Received: 1 March 2021 | Revised: 16 November 2021 | Accepted: 27 November 2021

DOI: 10.1002/widm.1444

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

Validation of cluster analysis results on validation data: A systematic framework

Theresa Ullmann[1] | **Christian Hennig**[2] | **Anne-Laure Boulesteix**[1]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany

[2]Dipartimento di Scienze Statistiche "Paolo Fortunati", Universita di Bologna, Bologna, Italy

Correspondence
Theresa Ullmann, Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Marchioninistraße 15, 81377, Munich, Germany.
Email: tullmann@ibe.med.uni-muenchen.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01IS18036A; Deutsche Forschungsgemeinschaft, Grant/Award Number: BO3139/7-1

Edited by: Witold Pedrycz, Editor-in-Chief

Abstract

Cluster analysis refers to a wide range of data analytic techniques for class discovery and is popular in many application fields. To assess the quality of a clustering result, different cluster validation procedures have been proposed in the literature. While there is extensive work on classical validation techniques, such as internal and external validation, less attention has been given to validating and replicating a clustering result using a validation dataset. Such a dataset may be part of the original dataset, which is separated before analysis begins, or it could be an independently collected dataset. We present a systematic, structured review of the existing literature about this topic. For this purpose, we outline a formal framework that covers most existing approaches for validating clustering results on validation data. In particular, we review classical validation techniques such as internal and external validation, stability analysis, and visual validation, and show how they can be interpreted in terms of our framework. We define and formalize different types of validation of clustering results on a validation dataset, and give examples of how clustering studies from the applied literature that used a validation dataset can be seen as instances of our framework.

This article is categorized under:
Technologies > Structure Discovery and Clustering
Algorithmic Development > Statistics
Technologies > Machine Learning

KEYWORDS
cluster stability, cluster validation, clustering, independent data, replication

1 | INTRODUCTION

Cluster analysis refers to data analytic techniques for structure and class discovery. It is popular in a range of fields, for example, medicine, biology, market research, social science, and data compression. However, when conducting cluster analysis, researchers are confronted with an overwhelming number of existing methods. They must preprocess the data, choose a clustering algorithm, and set parameters, such as the number of clusters (Van Mechelen et al., 2018;

Zimmermann, 2020). It is often unclear a priori which choice should be made for the analysis, and even once a choice is made, it may remain unclear how good the quality of the resulting clustering is.

These problems have prompted the development of so-called *cluster validation* techniques, see Handl et al. (2005) and Hennig (2015a) for overviews. The literature distinguishes between internal validation (where the clustering is evaluated based on internal properties, such as compactness and separateness of the clusters) and external validation (where the clustering is evaluated by comparing the clusters with respect to one or more variables not used for clustering, e.g., a survival time or a true class membership). Less attention has been given to the validation and replication of clustering results on a *validation dataset*, for which we introduce a structured framework that summarizes the existing literature in a systematic manner. A validation dataset could be part of the original dataset, set apart before the start of the analysis, or it could be a separate dataset, obtained, for example, from a different study centre.

The idea of validating a clustering on another dataset is not new and has appeared in the methodological literature decades ago (Breckenridge, 1989; McIntyre & Blashfield, 1980). In applied literature involving cluster analysis, it is not uncommon for authors to validate their clustering results on new data, be it with the procedure of McIntyre and Blashfield (1980) or another method. To the best of our knowledge, these approaches have never been systematically structured and evaluated, and different validation strategies are scattered across different works and application fields. This contrasts with the abundant methodological literature devoted to validation in the context of *supervised* classification (or more generally, supervised learning). This contrast may be partly due to the fact that cluster analysis—as opposed to supervised classification—is often viewed as exploratory research. The validation of clustering results is rightly considered to be less straightforward than the validation of a prediction model because "true labels" are unknown (Von Luxburg et al., 2012). Indeed, it is difficult to define exactly what is meant by validating a clustering on validation data. Answering this question is the key aspect of our framework.

In this article, we aim to give a systematic review of the various strategies used in the literature for validating clustering results on validation data. These existing approaches are combined into a structured framework. In this framework, we define and formalize the concept of validation on a validation dataset. In particular, we demonstrate that many classical validation techniques, such as internal and external validation, stability analysis, and visual validation, can be linked to evaluation on validation data: using validation datasets does not replace these approaches; rather, classical validation can be combined with validation data. Moreover, we show how clustering studies from the applied literature that used a validation dataset can be classified into our framework.

Why do researchers consider validation and replication of clustering results on a validation dataset to be important? The answer is closely tied to the clustering aim, which could either be *inferential* or *descriptive*. We define these terms as follows:

- *Inferential clustering*: The objects being clustered form a sample drawn from an underlying population for which inference is of interest, rather than making statements about the specific objects in the original dataset.
- *Descriptive clustering*: The data form a fixed set of entities of specific interest, and statements such as objects 1, 5, and 99 form a cluster are of interest.

As an example of the difference between inferential and descriptive clustering, consider an $n \times p$ dataset including the expression levels (continuous values) of p genes for n patients suffering from a particular disease, see Figure 1.

On the one hand, it may be of interest to perform clustering analyses of the patients to see if there are subpopulations of patients with systematically different gene expressions. This would be *inferential clustering*. For example, researchers have frequently used gene expression data to detect distinct breast cancer subtypes (Burstein et al., 2015; Curtis et al., 2012; Kapp et al., 2006; Lehmann et al., 2011; Sørlie et al., 2003; Sotiriou et al., 2003). Such subtypes can have clinical implications and may guide targeted treatment (Garrido-Castro et al., 2019; Prat et al., 2015). On the other hand, an $n \times p$ gene expression dataset could also be used to perform clustering of the (fixed set of) p genes to see if there are groups of specific genes that behave similarly, which might suggest a similar function or involvement in a common molecular process. This is an example of *descriptive clustering*. For example, researchers have used cluster analysis to find different groups of cancer-related genes (Freudenberg et al., 2009; Yang et al., 2014; Zhang et al., 2014).

For both clustering aims, using validation data is of crucial importance. To illustrate this, we again use the gene expression example, see Figure 1. First, we consider inferential clustering of breast cancer patients. All the papers for breast cancer subtype detection cited above used validation datasets to confirm the results of their own analyses and/or to validate previously reported subtypes. Indeed, a clustering of cancer patients would not be of much use if it only held on a single dataset. Due to the inferential nature of the clustering, the researchers' aim is to better understand the

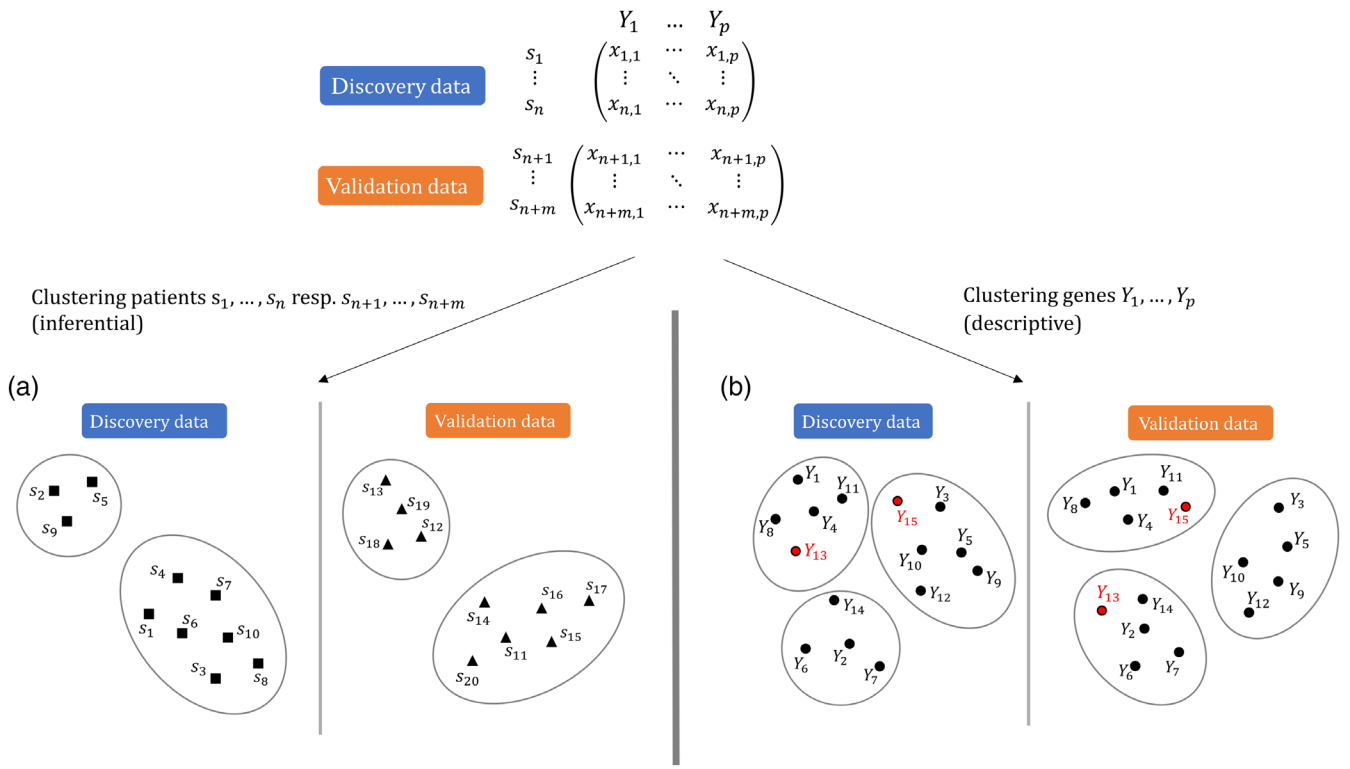

FIGURE 1 Schematic representation of clustering on a gene expression dataset. (a) Inferential clustering of the patients and (b) descriptive clustering of the genes. For illustration purposes, there are 10 patients $s_1,\dots,s_{10}$ in the discovery data, 10 patients $s_{11},\dots,s_{20}$ in the validation data, and 15 genes $Y_1,\dots,Y_{15}$. For inferential clustering, the objects to cluster (here, patients) are different between discovery and validation data, as indicated by using different symbols (squares vs. triangles). The two resulting clusterings nevertheless look somewhat similar: a smaller cluster on the top left, and a larger cluster on the bottom right. For descriptive clustering, the objects to cluster (here, genes $Y_1,\dots,Y_{15}$, marked by circles) remain the same across both datasets. However, their positions are slightly shifted in the validation data, because the gene expression values now stem from patients $s_{11},\dots,s_{20}$. Consequently, genes Y_{13} and Y_{15} (marked in red) are clustered differently on the validation data

disease and to find options for treatment with respect to the underlying population, for example, the population of *all* breast cancer patients. In particular, the clustering should not only hold for patients from a single hospital or a single country. To make sure that the clustering is not just an artifact of a single dataset, researchers thus use independently collected samples for validation, or at least split their dataset into discovery and validation sets.

Now consider the example of the descriptive clustering of cancer-related genes. While the set of genes is fixed, researchers typically want the gene clustering to hold more generally than only for the n specific patients. The genes' functions or involvement in molecular processes should reflect biological principles that hold for all patients with the particular cancer type, and researchers thus want to recover the clustering on datasets with other patients having the same disease. Again, validation datasets are used for this purpose. In this sense, descriptive clustering can have an inferential component, with the difference to "inferential clustering" (as defined above) being that the objects to be clustered are fixed and do not represent samples drawn from an underlying population. This has implications for choosing a suitable validation dataset and validation strategy, as will be discussed in more detail in Section 3 below.

Similar arguments about the importance of replicability and generalisability of clusterings results (as given for the example of gene expression data above) hold more generally for most cluster analysis applications, which can typically be classified as either inferential or descriptive clustering. For example, in market segmentation (inferential clustering of customers), the resulting clusters should be replicable such that managers can consistently market their products to the customer groups (Dolnicar & Leisch, 2010; Müller & Hamm, 2014). In text and keyword analysis, where words are clustered to reveal overarching topics (descriptive clustering), it is interesting to see whether topics stay stable on validation data, or whether some changes appear (Ding et al., 2001). Across different application fields, researchers usually want their results to be as generalizable as possible. Interesting properties of a clustering result should hold not only for a single specific dataset, but should also reappear when clustering validation data sampled from the same, or even

different distributions. Validating clusterings on validation data also enables researchers to evaluate results reported by other research teams. The confirmation of results on validation data is a vital part of research in general, and it has received considerable attention in recent years due to the so-called "replication crisis" (Hutson, 2018). In the context of classical hypothesis tests and effect estimates, many published results have turned out to be non-replicable, that is, they could not be confirmed on independent data [e.g., in psychology (Open Science Collaboration, 2015), cancer research (Begley & Ellis, 2012), or economics (Camerer et al., 2016)]. Replication is thus vital for assessing the credibility of scientific claims (Nosek & Errington, 2020). For cluster analysis, our article appears to be the first one to systematically review and discuss this topic.

Our framework, which is described in detail in Section 3, is based on the following two-step cluster analysis procedure (see also Figure 2):

1. *The primary cluster analysis and method selection step*: Using the original dataset or a part of it (in the following called "discovery data") a single clustering method is selected (where the "method" includes not only the choice of clustering algorithm, but also parameters such as the number of clusters and diverse pre/postprocessing steps), for example, via its performance with respect to internal/external validation indices.
2. *The validation step*: Important aspects of the clustering resulting from this method are validated on another dataset or the rest of the original dataset (in the following denoted by "validation data"). The validation data should be completely hidden from the method selection process of Step 1—analogously to the evaluation of supervised classifiers, where the selected model (including the chosen parameters) must be finally evaluated using validation data that was *not* used in any way for parameter tuning or model selection (Boulesteix et al., 2008; Simon et al., 2003).

The "important aspects" of the clustering that are checked in Step 2 usually depend on the research question and the field of application. Consider again the above example of clustering cancer patients, based on expression levels of cancer-related genes, for the purpose of finding subtypes of that disease. In this context, the following properties might be relevant aspects of the clustering:

- Suppose that Step 1 has resulted in two clusters. One cluster is much larger than the other, with about 80% percent of the patients in this cluster. One might be interested in whether this pattern of one large cluster and one smaller cluster can be replicated in Step 2.
- Assume it is found that the clustering chosen in Step 1 is related to survival time, that is, the patients' survival times differ depending on which cluster they belong to. Can this finding be replicated in Step 2 for patients in the validation data?

In the literature, the term "cluster validation" is sometimes used to refer to the use of validation techniques as a tool to compare different clusterings and select the most appropriate. This use of terminology would place validation within Step 1. But when validation techniques are used as selection tool, it is still an open issue whether the results generalize to new data, and this is addressed by Step 2.

The phrase "cluster validation" also appears in the literature about *benchmarking* of clustering methods (Boulesteix & Hatz, 2017; Van Mechelen et al., 2018; Zimmermann, 2020). A benchmarking study is a systematic comparison of different clustering *methods* on a class of data distributions or datasets. Validation techniques may be used to compare different methods. Benchmark studies thus analyze the "validity" of clustering methods and provide general guidance on which method to use. In contrast, our review considers the validation of specific results of applied clustering studies.

This article is structured as follows: in Section 2, we give an overview of the different uses of the term "validation" and perspectives on validity found in the clustering literature. We then present our validation framework in detail in Section 3. In Section 4, we demonstrate in an exemplary manner how clustering studies from the applied literature can

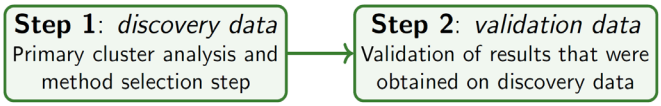

FIGURE 2 Two-step procedure for validating clustering results

be sorted into the framework. Section 5 contains a final discussion. In the Supporting Information, we present an illustration of the discussed validation strategies using openly available real-world data, where the data analysis is performed with thoroughly commented R code.

2 | DIFFERENT PERSPECTIVES ON "VALIDITY" IN CLUSTER ANALYSIS

We identified four approaches that address the validity of clusterings in the literature: (1) the comparison of "true" cluster labels with inferred clusters, (2) internal and external validity indices, (3) stability analyses, and (4) visual validation. These four approaches are briefly reviewed in the following subsections. An additional approach, hypothesis testing, is briefly discussed in Section 5. Internal and external validation, stability, and visual validation form the building blocks of our framework, see Section 3.

2.1 | Recovery of "true" clusters and analogies to the validity of supervised classification models

According to this perspective, a clustering of a dataset is "valid" if it corresponds to the "true" cluster structure in the data. Correspondingly, a clustering method is called "valid" if it can recover the "true" clusters in the data (Breckenridge, 1989; Milligan & Cooper, 1987). A related view is presented in the paper of Dougherty et al. (2007), which shows a connection to the term "validity" in the context of supervised classification. For supervised classification models, the validation of a classifier relates to estimating the *prediction error* on a test set, that is, how well the classifier can predict the known "true" labels of the instances in the test set. Dougherty et al. (2007) demonstrate that this approach can be transferred to cluster analysis. However, this requires datasets with *known* cluster labels. Yet, in practice, cluster analysis is usually applied to real datasets for which the "true" cluster labels of the data points are unknown. Note that even in the rare case of a cluster analysis performed on a dataset with given "true" cluster labels, these may not be unique, and there might be other equally legitimate cluster structures in the data, which can be even more interesting and useful as a result of the analysis than the one previously know (see Färber et al., 2010; Hennig, 2015b). When validating a clustering on validation data, the validation step used in supervised classification usually cannot be mimicked. The idea of Dougherty et al. (2007) thus mainly makes sense in the context of benchmark studies comparing clustering methods using simulated data with known "true" cluster labels. The ability of the methods to recover the true clusters may then be used as a performance criterion. To evaluate clusterings in applied studies, other options for validation are needed.

2.2 | Internal and external validation

In the absence of "true" cluster labels, assessing "cluster validity" often uses so-called internal indices or external information—leading to the terms "internal validation" and "external validation," respectively.

- *Internal validation* uses only the data that was used for clustering. Typically, internal validation consists of calculating an index that is supposed to measure how well the clustering fits the data (Halkidi et al., 2015). Such indices often exploit the proximity structure of the data, for example, by measuring the homogeneity and/or the separation of the clusters. Examples are the Average Silhouette Width index (Kaufman & Rousseeuw, 2009) and the Caliński–Harabasz index (Caliński & Harabasz, 1974). These indices combine measurements of the homogeneity and the separation of a clustering into a single value, in order to balance a small within-cluster heterogeneity and a large between-clusters heterogeneity. There are also indices that measure only isolated aspects of a clustering (e.g., only the homogeneity or only the separation of the clusters), see Akhanli and Hennig (2020).
- *External validation* makes use of additional (external) information that was *not* used for clustering. For example, when clustering a cancer gene expression dataset, one may use the survival time of patients to determine whether the clustering of patients based on gene expression can predict survival. The term "external validation" also encompasses the recovery of previously known "true labels" as presented in Section 2.1.

2.3 | Stability

Many authors consider *stability* to be a crucial aspect of cluster validity. The idea is that a good clustering method should yield similar partitions when applied to multiple datasets drawn from the same data distribution (Ben-David et al., 2006; Von Luxburg, 2010). In this spirit, a specific clustering of a single real dataset may be considered as validated if the clusterings obtained from datasets generated from the same data distribution are similar. There are several methods of generating multiple datasets to emulate the data distribution of the dataset to be analyzed, for example, by drawing subsamples from the original dataset (Hennig, 2007).

Stability analysis dates back to McIntyre and Blashfield (1980), Morey et al. (1983), and Breckenridge (1989). These authors considered the replicability of a clustering result on a validation dataset. To generate the validation dataset, the original data is split into two halves (by splitting along the objects to be clustered for inferential clustering, or by splitting across the variables of the dataset for descriptive clustering). This is followed by assessing whether the clustering obtained in the first half can be replicated in the second half. For descriptive clustering, because the objects in the two halves are the same, replicability can be assessed directly with a partition similarity index such as the Adjusted Rand Index (ARI; Hubert & Arabie, 1985; Rand, 1971), the Jaccard index (Jaccard, 1908), or the FM index (Fowlkes & Mallows, 1983). See Meila (2015) and Albatineh et al. (2006) for overviews of partition similarity indices. For inferential clustering, the objects to cluster are not the same in the two data halves, and thus the objects from the second half have to be classified into the clusters of the first half, before the clusterings can be compared with a partition similarity index (see Section 3.3 for details). Such stability analyses will indeed be a special case of the broader validation framework presented in Section 3.

In the decades that followed, however, the focus of stability analysis shifted away from this concept and more towards *method or model selection*. Like other validation techniques, stability analyses are used in Step 1 (see Figure 2) as a basis for the selection of a suitable clustering method and its parameters, such as the number of clusters (Ben-Hur et al., 2002; Bertrand & Mufti, 2006; Dolnicar & Leisch, 2010; Dudoit & Fridlyand, 2002; Fang & Wang, 2012; Fu & Perry, 2020; Lange et al., 2004; Levine & Domany, 2001; Monti et al., 2003; Tibshirani & Walther, 2005; Wang, 2010). In these approaches, stability analysis selects the clustering method that is most stable over multiple subsamples. The subsamples are drawn without replacement or in a cross-validation manner, or are bootstrap samples drawn with replacement from the data. For example, different numbers of clusters k can be considered in turn, and the k that leads to the most stable clustering, or the smallest k that exceeds a stability threshold, can be chosen. These studies typically consider inferential clustering, such that the term "subsamples" refers to subsets of objects to be clustered. Some schemes require the comparison of clusterings on subsets of objects that consist of disjunct subsamples of the original dataset and thus have no overlap (e.g., Dudoit & Fridlyand, 2002; Fang & Wang, 2012; Lange et al., 2004; Tibshirani & Walther, 2005; Wang, 2010). This requires the aforementioned supervised classification step for classifying observations of one sample to the clusters of the other sample. However, the approaches could in principle be modified to also apply to descriptive clustering.

When splitting the dataset multiple times to determine the stability of a clustering method or parameter, eventually information from the whole dataset enters the method selection process. Thus putting aside a validation dataset that is only used *after* the method selection is advised. Even if a clustering is chosen by stability analysis on a discovery dataset, it is *not* guaranteed that this clustering can be validated on a validation dataset.

Stability analysis can also be combined with classical internal validation indices by checking whether internal validation indices have similar values for multiple clusterings calculated on subsamples of the data (Jain & Moreau, 1987), see also Dangl and Leisch (2020) for a related approach. This idea will also be part of our framework in Section 3.

2.4 | Visual validation

Cluster analysis is often exploratory without fixed predefined expectations from the user. Patterns in the data that qualify to be interpreted as clusters can have very diverse appearances. Some key characteristics of clusters, such as being areas of high density separated by areas of lower density, are difficult to translate into easily computable statistics. Furthermore, many clustering methods rely on model assumptions and cluster concepts, the appropriateness of which is hard to diagnose by means other than visual. This explains why visual validation is important in cluster analysis.

Clusters can be declared valid based on visualization if they correspond to clearly visible patterns in the data, or in some cases if the assumptions required for the chosen clustering method look valid.

Useful plots for visual cluster validation can be distinguished into:

1. General purpose data plots in which found clusters can be indicated by colors or glyphs, such as scatterplots, matrix plots, principal components biplots, multidimensional scaling, or parallel coordinates plots (Cook & Swayne, 2007, chapter 5). There are also projection pursuit approaches that generate "interesting" data projections, potentially showing clustering structure, without requiring the clustering as input (e.g., Tyler et al., 2009).
2. Plots set up to visualize a specific clustering, which can be further classified as:

a. Plots that visualize the original data directly, such as cluster heatmaps (Hahsler & Hornik, 2011; Wilkinson & Friendly, 2009) or projections to optimally discriminate clusters (Hennig, 2004).
b. Plots that visualize the clustering solution without representing the original observations directly such as dendrograms, silhouette plots, and neighborhood graphs (Leisch, 2008).

We refer to the Supporting Information for an illustration of some of these methods.

Plots that visualize the original data directly can be used to assess patterns in data space, although these plots come with either information loss by dimension reduction, or heavy reliance on aspects such as variable and observation ordering. The advantage of plots that optimize objective functions dependent on the clustering, such as discriminant projections or heatmaps with orderings determined by the clustering, is that they have better chances to bring out the data patterns corresponding to the clustering than general purpose plots. On the other hand, they may lead to an overoptimistic assessment of the validity of the clustering, or an interpretation of spurious patterns. Validation data that is kept separate from the beginning of the analysis may help to avoid overoptimism, see Section 3.4.

Some of the plots that do not represent the original observations directly can also be valuable for cluster validation. The silhouette plot accompanies the Average Silhouette Width index (Kaufman & Rousseeuw, 2009) and gives observation-wise information about the quality of assignment in the given clustering; dendrograms visualize the hierarchical merging process and can sometimes reveal issues, such as potentially meaningful clusters disappearing at higher levels of the hierarchy.

3 | A SYSTEMATIC FRAMEWORK FOR VALIDATING A CLUSTERING ON A VALIDATION DATASET

In this section, we present a systematic framework for validating a clustering on a validation dataset that includes many existing approaches from the literature as special cases and revisits them more formally. We also show how the validation methods that we reviewed in the last section are incorporated into the framework.

We first discuss what is meant by a "validation dataset" in Section 3.1. In Section 3.2, we give an overview of properties of a clustering result that may be validated on the validation set (these properties are strongly related to the classical validation procedures discussed in Section 2). In Section 3.3, we outline the distinction between method-based and result-based validation on a validation dataset. In Section 3.4, we combine the concepts of Sections 3.2 and 3.3 into an overview of strategies for validation on validation data. In Section 3.5, we discuss how to judge whether "successful" validation has been achieved.

3.1 | Validation datasets

The term "validation dataset" can refer to a dataset composed of independently collected data (e.g., collected by other researchers or in a different laboratory) which is similar enough to the original data for cluster evaluation to be possible. In practice, however, genuinely independent data is often not available. In this case, one might split a single dataset into a discovery and a validation set.

Apart from this consideration, the structure of the validation data depends on two further aspects: (a) The data for clustering can either be object by variable data or object by object proximity data (where the term "proximity" denotes

either similarities or dissimilarities), see Van Mechelen et al. (2018). Here, "objects" denote the entities which are to be clustered. (b) The aim of the clustering could either be inferential or descriptive, as defined in the introduction (Section 1).

For inferential clustering, the validation data consists of more objects to cluster. On the other hand, for descriptive clustering, validation data does *not* consist of more objects because the set of objects to be clustered is fixed. Consider the example of the $n \times p$ gene expression dataset as described in the introduction. This dataset can be understood as an object by variable dataset in two ways. For inferential clustering of the patients, the patients are the "objects" and the genes the "variables." A validation dataset consists of more patients. For descriptive clustering of the genes, now the genes constitute the "objects," and the patients are the "variables." A validation dataset consists of more variables, that is, again of more patients.

In Table 1, we give a general overview of the structure of the validation data, where we distinguish between inferential and descriptive clustering as well as between object by variable and object by object data.

If separately collected data is not available, and the dataset must be split into discovery and validation sets, a 50/50 split ratio is usually chosen. Indeed, we believe that this choice makes sense in most cases: validation strategies often require the number of data points in the validation set to not be too small when trying to validate certain properties obtained from the clustering on the discovery set. A similar argument has been made in the context of stability analysis (Lange et al., 2004).

3.2 | Clustering properties to be validated

In the literature, we identified four categories of properties of clusterings that researchers may want to validate.

(Int) Internal properties of the clusters (that turn up when clustering the discovery data), for example:

- descriptive measures of the clusters such as the values of the cluster centroids or the relative sizes of the clusters,
- the value of an internal validation index calculated for the clustering result, and
- subsets of variables that characterize the clusters.

(Ext) Associations of the clusters with external variables or agreement of the clustering with an externally known partition. Some examples:

- Clusters of cancer patients have different mean survival rates.
- A clustering of genes shows some agreement with known functional gene labels. For example, a clustering may be compatible with known partitions of the genes into functional categories. Less restrictively, some particular genes, of which this was previously expected, may be in the same cluster.

(Vis) Characteristics that can be assessed using visualization: do the clusters correspond to distinctive meaningful patterns in the data? Do the clusters look how they were supposed to look like? This could refer to model assumptions for the clustering method, or a priori hypotheses or requirements by the researcher.

TABLE 1 Structure of the validation data depending on inferential versus descriptive clustering and object by variable versus object by object data

	Inferential clustering	Descriptive clustering
Object by variable data	Validation data: further objects, same variables	Validation data: further variables, same objects
If a single dataset is split	Split performed along the objects	Split performed along the variables
Object by object data	Validation data: proximity matrix of further objects	Validation data: proximity matrix of same objects, but with proximities derived from another source (e.g., based on different underlying variables).
If a single dataset is split	Objects can be split into two disjoint sets, yielding two smaller proximity matrices (one representing the discovery data, the other the validation data).	Impossible to split proximity data directly into discovery and validation data, but may be possible to split underlying variables.

(Stab) Stability of cluster membership: Does cluster membership remain stable when the same method (algorithm, number of clusters, etc.) is applied to the validation data? Since the objects in the discovery and the validation set are disjunct in the case of inferential clustering, this involves supervised classification of objects of one dataset to clusters of the other dataset.

Most subsections of Section 2 correspond to a category in the above list, with the exception of Section 2.1 (recovery of "true" clusters). If the "true" cluster labels are indeed known, this can be considered as a part of (Ext).

3.3 | Method-based and result-based validation

The validation of a clustering on a specific dataset can refer either to the validity of the used clustering *method*, or the validity of the clustering *result* itself. While this distinction is often not made clear in the literature on classical validation procedures, it has important implications for how validation on a validation dataset is performed. We thus distinguish between *method-based* and *result-based* validation on validation data, as illustrated in Figure 3. In the following, we explain these terms in more detail.

We denote the discovery data by D_1 and the validation data by D_2. The clustering chosen on D_1 in Step 1 (method selection, see Figure 2) is called C_1. Given C_1, the validation dataset can be handled in two different ways:

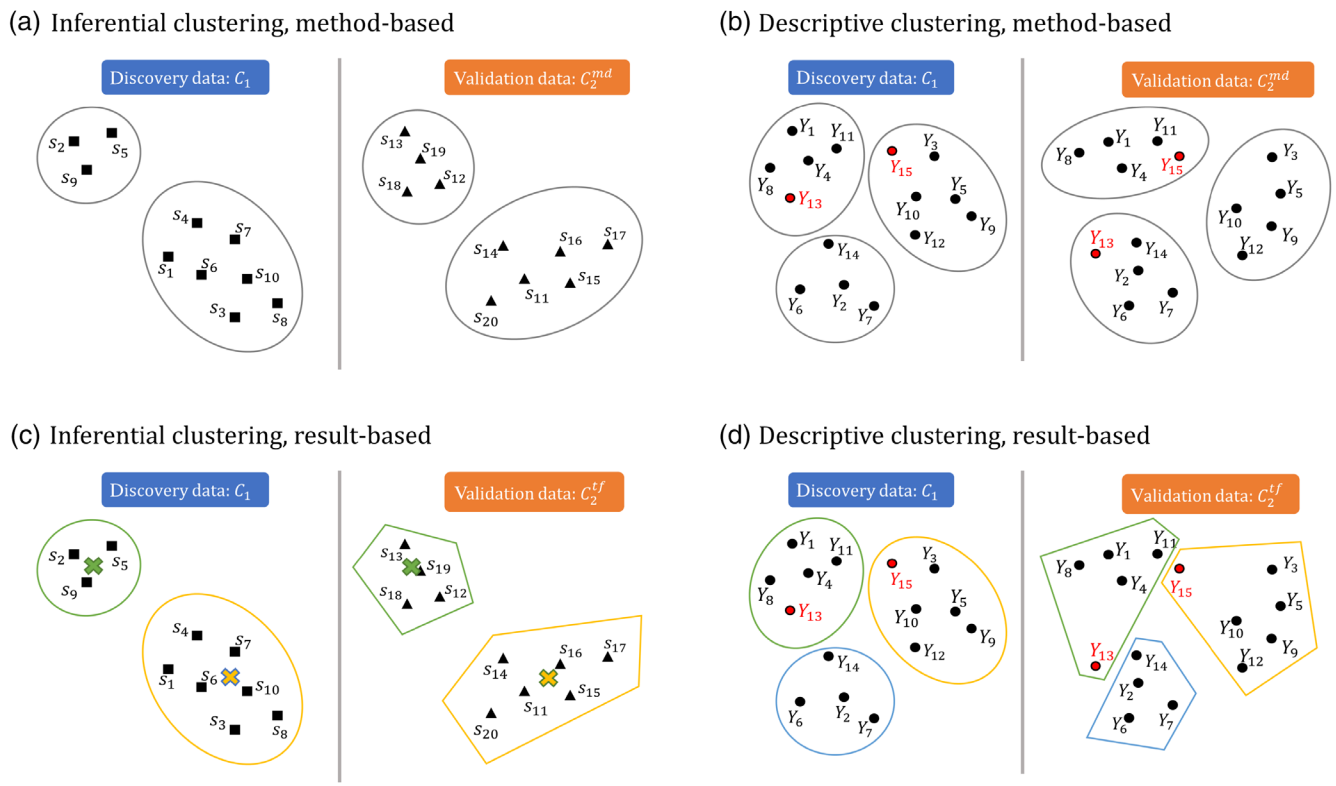

FIGURE 3 Method- and result-based validation for inferential and descriptive clustering. We use the same data example as in Figure 1. The top panel (a) and (b) (method-based validation) is from Figure 1. For inferential clustering (a), re-applying the clustering method to the validation data again detects a smaller cluster on the top left and a larger one on the bottom right. For descriptive clustering (b), the clustering C_2^{md} on the validation data groups the elements Y_{13} and Y_{15} (marked in red) differently than the clustering C_1 on the discovery data. The bottom panel (c) and (d) (result-based validation) illustrates the classification procedures that yield C_2^{tf}. The clusterings C_2^{tf} are depicted as polygons. The colors of the polygons match the corresponding clusters on the discovery data. For inferential clustering, nearest-centroid classification is depicted: the green and yellow crosses represent the centroids of C_1. The samples in the validation data are then assigned to the nearest centroid. In this particular example, the resulting clustering C_2^{tf} in (c) is equal to C_2^{md} in (a): in our terminology, the criterion (Stab) is perfectly fulfilled. For descriptive clustering (d), the most obvious way of transferring C_1 to the validation data is to set the cluster memberships in C_2^{tf} equal to those of C_1. In particular, the elements Y_{13} and Y_{15} are clustered as in C_1. Comparing C_2^{tf} with C_2^{md} in (b) shows that the cluster memberships are not perfectly stable according to criterion (Stab)

a. The same clustering method that yielded C_1 (i.e., same algorithm, same number of cluster k, etc.) can be applied to D_2, yielding a clustering C_2^{md} on D_2 ("md" for "method"). C_1 and C_2^{md} can then be compared with respect to aspects (Int), (Ext), or (Vis). We call this approach *method-based validation*. It puts a focus on the structural similarity of the clustering results as generated by the method.
b. Instead of applying the clustering method again, C_1 can be "transferred" to the validation data by using a supervised classifier to predict the cluster labels of the validation set (explained in more detail below). This results in a clustering C_2^{tf} on D_2 ("tf" for "transferred"). The transferred clustering can be compared to the original clustering C_1 with respect to aspects (Int), (Ext), or (Vis). We call this approach *result-based validation*. It puts a focus on whether the specific clustering result is also sensible for the validation data.

We now explain what we mean by "transferring" the clustering. For descriptive clustering, C_2^{tf} is simply C_1 (recall that for descriptive clustering, the objects to be clustered are the same for D_1 and D_2, and thus C_1 can immediately be considered to be a clustering of D_2). For inferential clustering, the objects to be clustered are different in the discovery and validation sets, so some proper "transfer" is required. This can be done using a supervised classifier (using the labeled discovery set (D_1, C_1) as "training set") to assign the objects in D_2 to the clusters in C_1 (Akhanli & Hennig, 2020; Lange et al., 2004). For example, one can calculate the centroids of the clusters in C_1, and then assign each sample in D_2 to its nearest centroid ("nearest-centroid classifier"). As C_2^{tf} is supposed to be an "extension" or "transfer" of the original clustering to the validation data, one should use a classifier that fits the assignment rule of the chosen clustering algorithm as closely as possible. The nearest-centroid classifier is suitable for k-means, which indeed clusters points by assigning them to the nearest centroid (Lloyd, 1982). For suitable classifiers for other clustering algorithms see Akhanli and Hennig (2020).

For (Stab) (stability of cluster membership), the clustering method needs to be applied again to D_2. We check whether the cluster memberships resulting from applying the method to the validation data are similar to the cluster memberships resulting from transferring the original clustering to the validation data. This combines (a) and (b).

3.4 | Overview of validation strategies

Table 2 combines the concepts of Sections 3.2 and 3.3 into an overview of strategies for validation on validation data. The precise choice of indices, plots, and so forth depends on the specific context of the analysis. We refer to Section 4 for illustrative examples from the applied literature.

Here are some considerations regarding the different strategies. The commented R code in the Supporting Information illustrates the following paragraphs with real-world datasets and concrete choices for indices and visualization tools.

3.4.1 | Validating (Int): Internal properties of the clustering

When applying result-based validation, the clusters of C_2^{tf} correspond to those of C_1. This makes the comparison easier. For method-based validation, the clusters of C_2^{md} are not automatically associated one-to-one with the clusters of C_1. Such an association is not needed when calculating internal indices that refer to a whole clustering, and comparing the index values between the clusterings on discovery and validation data. However, one may also be interested in comparing characteristics of specific clusters such as cluster centroids. In this case, there needs to be a matching of the clusters of C_2^{md} to the clusters of C_1, usually assuming that their number is the same. There are various methods to do this. For

TABLE 2 Strategies for validation on validation data

	Method-based validation	Result-based validation
	Compare C_1, C_2^{md} with respect to:	Compare C_1, C_2^{tf} with respect to:
(Int)	Internal properties	Internal properties
(Ext)	External associations	External associations
(Vis)	Visual properties	Visual properties
(Stab)	Compare C_2^{md}, C_2^{tf} with respect to cluster membership	

example, in centroid-based clustering one could match the centroids so that the sum of distances between centroids of matched clusters is minimal (Mirkin, 2005). Breckenridge (2000) suggests associating each cluster of C_2^{md} to a cluster of C_2^{tf} (e.g., by choosing the cluster association that maximizes the sum of the intersections of the clusters). The one-to-one cluster association of C_2^{tf} to C_1 can then be used to assign each cluster of C_2^{md} to one of C_1.

3.4.2 | Validating (Ext): Associations with external variables or agreement with externally known partitions

As for method-based validation of internal properties (Int), here too it may be necessary to match the clusters of C_2^{md} to those in C_1 and the remarks made above apply again. Note that this is not necessarily required. For example, testing whether the clusters are associated with an external variable, such as survival time, without interpreting the association of specific clusters, does not require matching.

For result-based validation of descriptive clustering, the partition C_2^{tf} is actually equal to C_1. This makes certain approaches such as testing an association between cluster membership and an external variable on both discovery and validation data meaningless.

3.4.3 | Validating (Vis): Visual patterns

Using the same variables for D_1 and D_2 as in inferential clustering, some plots such as scatterplots or parallel coordinates plots can visualize both C_2^{md} and C_2^{tf} in a straightforward manner comparable to C_1. Some other plots such as principal components biplots, other linear projection plots such as those in Hennig (2004), and multidimensional scaling require a selection of an optimal projection space for the dataset to be plotted. Although this could be done on the validation data, for inferential clustering, plotting the validation dataset on the projection space defined by the discovery dataset (and its clustering, if the projection space depends on it) allows for a more direct comparison. For linear projection methods, this requires a standard linear projection given the coordinate axes determined from D_1. For multidimensional scaling, there are techniques to embed new observations into the projection space defined by the original observations, for example, Gower (1968). For descriptive clustering, on the other hand, embedding the observations of D_2 in the space defined by D_1 is not informative as the points would be identical, so here an optimized projection space for D_2 must be found.

Some other plots, such as the silhouette plot and cluster heatmaps (as long as observations are ordered only by a partition rather than a full dendrogram), may benefit from matching clusters for determining their order, see the comments on internal validation (Int) in Section 3.4.1.

The results of visual validation are subjective, and although plots are reproducible given both discovery and validation datasets, the way the researcher arrives at a validity verdict will not be reproducible. Displaying the involved plots will give the reader the chance to form their own conclusions.

3.4.4 | Validating (Stab): Stability of cluster membership

Here one needs to compute both C_2^{tf} and C_2^{md}. These are then compared with an index for comparing partitions. The rationale behind this is as follows: cluster memberships in C_1 and C_2^{md} are compared to check whether repeated application of the clustering method leads to stable cluster memberships. For descriptive clustering, C_1 can be compared to C_2^{md} directly (here C_1 is equal to C_2^{tf}). For inferential clustering, C_1 and C_2^{md} cannot be compared directly because they are partitions of different sets of objects. Thus C_2^{tf} is used as a surrogate for C_1 on D_2. Different choices of a partition similarity index are possible, for example, the ARI, the Jaccard index, or the FM index (for overviews, see Meila, 2015; Albatineh et al., 2006).

3.5 | When is a clustering successfully validated?

Due to random variation, researchers will hardly ever achieve the exact same results on discovery and validation data. So far, there seem to be no systematic approaches for judging "validation success" in the context of validating clustering

results on validation data. In this section, we review the current status and outline which aspects would be interesting to study in further research.

The problem of defining "successful" validation does not only arise in cluster analysis, but generally in validation or replication studies. Here we consider "validation" to be the broader term, and "replication" as more specific, for which strategies of the validation framework can be used. "Replication" refers to using new data to re-assess scientific claims made in a previous publication (Nosek & Errington, 2020). The discussion about judging replication success is ongoing in the field of methodological research on replication studies, mostly in the context of hypothesis tests and effect estimates. For example, Hedges (2019) and Held (2020) argue that, when trying to replicate a hypothesis test (that was significant on the original data), it is not enough to check whether the test on the replication data is significant again. Actually, the binary distinction between significance and insignificance may not be helpful, for example, when comparing p values of 0.04 and 0.06 (given a significance level of 0.05). Rather, we should also check whether the effect estimate in the replication study provides evidence for the claim about the effect in the original study. Some clustering validation aspects are connected to significance tests, particularly testing for external associations in (Ext). The same caveats apply here regarding general replication of test results.

The consideration of differences between (internal or external) validity measurements on discovery and validation data, or the consideration of an index value for stability between discovery and validation sets, could in principle also be framed as a testing problem of a null hypothesis formalizing some kind of equality of structure. To our knowledge, this has not been performed yet and is left as a potential direction of future research. It can be expected that validation data results will not be quite as good due to selection bias originating from basing selection of the final clustering on results of the discovery data: the more different clustering algorithms or parameters are tried during the analysis on the discovery data, the more likely it is that one of them yields a satisfying result. If only the best result is chosen, this might be "overoptimistic" to some extent. In other words, the multiplicity of possible analysis strategies may hinder replicability (Hoffmann et al., 2021), see also the discussion in Section 5. Observing slightly worse values on the validation data is thus to be expected and does not necessarily mean that the validation has failed. However, if the results are severely worse, then this suggests problematic overoptimism on the discovery data.

As it stands, it must be acknowledged that the question "is validation successful?" cannot simply be answered with "yes" or "no". The validation dataset may deliver high or low agreement regarding various aspects (internal and external validity, stability, visual aspects) with what was found on the discovery data—where the clustering on the discovery data may already have been assessed as a weaker or stronger clustering in Step 1. For example, regarding an internal index, such as the Average Silhouette Width, it is of interest both whether the value is reasonably high on the discovery dataset alone, and whether the validation dataset supports whatever value was found on the discovery data. Guidelines or thresholds for interpreting index values are rarely given and in fact mostly arbitrary, so the researcher must rely on their understanding of the index, experience, and judgment.

4 | EXAMPLES FROM THE APPLIED LITERATURE

In this section, we review application studies that conducted cluster analysis on a discovery set and then validated the results with a validation set. Our aim is to demonstrate how these studies fit into the framework outlined above. Given the vast amount of applied cluster analysis studies, it is impossible to list every cluster study that used a validation set. Rather, we start by giving a short historical overview and then present some exemplary studies in Table 3.

The appearance of clustering studies that used a discovery and a validation set dates back to at least the 1960s. One of the first clustering studies that used a validation set was Goldstein and Linden (1969) who clustered patients with alcohol use disorder. In our terms, they performed method-based validation with respect to internal properties. Rogers and Linden (1973) provided an early implementation of stability-based validation, (Stab). They clustered college freshwomen based on personality features and used discriminant analysis as the classifier to derive C_2^{tf}. (Stab) was then presented more systematically by McIntyre and Blashfield (1980) and Breckenridge (1989).

In recent decades, many more clustering studies that use validation data have appeared. In Table 3, we list exemplary applied studies for the different validation types as outlined in Table 2. The studies are taken from our main field of expertise, that is, medicine and health science. Some of these studies used multiple aspects of the validation framework, but for the sake of illustration, we only list one validation type per study. We did not find an example for result-based validation of (Vis). In general, there appear to be few studies which performed validation of visual properties on

TABLE 3 Study examples for each validation type

Validation type and study	Clustering aim and validation motivation	Validation data	Cluster algorithm	Validation procedure
(Int) result-based: Kapp and Tibshirani (2007)	Inferential clustering of breast cancer patients based on microarray gene expression to validate breast cancer subtypes that were previously found by Sørlie et al. (2003).	Validation data consisted of independently collected samples from different countries.	hierarchical clustering	C_2^{tf} was derived via a variant of nearest-centroid classification (each sample in the validation data was assigned to the original cluster whose centroid had the maximum Pearson's correlation coefficient with the sample). C_2^{tf} was then evaluated with a newly introduced internal validation index (the "in-group proportion" IGP). This index was combined with a statistical test procedure that consists of generating centroids randomly placed in the data, classifying the samples of the validation data to these centroids to obtain clusterings $\widetilde{C}_2^{tf}$, and comparing the values of the internal index for the $\widetilde{C}_2^{tf}$'s with the index value for C_2^{tf}. The IGP was not applied to C_1.
(Int) method-based: De Bourdeaudhuij and Van Oost (1998)	Inferential clustering of adolescents to find clusters of health behavior (with respect to smoking, alcohol use, sleeping, food choice, BMI, and physical activity). Validation was used by the authors to replicate their own findings.	Validation data was a separately collected dataset.	hierarchical clustering	Clusters of C_1 and C_2^{md} were matched manually. Compared means of health behavior variables (centroids) between C_1 and C_2^{md} (e.g., the mean amount of smoking for cluster 1 was compared between C_1 and C_2^{md} and so on). Overall, the centroids were similar between both clusterings. In particular, both the most "healthy" and the most "unhealthy" cluster could be recovered from the validation data.
(Ext) result-based: Curtis et al. (2012)	Inferential clustering of breast cancer patients based on copy number and gene expression data to discover novel breast cancer subtypes. The authors chose result-based validation because in clinical practice, doctors would typically want to assign a new patient to a subtype, and the validity of such a procedure can be analyzed by classification of the validation samples to yield C_2^{tf} and then comparing C_2^{tf} to C_1.	Validation data was a second cohort from the same tumor banks.	iCluster (Shen et al., 2009)	C_2^{tf} was derived via nearest shrunken centroid classification (Tibshirani et al., 2003), which is a modification of nearest-centroid classification where the cluster centroids are shrunk towards the overall centroid. Comparison of C_1 and C_2^{tf} w.r.t. their associations with survival: a Cox proportional hazards models was fitted to the discovery data (respectively validation data), with the cluster memberships of C_1 (resp. C_2^{tf}) as covariates. The hazard ratios of the clusters were similar between the model for the discovery data and the model for the validation data.

(Continues)

TABLE 3 (Continued)

Validation type and study	Clustering aim and validation motivation	Validation data	Cluster algorithm	Validation procedure
(Ext) method-based: Freudenberg et al. (2009)	Descriptive clustering of cancer-related genes to find biologically meaningful clusters of co-expressed genes, which may help to elucidate biological pathways and generate hypotheses about transcriptional regulatory mechanisms. Authors performed validation to check the replicability of the clustering results.	Validation data was an independently collected breast cancer dataset.	hierarchical clustering combined with the CSIMM algorithm (Liu et al., 2006)	Each gene in the clustering was assigned a CLEAN score, a newly introduced external measure for agreement with previously known functional categories. The correlation between the gene CLEAN scores obtained with C_1 and C_2^{md} was calculated.
(Vis) method-based: Sweatt et al. (2019)	Inferential clustering of pulmonary arterial hypertension (PAH) patients based on blood proteomic profiles to find distinct PAH immune phenotypes. The underlying idea was that patient subgroups might express distinct patterns of inflammation in blood, and the detection of these groups may in turn help to develop tailored treatments in future studies. Validation was performed to assess whether the results generalize to other patients.	Validation data consisted of independently collected samples from a different country.	Consensus Clustering (Monti et al., 2003)	Clusters of C_1 and C_2^{md} were matched manually. Compared heatmaps and PCA plots for C_1 and C_2^{md} which were generated separately for discovery and validation data, no common projection space was used. The heatmaps and PCA plots were deemed to be similar between discovery and validation data.
(Stab): Bergström et al. (2001)	Inferential clustering of spinal pain patients based on the Multidimensional Pain Inventory, which is a battery of questionnaires where patients self-report their pain severity, pain-related interference in everyday life, etc. Previous studies had detected distinct subgroups of spinal pain patients with respect to how well the patients coped with their disease, which has implications for tailored treatment. The authors sought to find similar clusters in their data, and performed validation to assess the replicability of their own findings.	Validation data was an independently collected dataset.	k-means	C_2^{tf} was derived via nearest-centroid classification. The kappa coefficient (a partition similarity index) was used to compare C_2^{md} and C_2^{tf}. The resulting index value was 0.82, which was judged as indicating very good agreement.

a validation dataset in a thorough manner. We believe future studies would benefit from considering the procedures for (Vis), outlined above.

The studies cited in Table 3 mostly treat validation and discovery data *asymmetrically* (with the exception of Freudenberg et al., 2009, and Bergström et al., 2001). This is more obvious for result-based validation: the clustering C_1 is transferred to the validation data (and not the other way around). Method-based validation may appear more symmetric because the same method is applied to both discovery and validation data and the results are typically compared descriptively in a symmetric fashion. However, method-based validation can be asymmetric to the extent of which the validation data is kept apart from the method selection on the discovery data, and is only used later without model selection to validate the results on the discovery data. Asymmetry could be made more explicit by using a suitable test procedure to judge validation success (inspired by the methodological research on judging replication success, for example, Held (2020) advocates for an asymmetric approach when comparing the replication study to the original study), but as discussed in Section 3.5, such approaches do not seem to exist yet for cluster analysis.

Many studies in the literature do not strictly set apart the validation data during Step 1 (method selection). That is, these studies use the result of the validation on the validation data for method selection (e.g., Brennan et al., 2012; Jamison et al., 1988; Sinclair et al., 2005). In contrast, we have argued in the introduction and in Section 2.3 that for the purpose of validating a clustering result on validation data in the sense of our framework, method selection should be finished after Step 1.

Another validation variant is also frequently found in the literature (e.g., Ailawadi et al., 2001; Gruber et al., 2010; Homburg et al., 2008; Kaluza, 2000; Phinney et al., 2005): method selection is performed on the whole dataset, after which the data is split into two sets. The chosen cluster method is applied to the first set, and then validation on the second set (the validation data) is assessed. Successful "validation" may indicate a certain robustness or stability of the result, but in order to avoid overoptimism on the validation data, method selection should be constrained to the first part of the split dataset, and not be performed on the whole data according to our framework.

Other studies (e.g., Alexe et al., 2006) perform a procedure that appears similar to method-based validation: they split a dataset into two halves, use the first half as the discovery set, but obtain C_2^{md} by clustering discovery and validation data *together* (instead of only clustering the validation data), which again will likely yield more optimistic validation results than if C_2^{md} had been obtained based on the validation data only.

5 | DISCUSSION

We have presented a systematic framework for validating clusterings on a validation dataset that encompasses procedures known from the literature. This framework might help researchers to identify a suitable approach to validate their clustering results in future studies. However, the procedure cannot be performed in an "automated" manner. Rather, it requires substantial input from the researchers who must decide which validation criteria are important for them depending on the substantive context. Furthermore, specific indices and plots need to be chosen, as well as whether the amount of agreement between results on the discovery and validation datasets is assessed as sufficient. We have given hints about when some aspects may be of interest, but as every application is different, there are no clear rules. This holds for the clustering process in general: while cluster analysis is often interpreted as being able to find meaningful structure in the data "on its own", the choice of cluster concept and method requires thorough consideration by researchers (Akhanli & Hennig, 2020; Hennig, 2015b). The same is true for our validation framework.

Performing validation on the validation data adds some computational complexity to the cluster analysis. However, the overall complexity is often less than twice the complexity that would result from only analyzing the discovery data: frequently method selection is performed on the discovery data, and this possibly time-consuming process is not applied to the validation data.

Regarding the choice of validation data, a validation dataset could be obtained by splitting the original dataset, or it could be a separately collected dataset. On one hand, if the validation dataset and the discovery dataset are obtained by splitting an originally collected dataset, it is unclear whether a successful validation allows for generalization to data from other sources. Moreover, this reduces the size of the data and can make it more difficult to find meaningful cluster structure in the data. On the other hand, if the validation data have been independently collected (potentially coming from a different distribution) and the validation fails, it can be difficult to determine whether this is due to the clustering not being meaningful, or due to systematic differences between discovery and validation data. Conversely, if validation *is* successful, then this is all the more encouraging, because it suggests that the clustering result may be valid in a more general context.

Notably, the validation of clustering results on a validation dataset may also allow detection of "overoptimism" due to "overfitting" effects: when researchers try different clustering algorithms or parameters during the analysis, they can use classical internal and external validation methods to choose a single clustering out of these. However, the more clustering methods tried, the more likely it is that one of them yields a satisfying result by chance. Consequently, the reported results may be less reliable than they seem, similarly to the results of multiple tests if no adjustment is performed. While this is well-understood in the context of multiple testing, this is less so in the context of clustering. Repeating the same cluster analysis on another dataset is a sensible approach to ensure that seemingly satisfactory results are not (solely) the product of such overfitting effects.

In future work, it would be interesting to study further aspects of cluster validation in relation to validation data use. *Hypothesis testing* is an approach to cluster validation that we have not embedded in our framework. For example, one can test if a clustering result is significantly "better" than clusterings generated by the same method on homogeneous datasets (for an overview, see Huang et al., 2015). This can involve internal validation indices (Dubes, 1993; Gordon, 1998; Halkidi et al., 2002; Hennig & Lin, 2015) or stability analysis (Bertrand & Mufti, 2006; Dudoit & Fridlyand, 2002; John et al., 2020; Smith & Dubes, 1980). We do not know of work where hypothesis testing for cluster validation has involved validation data, but it could be of interest to derive distributions under suitable null hypotheses for statistics that are evaluated on validation data.

In conclusion, our hope for this framework is to improve the interpretation of clustering studies that use validation data, and to stimulate the use of validation sets in cluster analysis.

ACKNOWLEDGMENTS

We thank Anna Jacob and Alethea Charlton for making valuable language corrections. This work has been partially supported by the German Federal Ministry of Education and Research (BMBF) [grant number 01IS18036A to Anne-Laure Boulesteix (Munich Center of Machine Learning)] and the German Research Foundation [grant number BO3139/7-1 to Anne-Laure Boulesteix]. The authors of this work take full responsibility for its content.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Theresa Ullmann: Conceptualization (equal); methodology (lead); writing – original draft (lead); writing – review and editing (equal). **Christian Hennig:** Conceptualization (supporting); methodology (supporting); supervision (supporting); writing – original draft (supporting); writing – review and editing (equal). **Anne-Laure Boulesteix:** Conceptualization (equal); funding acquisition (lead); methodology (supporting); supervision (lead); writing – original draft (supporting); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Theresa Ullmann https://orcid.org/0000-0003-1215-8561
Christian Hennig https://orcid.org/0000-0003-1550-5637
Anne-Laure Boulesteix https://orcid.org/0000-0002-2729-0947

RELATED WIREs ARTICLE

Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey

REFERENCES

Ailawadi, K. L., Neslin, S. A., & Gedenk, K. (2001). Pursuing the value-conscious consumer: Store brands versus national brand promotions. *Journal of Marketing*, *65*(1), 71–89.

Akhanli, S. E., & Hennig, C. (2020). Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Statistics and Computing*, *30*(5), 1523–1544.

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*(2), 301–313.

Alexe, G., Dalgin, G. S., Ramaswamy, R., DeLisi, C., & Bhanot, G. (2006). Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns. *Cancer Informatics*, *2*, 243–227.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Ben-David, S., Von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. In *International conference on computational learning theory* (pp. 5–19). Springer.

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, *7*, 6–17.

Bergström, G., Bodin, L., Jensen, I. B., Linton, S. J., & Nygren, A. L. (2001). Long-term, non-specific spinal pain: Reliable and valid subgroups of patients. *Behaviour Research and Therapy*, *39*(1), 75–87.

Bertrand, P., & Mufti, G. B. (2006). Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics & Data Analysis*, *50*(4), 992–1015.

Boulesteix, A.-L., & Hatz, M. (2017). Benchmarking for clustering methods based on real data: A statistical view. In F. Palumbo, A. Montanari, & M. Vichi (Eds.), *Data science—Innovative developments in data analysis and clustering* (pp. 73–82). Springer.

Boulesteix, A.-L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, *6*, 77–97.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, *24*(2), 147–161.

Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research*, *35*(2), 261–285.

Brennan, T., Breitenbach, M., Dieterich, W., Salisbury, E. J., & Van Voorhis, P. (2012). Women's pathways to serious and habitual crime: A person-centered analysis incorporating gender responsive factors. *Criminal Justice and Behavior*, *39*(11), 1481–1508.

Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., Mills, G. B., Lau, C. C., & Brown, P. H. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, *21*(7), 1688–1698.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Cook, D., & Swayne, D. F. (2007). *Interactive and dynamic graphics for data analysis with R and GGobi*. Springer.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., & Yuan, Y. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*(7403), 346–352.

Dangl, R., & Leisch, F. (2020). Effects of resampling in determining the number of clusters in a data set. *Journal of Classification*, *37*, 558–583.

De Bourdeaudhuij, I., & Van Oost, P. (1998). Family characteristics and health behaviours of adolescents and families. *Psychology and Health*, *13*(5), 785–803.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842.

Dolnicar, S., & Leisch, F. (2010). Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, *21*(1), 83–101.

Dougherty, E. R., Hua, J., & Bittner, M. L. (2007). Validation of computational methods in genomics. *Current Genomics*, *8*(1), 1–19.

Dubes, R. C. (1993). Cluster analysis and related issues. In C. H. Chen, L. F. Pau, & P. S. P. Wang (Eds.), *Handbook of pattern recognition and computer vision* (pp. 3–32). World Scientific Publishing Company.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, *3*(7), research0036.1–0036.21.

Fang, Y., & Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, *56*(3), 468–477.

Färber, I., Günnemann, S., Kriegel, H.-P., Kröger, P., Müller, E., Schubert, E., Seidl, T., & Zimek, A. (2010). On using class-labels in evaluation of clusterings. In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD*, Washington, DC.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Freudenberg, J. M., Joshi, V. K., Hu, Z., & Medvedovic, M. (2009). Clean: Clustering enrichment analysis. *BMC Bioinformatics*, *10*(1), 234.

Fu, W., & Perry, P. O. (2020). Estimating the number of clusters using cross-validation. *Journal of Computational and Graphical Statistics*, *29*(1), 162–173.

Garrido-Castro, A. C., Lin, N. U., & Polyak, K. (2019). Insights into molecular classifications of triple-negative breast cancer: Improving patient selection for treatment. *Cancer Discovery*, *9*(2), 176–198.

Goldstein, S. G., & Linden, J. D. (1969). Multivariate classification of alcoholics by means of the MMPI. *Journal of Abnormal Psychology*, *74*(6), 661–669.

Gordon, A. D. (1998). Cluster Validation. In C. Hayashi, K. Yajima, H. Bock, N. Ohsumi, Y. Tanaka, & Y. Baba (Eds.), *Data science, classification, and related methods. Proceedings of the fifth conference of the International Federation of Classification Societies (IFCS-96)* (pp. 22–39). Springer.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, *55*, 582–585.

Gruber, M., Heinemann, F., Brettel, M., & Hungeling, S. (2010). Configurations of resources and capabilities and their performance implications: An exploratory study on technology ventures. *Strategic Management Journal*, *31*(12), 1337–1356.

Hahsler, M., & Hornik, K. (2011). Dissimilarity plots: A visual exploration tool for partitional clustering. *Journal of Computational and Graphical Statistics*, *20*(2), 335–354.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: Part I. *ACM SIGMOD Record*, *31*(2), 40–45.

Halkidi, M., Vazirgiannis, M., & Hennig, C. (2015). Method-independent indices for cluster validation and estimating the number of clusters. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 616–639). Chapman and Hall/CRC.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, *21*(15), 3201–3212.

Hedges, L. V. (2019). The statistics of replication. *Methodology*, *15*, 3–14.

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(2), 431–448.

Hennig, C. (2004). Asymmetric linear dimension reduction for classification. *Journal of Computational and Graphical Statistics*, *13*, 930–945.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271.

Hennig, C. (2015a). Clustering strategy and method selection. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 703–730). Chapman & Hall/CRC.

Hennig, C. (2015b). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62.

Hennig, C., & Lin, C.-J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing*, *25*(4), 821–833.

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925.

Homburg, C., Jensen, O., & Krohmer, H. (2008). Configurations of marketing and sales: A taxonomy. *Journal of Marketing*, *72*(2), 133–154.

Huang, H., Liu, Y., Hayes, D. N., Nobel, A., Marron, J. S., & Hennig, C. (2015). Significance testing in clustering. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 336–357). Chapman and Hall/CRC.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *44*, 223–270.

Jain, A. K., & Moreau, J. V. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, *20*(5), 547–568.

Jamison, R. N., Rock, D. L., & Parris, W. C. (1988). Empirically derived symptom checklist 90 subgroups of chronic pain patients: A cluster analysis. *Journal of Behavioral Medicine*, *11*(2), 147–158.

John, C. R., Watson, D., Russ, D., Goldmann, K., Ehrenstein, M., Pitzalis, C., Lewis, M., & Barnes, M. (2020). M3C: Monte Carlo reference-based consensus clustering. *Scientific Reports*, *10*(1), 1–14.

Kaluza, G. (2000). Changing unbalanced coping profiles-a prospective controlled intervention trial in worksite health promotion. *Psychology and Health*, *15*(3), 423–433.

Kapp, A. V., Jeffrey, S. S., Langerød, A., Børresen-Dale, A.-L., Han, W., Noh, D.-Y., Bukholm, I. R., Nicolau, M., Brown, P. O., & Tibshirani, R. (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics*, *7*(1), 231.

Kapp, A. V., & Tibshirani, R. (2007). Are clusters found in one dataset present in another dataset? *Biostatistics*, *8*(1), 9–31.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation*, *16*(6), 1299–1323.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, *121*(7), 2750–2767.

Leisch, F. (2008). Visualizing cluster analysis and finite mixture models. In C.-H. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 561–587). Springer.

Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, *13*(11), 2573–2593.

Liu, X., Sivaganesan, S., Yeung, K. Y., Guo, J., Bumgarner, R. E., & Medvedovic, M. (2006). Context-specific infinite mixtures for clustering gene expression profiles across diverse microarray dataset. *Bioinformatics*, *22*(14), 1737–1744.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research*, *15*(2), 225–238.

Meila, M. (2015). Criteria for comparing Clusterings. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 640–657). Chapman and Hall/CRC.

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement*, *11*(4), 329–354.

Mirkin, B. (2005). *Clustering for data mining: A data recovery approach*. CRC Press.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1), 91–118.

Morey, L. C., Blashfield, R. K., & Skinner, H. A. (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research*, *18*(3), 309–329.

Müller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis–a methodological approach. *Food Quality and Preference*, *34*, 70–78.

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Phinney, J. S., Dennis, J. M., & Gutierrez, D. M. (2005). College orientation profiles of Latino students from low socioeconomic backgrounds: A cluster analytic approach. *Hispanic Journal of Behavioral Sciences*, *27*(4), 387–408.

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., Dez, M., Viladot, M., Arance, A., & Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, *24*, S26–S35.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rogers, G., & Linden, J. D. (1973). Use of multiple discriminant function analysis in the evaluation of three multivariate grouping techniques. *Educational and Psychological Measurement*, *33*(4), 787–802.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, *95*(1), 14–18.

Sinclair, R. R., Tucker, J. S., Cullen, J. C., & Wright, C. (2005). Performance differences among four organizational commitment profiles. *Journal of Applied Psychology*, *90*(6), 1280.

Smith, S. P., & Dubes, R. (1980). Stability of a hierarchical clustering. *Pattern Recognition*, *12*(3), 177–187.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., & Geisler, S. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8418–8423.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., & Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10393–10398.

Sweatt, A. J., Hedlin, H. K., Balasubramanian, V., Hsi, A., Blum, L. K., Robinson, W. H., Haddad, F., Hickey, P. M., Condliffe, R., & Lawrie, A. (2019). Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension. *Circulation Research*, *124*(6), 904–919.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, *18*(1), 104–117.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

Tyler, D. E., Critchley, F., Dümbgen, L., & Oja, H. (2009). Invariant co-ordinate selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 549–592.

Van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F. & Steinley, D. (2018). Benchmarking in cluster analysis: A white paper. arXiv preprint arXiv:1809.10496.

Von Luxburg, U. (2010). *Clustering stability: An overview*. Now Publishers Inc.

Von Luxburg, U., Williamson, R.C., & Guyon, I. (2012). Clustering: Science or art? In Guyon, I., Dror, G., Lemaire, V., and Taylor, G., editors, Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of* Proceedings of Machine Learning Research, pages 65–79, . PML Research Press.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *97*(4), 893–904.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, *63*(2), 179–184.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, *5*(1), 1–9.

Zhang, Q., Burdette, J. E., & Wang, J.-P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*, *8*(1), 1–18.

Zimmermann, A. (2020). Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *WIREs Data Mining and Knowledge Discovery*, *10*(2), e1330.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2022). Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1444. https://doi.org/10.1002/widm.1444