dummy binary features [18].

Feature Scaling

Feature scaling may be used to address 3 issues related to continuous features. The first is that continuous features may be measured in different units and should therefore be rescaled to bring them onto a common scale before calculating dissimilarities. The second is that continuous features measured in the same units may have different variances. In some cases, the differences in variance may be useful for clustering, but in others, these may obscure the true underlying cluster structure in the data. In the latter case, the continuous features should be rescaled. Common approaches to these 2 issues are to standardize features to have 0 mean and unit variance (referred to as *z-scores*) or to use range normalization techniques, for example, to scale each feature so that it is in the interval of 0 to 1.

The third issue is that the features may not follow the desired probability distribution properties for further analysis (eg, having Gaussian-distributed features). This issue needs to be considered when statistical methods make distributional assumptions. Although few dissimilarity-based clustering methods make distributional assumptions, several methods involve the calculation of cluster means (eg, k-means, hierarchical clustering with the Ward linkage). The mean is a poor choice of summary statistic for a feature that is skewed (or a feature with multiple modes), so a power transformation may be advantageous as a preprocessing step when using such clustering methods.

When dealing with mixed-type data, it may be necessary to scale the categorical features to avoid assigning categorical features greater weight over continuous features or vice versa. This issue is discussed in detail in the context of dissimilarity measures by Hennig and Liao [13].

Dimensionality Reduction

There are generally 2 motivations for reducing the dimensionality of a dataset before applying cluster analysis. First, as previously mentioned in the *Sample Size* subsection, datasets with a high feature to sample ratio may not produce stable cluster results. Second, the cluster structure may only be apparent using a subset of the information available in the data. Using all available information may introduce noise, which could obscure the true underlying cluster structure [19]. There are 2 approaches to dimensionality reduction: *feature selection* and *feature transformation*.

Feature Selection

Feature selection involves selecting a subset of the available features for use in cluster analysis. Herein, we have referred to the features selected for the cluster analysis as *cluster features*.

Feature Transformation

Feature transformation involves combining original features to create new features. Generally, a subset of these new features is selected for inclusion in the analysis. It is beyond the scope of this review to provide in-depth details on the methods of feature transformation (also known as *feature extraction*); we referred to van der Maaten et al's [20] work for a comprehensive review. Here, we briefly outlined *principal component analysis* (PCA), which is the most commonly used method for linear data projection. PCA may be applied to p continuous, correlated features to extract $m<p$ continuous, and uncorrelated features (known as *principal components*), each being a linear function of the original cluster features [21]. Related methods include factor analysis for continuous data, *multiple correspondence*

analysis (MCA) for categorical data [22], and multiple factor analysis for mixed-type data [23].

Cluster Analysis

Dissimilarity Measures

Model-free clustering methods rely on a *dissimilarity measure* to quantify how dissimilar 2 samples are from one another. Dissimilarity may also be referred to as a *distance measure* if it satisfies the triangle inequality. The most widely used dissimilarity measure is the squared Euclidean distance (henceforth referred to as *Euclidean distance*), which is intended for use with continuous features. A dissimilarity measure that can handle both categorical and continuous features is the Gower distance [24].

Cluster Analysis Methods

There are many different methods of cluster analysis (eg, k-means, hierarchical clustering with the Ward linkage, spectral clustering), and each method may be implemented using different algorithms. A comprehensive overview of the wide range of clustering methods can be found elsewhere [25].

Postprocessing

Choosing the Number of Clusters

A key challenge in cluster analysis is choosing the number of clusters to present in the final solution, which is typically unknown *a priori*. Often, researchers use their preferred clustering methods, running them for 2 to *k* clusters (where *k* is an integer number indicating the number of clusters) and then have a strategy to determine *k*.

Providing a detailed commentary on these strategies is beyond the scope of this review. An overview of strategies for choosing *k* is provided by Everitt et al [23]. Graphical techniques include dendrograms (when using hierarchical clustering methods) and silhouette plots [26]. An alternative approach is to choose the number of clusters that gives the most stable solution [27]. In practice, a key determinant in choosing the number of clusters is often the clinical interpretation of the solutions.

We highlighted the possibility that there might not be meaningful clustering of the data to form groups, and thus, the entire dataset is treated as 1 cluster. This may reflect the lack of statistical power (sufficiently large sample size) to determine clusters or that the investigated problem using that dataset is not amenable to clustering using the available sample size and features. Some statistics used for choosing k, such as the Gap statistic [28], can be calculated for k=1. However, statistics that require the calculation of between cluster differences or distances, such as the silhouette statistic, are not defined for k=1 [26].

Stability

Assessing the quality of a clustering solution produced using any cluster algorithm is challenging. Unlike supervised learning setups, there is no *ground truth* against which one can formally test their findings. However, there are several ways in which one can assess the integrity of their findings.

Most importantly, it is crucial to assess the *stability* of the resulting clusters. A definition of *cluster stability*, given by von Luxburg [27], is whether clustering different datasets sampled from the same underlying joint distribution will result in producing the same clusters. There are several ways in which this may be assessed in practice (eg, by comparing the cluster results of a dataset that has been randomly split into 2 or more subsets, and each subset is independently fed into the cluster algorithm).

Quality

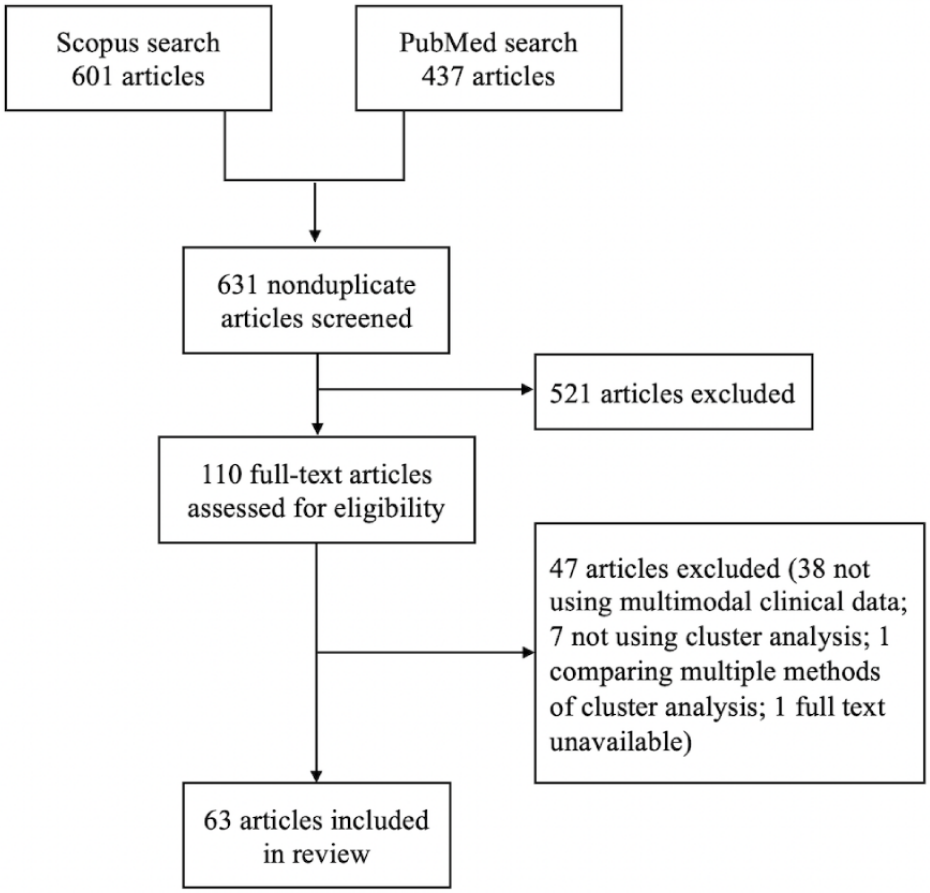
Beyond stability, there are numerous steps one may take to ensure the integrity of their cluster analysis findings, for example, repeating the analysis in a different cohort or at a different time point, or altering the encoding of a feature. These steps are often referred to as reproducibility testing. However, we avoided this term because it implies that we seek the exact same results, which we do not feel is reasonable in all scenarios. To extract this information from the studies in this review, 2 reviewers independently extracted details of postprocessing methods, which we felt assessed the quality of the cluster results, but did not come under stability. In our schematic and results, we referred to these methods as testing the quality of the cluster results.

Results

Literature Search Outcomes

We identified 63 studies that used cluster analysis to identify subtypes of asthma using multimodal clinical data (Figure 2). One of the excluded articles satisfied our inclusion criteria but investigated 85 combinations of cluster analysis steps in a hierarchical cluster analysis of 383 children with asthma [29]. We excluded this study from our review as including all 85 combinations of methods was deemed infeasible. For the 2 studies in which cluster analysis was carried out in multiple populations [14,28], we included only the analysis of the larger population. The characteristics of each study are presented in Multimedia Appendix 3.

Figure 2. Flow of studies into review.



Initial Considerations

Identifying Candidate Features

A total of 42 (67%) studies identified candidate features based on previous studies or clinical input (relevance to asthma subtypes, avoiding clinical redundancy, and easily measured in clinical practice). The numbers used in each method are summarized in Table 1.

Missing Data

A total of 42 (67%) studies detailed their methods for dealing with missing data; the methods used are shown in Table 1. The most common method was to carry out a complete case analysis by excluding all patients with any missing cluster feature entries (35% of studies).

Table 1. Initial considerations across the asthma studies we have included in this review (N=63).

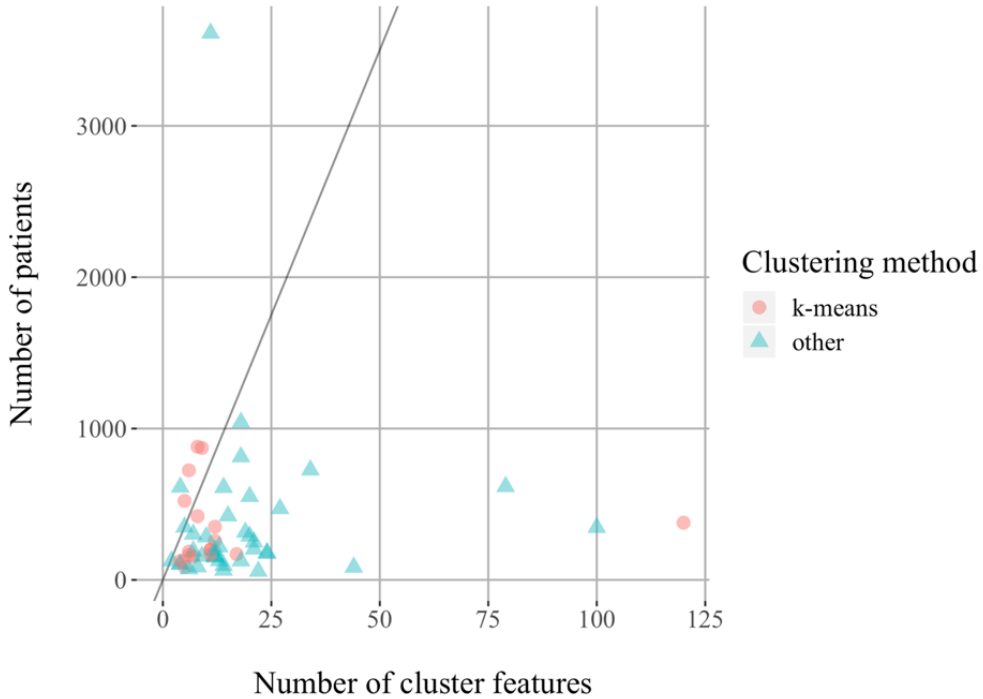
Method	Values, n (%)[a]
Identifying candidate features	
Clinical intuition and understanding	33 (52)
Avoid clinical redundancy	15 (24)
Previous studies	15 (24)
Easily measured in clinical practice	8 (13)
Missing data	
Complete case analysis	22 (35)
Features with >x%[b] missing values removed	14 (22)
Imputed	11 (17)
Patients with >x%[b] missing values removed	5 (8)
No missing data present	2 (3)
Clustering methods handle missing data	1 (2)

[a]One study may use multiple methods; some studies may use no methods.
[b]x>0.

Sample Size

The sample sizes for cluster analysis ranged from 40 to 3612, with a median of 195 patients. Figure 3 presents a scatter plot of the number of patients in the cluster analysis versus the final number of cluster features. The straight line corresponds to the number of samples per feature as recommended by Dolnicar et al [17]. As this estimate was derived from simulation studies using k-means as the clustering method, different markers are used for the studies which used clustering techniques other than k-means. Note that the studies that did not specify the final number of cluster features were omitted from the plot. Six studies (10%) had at least 70 times as many patients as cluster features, as recommended by Dolnicar et al [17].

Figure 3. Number of patients versus final number of cluster features. The line corresponds to the number of patients that is equal to 70 times the number of features.



Feature Engineering

Feature Scaling and Encoding

Judging whether feature scaling and encoding were appropriate depends on the methods of cluster analysis used and vice versa. Therefore, we reported the methods of feature scaling and encoding alongside the methods of cluster analysis in Tables 2-4 and Multimedia Appendix 4.

Table 2. Breakdown of methods used by studies applying hierarchical clustering with Ward's linkage (N=23).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
Not detailed	N/A[a]	1 (4)
Mixed		
Euclidean assumed		
Scaled but method unspecified	Yes	1 (4)
	No	1 (4)
Scaled to lie in the interval of 0 to 1	Yes	1 (4)
z-scores	Yes	1 (4)
	No	1 (4)
Not detailed	Yes	3 (13)
	No	6 (26)
Euclidean stated		
z-scores	Yes	2 (9)
	No	1 (4)
Gower[b]		
Gower standardisation	No	3 (13)
Scaled but method unspecified	No	1 (4)
treeClust		
Not detailed	No	1 (4)

[a]N/A: not applicable (irrelevant for continuous features).

[b]Computing the Gower coefficient normalizes the distance between feature samples by dividing by the feature range. Therefore, it is not necessary to normalize continuous features prior to computing the Gower coefficient.

Table 3. Breakdown of methods used by studies applying k-means (N=22).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
z-scores for one feature	N/A[a]	1 (5)
No details	N/A	3 (14)
Euclidean stated		
No details	N/A	1 (5)
Mixed		
Euclidean assumed		
Scaled but method unspecified	No	1 (5)
z-scores	Yes	6 (27)
z-scores for one feature	No	1 (5)
No details	Yes	1 (5)
	No	2 (9)
Euclidean stated		
z-scores	Yes	1 (5)
No details	No	1 (5)
Unclear		
Euclidean assumed		
No details	No	3 (14)
Euclidean stated		
z-scores	No	1 (5)

[a]N/A: not applicable (irrelevant for continuous features).

Table 4. Breakdown of methods used by studies applying SPSS TwoStep (N=7).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
No details	N/A[a]	1 (14)
Mixed		
Log-likelihood assumed		
Scaled to lie in the interval 0 to 1	Yes	1 (14)
z-scores	No	1 (14)
No details	Yes	2 (29)
Log-likelihood stated		
Scaled but method unspecified	No	1 (14)
No details	No	1 (14)

[a]N/A: not applicable (irrelevant for continuous features).

Univariate Feature Transformation

A total 23 (37%) studies applied univariate feature transformation to bring features closer to a normal distribution. The most common univariate feature transformation was logarithmic transformation, applied to nonnormally distributed features in 33% of studies. Lefaudeux et al [30] applied the Box-Cox transformation to all features, whereas Khusial et al [31] stated that data were transformed if necessary but gave no further details.

Feature Selection

A total of 22 (35%) studies detailed methods of feature selection to identify their cluster features. The number of features selected in the 63 studies included in this review ranged from 2 to 120, with a median of 12 features. In addition, 47 (75%) studies had mixed-type features, and 12 (19%) had continuous features, and in 4 (6%) studies, the type of features was unclear. Methods for feature selection are listed in Table 5.

A total of 13 (20%) studies used PCA or factor analysis for feature selection. These are not typically methods that should be used for feature selection; we defer further elaboration on the topic for the Discussion. All but one of these studies computed the components (or factors) that represent an underlying latent feature structure, then selected 1 (or in some cases multiple [32,33]) original feature corresponding to each component (or factor) of the latent feature structure. Just et al [34] stated that they used PCA to select features according to statistical significance. As PCA does not involve the computation of statistical significance (*P* values), more detail would be required here to fully understand the methods used for feature selection in this paper. Pérez-Losada et al [35] stated PCA based on Euclidean distances was carried out. It is unclear whether this was an error in reporting or whether PCA was applied to the matrix of Euclidean distances between features instead of the covariance matrix. To implement the latter approach, the Euclidean distances would have to be converted to similarities. Moreover, the authors stated that PCA was used *to identify key clinical components relevant to asthma diagnosis and assessment*. Overall, it is not clear how the authors processed the data using PCA, and there was no justification for using Euclidean distances in that computation. Although the application of PCA leads to the computation of features (principal components) that maximally explain the (remaining) variance in the data, there is no guarantee that the resulting principal components will be highly predictive of an outcome (in this case, asthma diagnosis and assessment).

Table 5. Feature engineering methods used in the asthma studies included in this review.

Method	Values, n (%)[a]
Univariate feature transformation	
Logarithmic transformation	21 (33)
Box-Cox transformation	1 (2)
Method not explained	1 (2)
Feature selection	
Factor analysis[b]	8 (13)
Principal component analysis[b]	5 (8)
Avoid collinearity	3 (5)
Avoid multicollinearity	3 (5)
Supervised learning methods	2 (3)
Multiple correspondence analysis	1 (2)
Feature transformation	
Principal component analysis	4 (6)
Factor analysis	1 (2)
Multiple correspondence analysis	1 (2)

[a]As a percentage of all 63 studies.

[b]These are not typically methods of feature selection but have been used in these studies.

Three (5%) studies considered collinearity via pairwise correlations, although the exact criteria for selection features based on this were unclear [36-38]. In addition, 3 (5%) studies avoided multicollinearity, but none detailed their methods for doing so [39-41].

Furthermore, 2 (3%) studies selected features using statistical hypothesis tests with respect to the outcome of interest. Sakagami et al [42] used mean annual decline in forced expiratory volume in 1 second as the outcome feature in a multiple regression analysis using stepwise feature selection. All features with coefficients statistically significantly different to 0 in the multiple regression model were included as cluster features. Seino et al [43] grouped participants according to whether or not they had symptoms of depression. Features were selected for cluster analysis if the difference between the 2 groups (tested using a Wilcoxon rank-sum or chi-square test for continuous and categorical features, respectively) was statistically significant.

Feature Transformation

A total of 6 (10%) studies performed feature transformation before cluster analysis; the methods are summarized in Table 5. Of the 4 studies that used PCA for feature transformation, 3 used continuous input features [30,44,45], whereas the fourth used mixed-type input features [46]. None of the studies stated

whether the covariance or correlation matrix was used as input for PCA. Only Newby et al [45] specified the number of transformed features retained, and the proportion of original variance accounted for.

Khusial et al [31] performed factor analysis on a subset of the selected features; it is unclear whether categorical features are included in this subset. Although the resulting factors were scaled to z-scores, the authors did not provide further information regarding whether the features were scaled before factor analysis. Four factors were retained, but neither the proportion of variance explained by these factors nor a table of the factor loadings is given.

Sendín-Hernández et al [47] performed MCA to transform 5 continuous and 14 categorical features. They gave the proportion of variance explained by the transformed features but gave neither the number of transformed features retained nor a table of the feature loadings.

Cluster Analysis

Hierarchical Clustering

A total of 23 (37%) studies applied hierarchical clustering with the Ward method [48] as the principal clustering technique. A breakdown of the methods used by these studies is given in Table 2. One study applied these methods to continuous data, and the remaining 22 studies used mixed-type data. Three studies stated that the Euclidean distance was used, 4 used Gower coefficient (issues with the Gower coefficient combined with the Ward method are addressed in the *Discussion* section), and 1 used tree-based dissimilarity measure [49]. For the remaining 15 studies, we assumed that the Euclidean distance was used. Of the 23 studies, 11 did not detail whether the features were rescaled. Of the 17 studies using the Euclidean distance with mixed-type features, 8 encoded categorical features as binary features.

A total of 3 (5%) further studies (in addition to the 23 studies introduced at the start of the paragraph) applied hierarchical clustering to continuous data. Amore et al [39] used the average linkage and the Euclidean distance, whereas 2 studies used hierarchical clustering but did not specify the linkage or dissimilarity measure used [44,50].

k-Means

A total of 22 (35%) studies used k-means clustering as the principal clustering technique. A breakdown of the methods used by these 3 studies is given in Multimedia Appendix 4. A breakdown of the methods used by these studies is given in Table 3. Five studies applied k-means to continuous data, and 13 studies applied it to mixed-type data. In 3 studies, the cluster features were not explicitly stated, and the data types therefore were unclear. Of the 22 studies, 4 explicitly stated that the Euclidean distance was used. As no other dissimilarity metrics were mentioned, we assumed that the Euclidean distance was used in the remaining 18 studies because it is often the default option for most algorithmic packages. Of the 22, 11 studies did not detail whether continuous features were scaled before cluster analysis. Of the 13 studies with mixed-type data, 8 encoded categorical features as binary features.

Preclustering Methods

When dealing with very large sample sizes, it can be advantageous to introduce a precluster step. The aim is to group samples and to use these groups or *preclusters* as input to a follow-on clustering algorithm (ie, using 2 steps with cascaded cluster algorithms). This step is used to reduce the computation time required to compute the cluster results.

A total of 7 (11%) studies used the SPSS TwoStep clustering method [51,52]. A breakdown of the preprocessing methods and distance measures used by these studies is given in Table 4. In the first (precluster) step, a cluster feature tree is identified. In the second step, the preclusters are merged stepwise until all clusters are in 1 cluster using the Euclidean or log-likelihood distance for continuous or mixed-type features, respectively. An advantage of the log-likelihood distance measure is that it is designed to handle mixed-type features. However, in doing so, it assumes that continuous (categorical) features follow a normal (multinomial) distribution within clusters.

None of the studies in this review adequately considered the distributional assumptions made by the SPSS TwoStep method. Ruggieri et al [53] acknowledged that the method assumes continuous features are normally distributed, but they did not explicitly report whether these assumptions were satisfied. Although Newby et al [45] acknowledged that the method assumes cluster features are statistically independent within clusters, they only go as far as to ensure that their cluster features are uncorrelated (by applying PCA), which does not necessarily imply independence. The remaining 5 studies that used the SPSS TwoStep method did not reference distributional assumptions.

Two (3%) further studies preclustered samples (Just et al [34] specified k-means, and Ye et al [54] did not specify the precluster method) and then applied hierarchical clustering with the Ward linkage method on the preclusters. A breakdown of the methods used by these 2 studies is given in Multimedia Appendix 4.

k-Medoid Methods

Three studies used k-medoid methods. A breakdown of the methods used by these 3 studies is given in Multimedia Appendix 4. Two used k-medoids implemented by the Partition Around Medoids algorithm [55]. Lefaudeux et al [30] used the Euclidean distance with center-scaled continuous data, and Sekiya et al [56] used the Gower metric with mixed-type data. Loza et al [57] applied fuzzy partition-around-medoid clustering with the Euclidean distance to continuous data scaled with average absolute deviation.

Kernel k-Means and Spectral Clustering

Kernel k-means and spectral clustering are different but related methods, which may be used to identify clusters that are not linearly separable in the input feature space [58]. As these methods were used by only 1 study each (Wu et al used multiple kernel k-means [59], and Howrylak et al used spectral clustering [37]), we do not explore them in detail in this review. However, details of the feature scaling, encoding, and distance measures used by these 2 studies is given in Multimedia Appendix 4.

Unclear Methods

Wang et al [41] described a 2-step clustering method in which the first step was to carry out hierarchical clustering using the Ward method, but with the log-likelihood distance in place of the Euclidean distance. This first step was used to determine the number of clusters, which was then used in the k-means method in the second step. However, the authors cite the SPSS TwoStep method [52], which is different from that described previously. It was therefore ambiguous which clustering method was applied in this study.

Postprocessing

Choosing the Number of Clusters

A total of 54 (86%) studies explained in detail the methods used to select the number of clusters. Of these, 20 (32%) studies used more than one method for choosing the number of clusters. The maximum number of methods used was 6.

A total of 27 (43%) studies used a dendrogram to choose the number of clusters to include in their study (Table 6). Note that 18 of the 22 studies that applied k-means clustering used hierarchical cluster as a first step to identify the likely number of clusters. Of these 18 studies, 11 explicitly stated that the dendrogram was used to choose the number of clusters.

Of the 8 (13%) studies that specified a maximum number of clusters, the maximum number ranged between 2 and 15 clusters. Seven (11%) studies used a statistic (or multiple statistics), including the c-index [60], Gap statistic [37], deviation from ideal stability [30], Calinski and Harabasz index [30], Dunn's partition [57], cubic cluster criterion (CCC) statistic [28], pseudo-F statistic [28,36], and pseudo-T2 statistic [28,36].

Four studies (6%) avoided very small clusters. Approaches to this include merging 2 clusters containing 6 and 12 samples [61], omitting small clusters containing 1 [35] and 6 [62] samples, and choosing the number such that no cluster contained less than 10% of the total samples [63].

Stability

A total of 11 (17%) studies tested the stability of their cluster solution; the methods are detailed in Table 6. Of these, 1 study used 2 methods, and the remaining 10 each used only 1 method to test stability.

Quality

A total of 24 (38%) studies assessed the quality of their solution using methods beyond those assessing stability. The methods are detailed in Table 6. Of these, 3 used more than one method. The maximum number of methods used in this study was 4.

Of the 30 studies that assessed the stability or quality of their cluster analysis, 21 (70%) reported their findings. However, the reporting of these results was in many cases brief, consisting of statements such as "the clusters were shown to be stable" without providing supporting evidence.

Table 6. Postprocessing methods used in the asthma studies included in this review.

Method	Values, n (%)[a]
Choosing the number of clusters	
Dendrogram	27 (43)
Hierarchical clustering with Ward linkage	19 (30)
Specify a maximum number of clusters[b]	8 (13)
Statistic(s)	7 (11)
Silhouette plot or average silhouette width	5 (8)
Bayesian information criterion	4 (6)
Specify a minimum size of smallest cluster[b]	4 (6)
Previous studies	3 (5)
Unclear	3 (5)
Clinical interpretation	2 (3)
Scree plot	1 (2)
Stability	
Repeated in random subset	3 (5)
Leave-one-out cross-validation	3 (5)
Bootstrap methods	3 (5)
Unclear methods	2 (3)
Train and test set	1 (2)
Quality	
Repeated in selected subset	8 (13)
Repeated with difference methods	6 (10)
Repeated with different initial configurations	5 (8)
Repeated in separate cohort	4 (6)
Repeated with altered features	3 (5)
Repeated at different time point	3 (5)
Repeated with different software	1 (2)

[a]Studies may have used more than 1 method.

[b]These methods were not included when calculating the number of methods used to choose the number of clusters.

Discussion

Principal Findings

We identified 63 studies that applied cluster analysis to multimodal clinical data to identify subtypes of asthma. We explored the clustering methodologies and their limitations in detail. The principal finding of this review was that the majority of the reviewed studies have flaws in the application of cluster analysis. Although some of these flaws were related to the multimodal nature of the clinical data, they extended to aspects of cluster analysis, which are agnostic of data type, such as sample size, stability, and reporting of the results.

These findings build on a previous review, which identified limitations such as lack of robustness in feature selection and neglect to specify distance measures in studies using cluster analysis to contribute to our understanding of the spectrum of asthma syndrome [11]. Our review investigated the methods of feature engineering more generally and identified not only neglect to specify dissimilarity measures but also instances in which the dissimilarity measure was inappropriate for the data to which it was applied. In addition, we identified issues related to sample size, cluster analysis methods, choosing the number of clusters, and testing the stability and quality of results. These issues are discussed in the following paragraphs.

A widespread limitation in the reviewed studies was the small sample size. Studies had overall sample sizes as small as 40 patients, with clusters as small as 6 patients. We argue that there is limited utility in clustering data with such small sample sizes: they may result in clusters that are unstable [64] and may therefore lead to selecting fewer clusters than are present in the underlying population from which the data are sampled.

In the following paragraphs, we discussed the limitations of 3 of the feature selection approaches applied by the reviewed

studies. The first approach was to avoid collinearity or multicollinearity or excluding features that were considered to be *clinically redundant*. Although one should avoid including features that are *redundant* (can be completely deduced from a combination of the other cluster features), this is rarely the case. Therefore, removing features inevitably leads to loss of information. We suggest that the removal of features based on redundancy needs to be carefully considered, for example, 2 or more features (some of which may appear univariately redundant) may jointly contribute toward determining a cluster (or similarly toward the estimation of a clinical outcome in a standard supervised learning setup).

The second was the use of PCA or factor analysis to select features, which has a similar motivation to the concept described earlier for discarding statistically correlated features. There are methodological justifications for the use of PCA, factor analysis, or other nonlinear embedding methods for feature transformation [19]. They aim to jointly combine the original features and project them in a new feature space, which may have some useful properties, including interpretation, determining latent feature structure, and improving the clustering or statistical mapping outcomes [16]. However, we suggest exercising caution toward using these methods for feature selection as described in some of the studies summarized in the Results section of this review because they were fundamentally developed toward different aims. Haldar et al used PCA for feature selection in the first publication to apply cluster analysis to identify asthma subtypes [14]. It is possible that other studies used this as a point of reference for these methods, leading to the common application of these methods in the field of asthma subtyping.

The third approach to feature selection was the use of statistical hypothesis tests with respect to outcomes of interest, as done in 2 studies [42,43]. Methods in which an outcome of interest is used to guide feature selection in cluster analysis have been described previously [65,66]. Although these approaches may be useful for situations in which there exists an outcome of particular interest to the clustering problem, the user should be aware of and acknowledge the assumptions made in the process. In the context of the 2 reviewed studies that used this approach, Sakagami et al did not acknowledge the linearity assumption in linear regression [42], whereas Seino et al's method does not account for potentially highly correlated features [43], a concept that is key in feature selection for cluster analysis.

Feature transformation was applied in only 6 studies, and the methods were generally poorly reported. As with cluster analysis, feature encoding and scaling are important considerations in feature transformation, but none of the studies gave adequate details in their methods. The results of feature transformation were also poorly reported. Although the key reason for applying feature transformation methods is to reduce the dimensionality of the dataset, only 2 [31,45] of the 6 studies provided details on the number of features retained. We suggest that the results of PCA, factor analysis, or MCA should include a table of component (or factor) loadings, the number of features retained, and the proportion of variance accounted for in the transformed features.

Most studies explicitly stated the clustering method that they used but were less explicit regarding the preprocessing steps and choice of dissimilarity measure. Hastie et al [16] state, "Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm."

We expand on this statement, further adding that preprocessing steps such as feature scaling and feature encoding are also more important in obtaining success than the choice clustering algorithm. This is in line with the conclusions of Prosperi et al, who demonstrated that clustering using different feature sets and encodings in asthma datasets can lead to different cluster solutions [29]. Both preprocessing steps and dissimilarity measures, along with their relation to clustering algorithms, have been given poor consideration in clustering applications in asthma, as discussed in the following 3 paragraphs.

First, the Euclidean distance was used with mixed-type data in over half of the studies (54%). Although the Euclidean distance is intended for use with continuous data, problems associated with applying it to mixed-type data may be mitigated by carefully considering feature scaling and feature encoding. However, in our review, we found that many studies did not specify their methods for rescaling, and many studies included ordinal and nominal categorical features but did not specify how these would be treated when calculating the Euclidean distances. The lack of consideration of feature scaling and encoding in these cases may have resulted in assigning an unintended weight structure to the cluster features.

Second, 4 studies used Gower coefficient in hierarchical clustering with Ward linkage [36,67-69], and 1 used tree-based distances [49,70]. These studies should be given some credit for using dissimilarities that can handle mixed-type data. However, the application of hierarchical clustering with Ward linkage relies on the properties of the Euclidean distance in the computations. These properties do not hold for Gower coefficient, and hence, errors are perpetuated at each level of the hierarchy. An example that demonstrates this issue is given in Multimedia Appendix 5.

A final point in the use of k-means and hierarchical clustering using the Ward method with mixed cluster features is that the theory underpinning these methods involves the calculation of cluster means. The mean is not an appropriate summary statistic for categorical features, which are more typically summarized by the mode. For this reason, we suggest that k-medoids may be a more appropriate method for mixed-type features used in clustering. Instead of computing each cluster's mean (as with hierarchical clustering using Ward's method and k-means), k-medoids compute each cluster's medoid, defined as the sample in the cluster for which the average dissimilarity to all other samples in the cluster is minimized [55]. In addition, k-medoids do not rely on the properties of the Euclidean distance in the computations, thus avoiding the issue described in the previous paragraph. Despite these advantages, only 2 studies in this review used k-medoids [30,56].

The SPSS TwoStep method was used in 7 of the 63 studies investigated here. We see 2 key limitations with the application of this method across the reviewed studies. First, none of the

studies gave adequate consideration to the distributional assumptions made when using the log-likelihood distance, and most did not mention the assumptions at all. Second, this method is designed for clustering several millions of samples with many features within an acceptable time and makes a key compromise in doing so [52]. This compromise is that the data are not stored in the main memory but are read sequentially, hence making the solution sensitive to the ordering of the data. None of the studies acknowledged this inherent shortcoming, nor did they confirm that their data were in a random order. Perhaps, more concerningly, the studies that applied these methods actually had very small datasets (range 84-349 samples) that could easily be stored, therefore making other standard techniques more appropriate. In our view, this compromise was therefore unnecessary.

Only 1 study [57] used a method that obtains a *fuzzy* cluster solution (in which a patient may be assigned a membership value to multiple clusters), as opposed to a *hard* cluster solution (in which each patient is assigned to a single cluster) [23]. A fuzzy cluster solution can indicate where a patient membership value is similar across multiple clusters, whereas this information is lost (or leads to lack of stability) in a hard cluster solution. Owing to the noisy nature of clinical data and the clinical complexity of grouping patients into distinct groups, we suggest that fuzzy cluster solutions may be more appropriate than hard cluster solutions in the review applications in asthma. However, it is important to acknowledge that there are added challenges in the interpretation and communication of fuzzy cluster solutions and that the methods may be more computationally intensive [71].

Selecting the number of clusters can be challenging and depends largely on the context of the application. In the case of the reviewed applications in asthma, the *true* number of clusters is unknown, and the analyses are exploratory. Although 86% of the review studies gave some details regarding their methods for choosing the number of clusters (*k*), they were generally poorly reported. The most popular approach was the dendrogram, but only Labor et al [72] specified their criteria for cutting the dendrogram. In 14 studies, the dendrogram was the only method mentioned. We suggest that more than one method should be used to select the number of clusters to validate this decision.

Our review shows that studies rarely tested the stability and quality of their results, with a particular lack of emphasis on stability. This is concerning, as many studies use methods such as k-means, which reach local minima, and apply them to small sample sizes, thus increasing the risk of obtaining unstable results. We argue that because of the unsupervised nature of cluster analysis, testing the stability and quality of the results should be a key theme and would like to urge researchers and peer reviewers in this research field to carefully consider these aspects. However, we do appreciate that assessing the stability and quality of a solution in the absence of *ground truth* is challenging and that there are currently no well-established frameworks for doing so [27].

Although this review focused on applications in subtyping asthma, the identified issues have been found in studies using cluster analysis to subtype other diseases. For example, recent studies in autism [73] and hypersomnolence [74] have applied cluster analysis to very small samples (55 and 17 patients, respectively). A recent study on Parkinson disease [75] stated in the main text that a *model-based* cluster analysis method was used, whereas the supplementary materials revealed that the method was in fact k-means, which is not model-based. In addition, supplementary materials listed 3 methods for choosing the number of clusters (CCC, pseudo-F, and R-squared statistics) but did not present the results from these 3 methods anywhere in the main text or supplementary materials. These findings demonstrate the widespread nature of the issues that this review has highlighted, and that the issues are not restricted to asthma-related studies.

For a recent example of a well-considered and well-reported application of cluster analysis to multimodal clinical data, we refer the reader to Pikoula et al's study of Chronic Obstructive Pulmonary Disease subtypes [76]. The main text and supplementary materials provide a transparent report of the methodology with respect to feature engineering and cluster analysis methods. In particular, Pikoula et al performed a rigorous assessment of the stability, reproducibility, and sensitivity of the resulting clusters, which could be used as a framework for future studies. The results that were key to the study's conclusions (eg, MCA feature loadings, silhouette plots, results from stability, reproducibility, and sensitivity analyses) are correctly reported in the manuscript, enabling readers to have a thorough understanding of the study's findings.

Limitations

The literature search presented in this study is comprehensive but practically cannot be exhaustive. We restricted the search to articles that included the terms *cluster analysis* or *clustering**. Although it is not strictly speaking correct to do so, some studies in the medical literature use the term *classification* to refer to cluster analysis, often confusing the 2 terms and sometimes using them almost interchangeably, for example, see the studies by Just et al [34] and Kim et al [46]. Widening the search to identify studies that use the term *classification* would have greatly increased the initial number of results of the PubMed search, but we suspect that the increase in the number of eligible studies for cluster analysis identified would have been small. Similarly, the terms *latent class analysis* and *mixture model analysis* might sometimes be erroneously used to refer to cluster analysis: we clarify that these terms were not included in our search strategy. As this is not a systematic review, we feel that our search criteria are fully sufficient for this study's purposes.

We did not fully explore multiple kernel k-means [77] or spectral clustering [78] methods, each used by 1 study in this review. As with all other cluster analysis methods mentioned here, careful consideration must be taken when applying these methods to mixed-type data. There are numerous other considerations that are important to these methods, such as the choice of kernel function, but these are beyond the scope of this review.

Conclusions

This review highlights a number of issues in previous applications of cluster analysis to multimodal clinical data in asthma. We make the following key recommendations based on these findings:

- Careful consideration should be given to the preprocessing of multimodal clinical data and how the scaling and encoding of features may affect their weighting in the analysis.
- The choice of dissimilarity measures and cluster analysis methods are dependent on one another as well as on the scaling and encoding of the data. Certain combinations of these data analytics components may be incompatible and give unreliable results.
- The stability and quality of the cluster results should be thoroughly evaluated.

The abovementioned recommendations focus on the application of cluster analysis, but we put similar emphasis on the clear reporting of each of the abovementioned points, as this was also found to be lacking in the reviewed papers.

Acknowledgments

This study was supported by the Health Data Research, United Kingdom (HDR UK), which receives funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), the British Heart Foundation, and the Wellcome Trust and by the Asthma UK Centre for Applied Research, which is funded by Asthma UK. The funders had no role in the study or the decision to submit this work to be considered for publication.

Authors' Contributions

EH was responsible for conducting the study. EH conducted the identification of articles and screened them for eligibility. EH and HT independently extracted data according to the described methodology and synthesized the findings. EH wrote up the first draft of the manuscript, and AT, AS, and HT contributed to the final version.

Conflicts of Interest

AS is supported by a research grant from the Asthma UK Centre for Applied Research. All other authors have no conflict of interest pertaining to this study to declare.

Multimedia Appendix 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist.
[DOC File , 64 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Data dictionary.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Study characteristics.
[DOCX File , 85 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Breakdown of methods used by the 11 studies that did not use the three most common clustering methods.
[DOCX File , 16 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Illustrative example of the use of Gower coefficient with hierarchical clustering and Ward linkage.
[DOCX File , 12 KB-Multimedia Appendix 5]

References

1. Lawton M, Ben-Shlomo Y, May MT, Baig F, Barber TR, Klein JC, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. J Neurol Neurosurg Psychiatry 2018 Dec;89(12):1279-1287 [FREE Full text] [doi: 10.1136/jnnp-2018-318337] [Medline: 30464029]

2. Ousley O, Cermak T. Autism spectrum disorder: defining dimensions and subgroups. Curr Dev Disord Rep 2014 Mar 1;1(1):20-28 [FREE Full text] [doi: 10.1007/s40474-013-0003-1] [Medline: 25072016]
3. Li D, Haritunians T, Landers C, Potdar AA, Yang S, Huang H, et al. Late-onset Crohn's disease is a subgroup distinct in genetic and behavioral risk factors with UC-like characteristics. Inflamm Bowel Dis 2018 Oct 12;24(11):2413-2422 [FREE Full text] [doi: 10.1093/ibd/izy148] [Medline: 29860388]
4. Bowman P, Flanagan SE, Hattersley AT. Future roadmaps for precision medicine applied to diabetes: rising to the challenge of heterogeneity. J Diabetes Res 2018;2018:3061620 [FREE Full text] [doi: 10.1155/2018/3061620] [Medline: 30599002]
5. Sidhaye VK, Nishida K, Martinez FJ. Precision medicine in COPD: where are we and where do we need to go? Eur Respir Rev 2018 Sep 30;27(149) [FREE Full text] [doi: 10.1183/16000617.0022-2018] [Medline: 30068688]
6. Zhang J, Späth SS, Marjani SL, Zhang W, Pan X. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med 2018 Jun;1(1):29-48 [FREE Full text] [doi: 10.1093/pcmedi/pby007] [Medline: 30687561]
7. Pavord ID, Beasley R, Agusti A, Anderson GP, Bel E, Brusselle G, et al. After asthma: redefining airways diseases. Lancet 2018 Jan 27;391(10118):350-400. [doi: 10.1016/S0140-6736(17)30879-6] [Medline: 28911920]
8. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
9. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1736-1788 [FREE Full text] [doi: 10.1016/S0140-6736(18)32203-7] [Medline: 30496103]
10. Rackemann FM. A working classification of asthma. Am J Med 1947 Nov;3(5):601-606. [doi: 10.1016/0002-9343(47)90204-0] [Medline: 20269240]
11. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. Pulm Ther 2016;2:19-41 [FREE Full text] [doi: 10.1007/s41030-016-0017-z] [Medline: 27512723]
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, USA: Springer; 2009.
13. Hennig C, Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J R Stat Soc Ser C Appl Stat 2013;62(3):309-369. [doi: 10.1111/j.1467-9876.2012.01066.x]
14. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med 2008 Aug 1;178(3):218-224 [FREE Full text] [doi: 10.1164/rccm.200711-1754OC] [Medline: 18480428]
15. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. Curr Allergy Asthma Rep 2015 Jul;15(7):38 [FREE Full text] [doi: 10.1007/s11882-015-0542-0] [Medline: 26143394]
16. Hastie T, Tibshirani R, Friedman J. Unsupervised learning. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009:485-552.
17. Dolnicar S, Grün B, Leisch F, Schmidt K. Required sample sizes for data-driven market segmentation analyses in tourism. J Travel Res 2014;53(3):296-306. [doi: 10.1177/0047287513496475]
18. Hastie T, Tibshirani R, Friedman J. Overview of supervised learning. In: Hastie T, Tibshirani R, Friedman J, editors. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (chapter 2). New York: Springer; 2009:9-38.
19. Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis. Methods Mol Biol 2003;224:159-182. [doi: 10.1385/1-59259-364-X:159] [Medline: 12710673]
20. van der Maaten L, Postma E, van den Herik J. Dimensionality reduction: a comparative review. J Mach Learn Res 2009;10:66-71 [FREE Full text]
21. Jackson E. A User's Guide to Principal Components. Jersey City, USA: Wiley-Blackwell; 1991.
22. Pagès J. Multiple correspondence analysis. In: Multiple factor Analysis by Example using R. Boca Raton, Florida: Chapman and Hall/CRC; 2018:39-66.
23. Pagès J. Multiple factor analysis and procrustes analysis. In: Multiple factor Analysis by Example using R. Boca Raton, Florida: Chapman and Hall/CRC; 2018:189-208.
24. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971 Dec;27(4):857-871. [doi: 10.2307/2528823]
25. Everitt B, Landau S, Leese M. Cluster Analysis. Fifth Edition. New York, USA: Wiley Publishing; 2011.
26. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987 Nov;20(5):53-65. [doi: 10.1016/0377-0427(87)90125-7]
27. von Luxburg U. Clustering stability: an overview. Found Trends Mach Learn 2010;2(3):235-274. [doi: 10.1561/2200000008]
28. Schatz M, Hsu JY, Zeiger RS, Chen W, Dorenbaum A, Chipps BE, et al. Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. J Allergy Clin Immunol 2014 Jun;133(6):1549-1556. [doi: 10.1016/j.jaci.2013.10.006] [Medline: 24315502]

29. Prosperi MC, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, Simpson A, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med 2013 Dec 1;188(11):1303-1312 [FREE Full text] [doi: 10.1164/rccm.201304-0694OC] [Medline: 24180417]
30. Lefaudeux D, de Meulder B, Loza MJ, Peffer N, Rowe A, Baribaud F, U-BIOPRED Study Group. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. J Allergy Clin Immunol 2017 Jun;139(6):1797-1807. [doi: 10.1016/j.jaci.2016.08.048] [Medline: 27773852]
31. Khusial RJ, Sont JK, Loijmans RJ, Snoeck-Stroband JB, Assendelft PJ, Schermer TR, ACCURATE Study Group. Longitudinal outcomes of different asthma phenotypes in primary care, an observational study. NPJ Prim Care Respir Med 2017 Oct 3;27(1):55 [FREE Full text] [doi: 10.1038/s41533-017-0057-3] [Medline: 28974677]
32. Hsiao H, Lin M, Wu C, Wang C, Wang T. Sex-specific asthma phenotypes, inflammatory patterns, and asthma control in a cluster analysis. J Allergy Clin Immunol Pract 2019 Feb;7(2):556-67.e15. [doi: 10.1016/j.jaip.2018.08.008] [Medline: 30170162]
33. Moore WC, Hastie AT, Li X, Li H, Busse WW, Jarjour NN, National Heart, Lung, Blood Institute's Severe Asthma Research Program. Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. J Allergy Clin Immunol 2014 Jun;133(6):1557-63.e5 [FREE Full text] [doi: 10.1016/j.jaci.2013.10.011] [Medline: 24332216]
34. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. Eur Respir J 2012 Jul;40(1):55-60 [FREE Full text] [doi: 10.1183/09031936.00123411] [Medline: 22267763]
35. Pérez-Losada M, Authelet KJ, Hoptay CE, Kwak C, Crandall KA, Freishtat RJ. Pediatric asthma comprises different phenotypic clusters with unique nasal microbiotas. Microbiome 2018 Oct 4;6(1):179 [FREE Full text] [doi: 10.1186/s40168-018-0564-7] [Medline: 30286807]
36. Ding L, Li D, Wathen M, Altaye M, Mersha TB. African ancestry is associated with cluster-based childhood asthma subphenotypes. BMC Med Genomics 2018 May 31;11(1):51 [FREE Full text] [doi: 10.1186/s12920-018-0367-5] [Medline: 29855310]
37. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA, Childhood Asthma Management Program Research Group. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. J Allergy Clin Immunol 2014 May;133(5):1289-300, 1300.e1 [FREE Full text] [doi: 10.1016/j.jaci.2014.02.006] [Medline: 24892144]
38. Loureiro CC, Sa-Couto P, Todo-Bom A, Bousquet J. Cluster analysis in phenotyping a Portuguese population. Rev Port Pneumol (2006) 2015 Sep 3 [Online ahead of print]. [doi: 10.1016/j.rppnen.2015.07.006] [Medline: 26344641]
39. Amore M, Antonucci C, Bettini E, Boracchia L, Innamorati M, Montali A, et al. Disease control in patients with asthma is associated with alexithymia but not with depression or anxiety. Behav Med 2013;39(4):138-145. [doi: 10.1080/08964289.2013.818931] [Medline: 24236811]
40. Cabral AL, Sousa AW, Mendes FA, Carvalho CR. Phenotypes of asthma in low-income children and adolescents: cluster analysis. J Bras Pneumol 2017;43(1):44-50 [FREE Full text] [doi: 10.1590/S1806-37562016000000039] [Medline: 28125150]
41. Wang L, Liang R, Zhou T, Zheng J, Liang BM, Zhang HP, et al. Identification and validation of asthma phenotypes in Chinese population using cluster analysis. Ann Allergy Asthma Immunol 2017 Oct;119(4):324-332. [doi: 10.1016/j.anai.2017.07.016] [Medline: 28866310]
42. Sakagami T, Hasegawa T, Koya T, Furukawa T, Kawakami H, Kimura Y, et al. Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline. J Asthma 2014 Mar;51(2):113-118. [doi: 10.3109/02770903.2013.852201] [Medline: 24102534]
43. Seino Y, Hasegawa T, Koya T, Sakagami T, Mashima I, Shimizu N, Niigata Respiratory Disease Study Group. A cluster analysis of bronchial asthma patients with depressive symptoms. Intern Med 2018 Jul 15;57(14):1967-1975 [FREE Full text] [doi: 10.2169/internalmedicine.9073-17] [Medline: 29526967]
44. Agache I, Strasser DS, Klenk A, Agache C, Farine H, Ciobanu C, et al. Serum IL-5 and IL-13 consistently serve as the best predictors for the blood eosinophilia phenotype in adult asthmatics. Allergy 2016 Aug;71(8):1192-1202. [doi: 10.1111/all.12906] [Medline: 27060452]
45. Newby C, Heaney LG, Menzies-Gow A, Niven RM, Mansur A, Bucknall C, British Thoracic Society Severe Refractory Asthma Network. Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability. PLoS One 2014;9(7):e102987 [FREE Full text] [doi: 10.1371/journal.pone.0102987] [Medline: 25058007]
46. Kim MA, Shin SW, Park JS, Uh ST, Chang HS, Bae DJ, et al. Clinical characteristics of exacerbation-prone adult asthmatics identified by cluster analysis. Allergy Asthma Immunol Res 2017 Nov;9(6):483-490 [FREE Full text] [doi: 10.4168/aair.2017.9.6.483] [Medline: 28913987]
47. Sendín-Hernández MP, Ávila-Zarza C, Sanz C, García-Sánchez A, Marcos-Vadillo E, Muñoz-Bellido FJ, et al. Cluster analysis identifies 3 phenotypes within allergic asthma. J Allergy Clin Immunol Pract 2018;6(3):955-61.e1. [doi: 10.1016/j.jaip.2017.10.006] [Medline: 29133218]
48. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]

49. Buttrey SE, Whitaker LR. treeClust: an R package for tree-based clustering dissimilarities. R J 2015;7(2):227-236 [FREE Full text] [doi: 10.32614/rj-2015-032]
50. Meyer N, Nuss SJ, Rothe T, Siebenhüner A, Akdis CA, Menz G. Differential serum protein markers and the clinical severity of asthma. J Asthma Allergy 2014;7:67-75 [FREE Full text] [doi: 10.2147/JAA.S53920] [Medline: 24851055]
51. Zhang T, Ramakrishnan R, Livny M. BIRCH: a new data clustering algorithm and its applications. Data Min Knowl Discov 1996;25(2):141-182. [doi: 10.1145/233269.233324]
52. Bacher J, Wenzig K, Vogler M. SPSS TwoStep Cluster - a first evaluation. Soc Sci Open Access Repos 2004 [FREE Full text]
53. Ruggieri S, Drago G, Longo V, Colombo P, Balzan M, Bilocca D, RESPIRA Project Group. Sensitization to dust mite defines different phenotypes of asthma: a multicenter study. Pediatr Allergy Immunol 2017 Nov;28(7):675-682. [doi: 10.1111/pai.12768] [Medline: 28783215]
54. Ye W, Xu W, Guo X, Han F, Peng J, Li X, et al. Differences in airway remodeling and airway inflammation among moderate-severe asthma clinical phenotypes. J Thorac Dis 2017 Sep;9(9):2904-2914 [FREE Full text] [doi: 10.21037/jtd.2017.08.01] [Medline: 29221262]
55. Kaufman L, Rousseeuw PJ. Partition Around Medoids (Program PAM). In: Kaufman L, Rousseeuw PJ, editors. Finding Groups in Data: An Introduction to Cluster Analysis. NJ: John Wiley & Sons; 2005.
56. Sekiya K, Nakatani E, Fukutomi Y, Kaneda H, Iikura M, Yoshida M, et al. Severe or life-threatening asthma exacerbation: patient heterogeneity identified by cluster analysis. Clin Exp Allergy 2016 Aug;46(8):1043-1055. [doi: 10.1111/cea.12738] [Medline: 27041475]
57. Loza MJ, Djukanovic R, Chung KF, Horowitz D, Ma K, Branigan P, ADEPT (Airways Disease Endotyping for Personalized Therapeutics), U-BIOPRED (Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium) investigators. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respir Res 2016 Dec 15;17(1):165 [FREE Full text] [doi: 10.1186/s12931-016-0482-9] [Medline: 27978840]
58. Dhillon IS, Guan Y, Kulis B. Kernel K-Means: Spectral Clustering and Normalized Cuts. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004 Presented at: KDD'04; August 22 - 25, 2004; Seattle, WA, USA p. 551-556. [doi: 10.1145/1014052.1014118]
59. Wu W, Bang S, Bleecker ER, Castro M, Denlinger L, Erzurum SC, et al. Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma. Am J Respir Crit Care Med 2019 Jun 1;199(11):1358-1367. [doi: 10.1164/rccm.201808-1543OC] [Medline: 30682261]
60. Gomez JL, Yan X, Holm CT, Grant N, Liu Q, Cohn L, SARP Investigators. Characterisation of asthma subgroups associated with circulating YKL-40 levels. Eur Respir J 2017 Oct;50(4) [FREE Full text] [doi: 10.1183/13993003.00800-2017] [Medline: 29025889]
61. Amelink M, de Nijs SB, de Groot JC, van Tilburg PM, van Spiegel PI, Krouwels FH, et al. Three phenotypes of adult-onset asthma. Allergy 2013;68(5):674-680. [doi: 10.1111/all.12136] [Medline: 23590217]
62. Benton AS, Wang Z, Lerner J, Foerster M, Teach SJ, Freishtat RJ. Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort. J Asthma 2010 Sep;47(7):728-734 [FREE Full text] [doi: 10.3109/02770903.2010.491142] [Medline: 20684733]
63. Lemiere C, NGuyen S, Sava F, D'Alpaos V, Huaux F, Vandenplas O. Occupational asthma phenotypes identified by increased fractional exhaled nitric oxide after exposure to causal agents. J Allergy Clin Immunol 2014 Nov;134(5):1063-1067. [doi: 10.1016/j.jaci.2014.08.017] [Medline: 25262466]
64. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB. Reproducible clusters from microarray research: whither? BMC Bioinformatics 2005 Jul 15;6(Suppl 2):S10 [FREE Full text] [doi: 10.1186/1471-2105-6-S2-S10] [Medline: 16026595]
65. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004 Apr;2(4):E108 [FREE Full text] [doi: 10.1371/journal.pbio.0020108] [Medline: 15094809]
66. Bair E. Semi-supervised clustering methods. Wiley Interdiscip Rev Comput Stat 2013;5(5):349-361 [FREE Full text] [doi: 10.1002/wics.1270] [Medline: 24729830]
67. Just J, Gouvis-Echraghi R, Couderc R, Guillemot-Lambert N, Saint-Pierre P. Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze. J Allergy Clin Immunol 2012 Jul;130(1):103-10.e8. [doi: 10.1016/j.jaci.2012.02.041] [Medline: 22502798]
68. Just J, Saint-Pierre P, Gouvis-Echraghi R, Boutin B, Panayotopoulos V, Chebahi N, et al. Wheeze phenotypes in young children have different courses during the preschool period. Ann Allergy Asthma Immunol 2013 Oct;111(4):256-61.e1. [doi: 10.1016/j.anai.2013.07.002] [Medline: 24054360]
69. Just J, Saint-Pierre P, Gouvis-Echraghi R, Laoudi Y, Roufai L, Momas I, et al. Childhood allergic asthma is not a single phenotype. J Pediatr 2014 Apr;164(4):815-820. [doi: 10.1016/j.jpeds.2013.11.037] [Medline: 24412137]
70. Zoratti EM, Krouse RZ, Babineau DC, Pongracic JA, O'Connor GT, Wood RA, et al. Asthma phenotypes in inner-city children. J Allergy Clin Immunol 2016 Oct;138(4):1016-1029 [FREE Full text] [doi: 10.1016/j.jaci.2016.06.061] [Medline: 27720016]
71. Kaufman L, Rousseeuw PJ. Fuzzy Analysis (Program FANNY). In: Kaufman L, Rousseeuw PJ, editors. Finding Groups in Data: An Introduction to Cluster Analysis. NJ: John Wiley & Sons; 2005:164-198.

72. Labor M, Labor S, Jurić I, Fijačko V, Grle SP, Plavec D. Mood disorders in adult asthma phenotypes. J Asthma 2018 Jan;55(1):57-65. [doi: 10.1080/02770903.2017.1306546] [Medline: 28489959]
73. Obara T, Ishikuro M, Tamiya G, Ueki M, Yamanaka C, Mizuno S, et al. Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. Sci Rep 2018 Oct 4;8(1):14840 [FREE Full text] [doi: 10.1038/s41598-018-33110-w] [Medline: 30287864]
74. Cook JD, Rumble ME, Plante DT. Identifying subtypes of Hypersomnolence Disorder: a clustering analysis. Sleep Med 2019 Dec;64:71-76. [doi: 10.1016/j.sleep.2019.06.015] [Medline: 31670163]
75. Wolters AF, Moonen AJ, Lopes R, Leentjens AF, Duits AA, Defebvre L, et al. Grey matter abnormalities are associated only with severe cognitive decline in early stages of Parkinson's disease. Cortex 2020 Feb;123:1-11. [doi: 10.1016/j.cortex.2019.09.015] [Medline: 31733342]
76. Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Med Inform Decis Mak 2019 Apr 18;19(1):86 [FREE Full text] [doi: 10.1186/s12911-019-0805-0] [Medline: 30999919]
77. Bang S, Yu Y, Wu W. Robust multiple kernel k-means clustering using min-max optimization. arXiv preprints 2018 preprint; arXiv:1803.02458 [FREE Full text]
78. Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an Algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001 Presented at: NIPS'01; December 3-8, 2001; Vancouver, Canada p. 849-856 URL: https://dl.acm.org/doi/10.5555/2980539.2980649

Abbreviations

CCC: cubic cluster criterion
HDR: Health Data Research
MCA: multiple correspondence analysis
PCA: principal component analysis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 30.09.19; peer-reviewed by M Pikoula, C Newby, K Usop; comments to author 11.11.19; revised version received 10.12.19; accepted 10.02.20; published 28.05.20

Please cite as:
Horne E, Tibble H, Sheikh A, Tsanas A
Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping
JMIR Med Inform 2020;8(5):e16452
URL: http://medinform.jmir.org/2020/5/e16452/
doi: 10.2196/16452
PMID: 32463370