# Systematic design, generation, and application of synthetic datasets for flow cytometry

PDA Journal
of Pharmaceutical Science and Technology

# Systematic design, generation, and application of synthetic datasets for flow cytometry

Melissa Cheung, Jonathan J Campbell, Robert J Thomas, et al.

*PDA Journal of Pharmaceutical Science and Technology* **2022**,
Access the most recent version at doi:10.5731/pdajpst.2021.012659

# Systematic design, generation, and application of synthetic datasets for flow cytometry

Melissa Cheung[1]*, Jonathan J. Campbell[2], Robert J. Thomas[1], Julian Braybrook[2] and Jon Petzing[1]

[1] Centre for Biological Engineering, Loughborough University, Loughborough, Leicestershire, UK

[2] National Measurement Laboratory, LGC, Teddington, UK.

* Corresponding author

Email: M.Cheung@lboro.ac.uk

## Abstract

Application of synthetic datasets in training and validation of analysis tools have led to improvements in many decision-making tasks in a range of domains from computer vision to digital pathology. Synthetic datasets overcome the constraints of real-world datasets, namely difficulties in collection and labelling, expense, time and privacy concerns.

In flow cytometry, real cell-based datasets are limited by properties such as size, number of parameters, distance between cell populations and distributions, and are often focused on a narrow range of disease or cell types. Researchers in some cases have designed these desired properties into synthetic datasets, however operators have implemented them in inconsistent approaches and there is a scarcity of publicly available, high-quality synthetic datasets.

In this research, we propose a method to systematically design and generate flow cytometry synthetic datasets with highly controlled characteristics. We demonstrate the generation of two-cluster synthetic datasets with specific degrees of separation between cell populations, and of non-normal distributions with increasing levels of skewness and orientations of skew pairs. We apply our synthetic datasets to test the performance of a popular automated cell populations identification software, SPADE3, and define the region where the software performance decreases as the clusters get closer together.

Application of the synthetic skewed dataset suggests the software is capable of processing non-normal data. We calculate the classification accuracy of SPADE3 with robustness not achievable with real-world datasets. Our approach aims to advance research towards generation of high-quality synthetic flow cytometry datasets, and to increase their awareness among the community.

The synthetic datasets can be utilised in benchmarking studies that critically evaluate cell population identification tools and help illustrate potential digital platform inconsistencies.

These datasets have the potential to improve cell characterisation workflows that integrate automated analysis in clinical diagnostics and cell therapy manufacturing.

Keywords: flow cytometry, synthetic datasets, clusters, separation, skew, accuracy, repeatability.

# 1 Introduction

An important challenge in manufacturing of emerging therapies is the need to develop and satisfy appropriate regulatory standards. To support advanced analytics, toolsets are required to establish quality assurance, such as data reference sets for data analysis, presentation, and interpretation (1).

Synthetic datasets are datasets generated by computer simulation, rather than collected through real world observations or experiments. These datasets are often created from mathematical models that approximate aspects of real-world data. Synthetic datasets can be referred to as ‘simulated’, ‘artificial’, ‘mock’, ‘toy’, and more colloquially, ‘dummy data’.

In unsupervised machine learning, well known ‘toy’ datasets used to compare different clustering algorithms include clusters in the shape of two rings, crescent moons, spirals, and data with no structures (2). Sophisticated synthetic datasets include urban street images applied to object detection for autonomous driving (3,4), and household objects images for object detection in the field of robotic manipulation (5). In medical imaging fields, synthetic datasets generated based on real images (e.g. magnetic resonance imaging (MRI), mammography, and whole-slide histopathology datasets) have demonstrated utility within computer-aided detection or computer-aided diagnosis systems, and may be useful for educational purposes and in quality control (6–8).

Further strategies to generate artificial datasets have made use of data augmentation methods (9). Similarly, in flow cytometry, cell subsets from real samples can be selected and computationally mixed in a copy-and-paste strategy with other real or synthetic cell populations, to create augmented and reprocessed ‘semi-synthetic’ datasets (10).

Real data are generally required during the development of computational analysis tools to provide means for training and validation, as well as potential decision-making reasons. An analysis tool implies any method that performs detection, recognition, identification, classification, tracking, prediction, or any other function that enables subsequent decision making. Real data also play an important role in benchmarking studies that evaluate the performance of these tools.

The key advantage of real data is that the information contains true characteristics of a biological system. Albeit biological systems are very large and have complex signalling pathway interactions between multiple cellular and molecular components. Real data have various limitations that necessitates the creation of synthetic datasets to overcome them. Being mathematical models, synthetic datasets can reduce the complexity of real data to gain insight into how they are processed by computational analysis tools that have hidden ‘black box’ algorithms.

A disadvantage of synthetic datasets is the issue of how accurately they represent the real data. The purpose of synthetic datasets is to simplify real data, which requires assumptions to be made and boundary conditions to be set around it, hence it is impossible to produce a faithful replication of the real data. Whilst it may not be necessary to capture all the complexities and features of real data in the design of synthetic ones in order to establish their utility and credibility, there is potentially an expectation from users that higher complexity equates to higher quality, leading to a lack of acceptance of synthetic data amongst the biomedical community.

Often, real datasets with predetermined criteria are difficult to collect because of limited availability at certain conditions and time periods. Synthetic data can be designed to mirror existing real data and further optimise the dataset by including rare cases and those at

extreme conditions, thereby enhancing the realistic range of features or parameters. Additionally, a high level of control is potentially achievable with synthetic datasets, where designers can quickly change one factor at a time or build up layers of complexity through controlled addition of factors.

Once collected, a major shortcoming of real data is the laborious and time-consuming task of labelling observations with meaningful information (e.g. healthy vs abnormal) performed by experienced personnel. Synthetic datasets can be designed with the labels inherent in the data, side-stepping this task. The desired property of the synthetic data is also known and can be applied in performance assessment of analysis tools.

A drawback of large-scale real datasets is that they sometimes take a large amount of time to acquire (particularly true concerning collection of rare events). An equivalent large-scale, complex synthetic dataset may also require a large amount of computing time to generate, however this problem is negated as computers become faster.

Further benefits of synthetic data are: the potential lower costs associated with use of a modern computer rather than expensive technical equipment, reagents and raw materials; the reproducibility of computer code; and the absence of personal data which means that the processing of synthetic data does not have the same privacy concerns and legal compliance requirements as that of real data (11).

In flow cytometry, synthetic datasets usually aim to mimic the properties of real cell populations. The properties of these randomly generated datasets range from simple two-dimensional datasets with four clusters (12), to up to 30 populations in 35 dimensions (13). The statistical distributions of synthetic clusters vary from normal (Gaussian), to non-normal generated from mixtures of several Gaussians, and skewed (14–16). Simulated background noise also features (17,18). Prior synthetic work approaches, however, have not explored

other possible characteristics specifically such as distance between clusters (both standard and rare), which is modelled in real data through the comparison of median fluorescence intensities between a stained and an unstained population in terms of population widths or standard deviations, in order to estimate the relative brightness of a fluorophore (19,20). Moreover, in a somewhat fragmented space, there is reason to apply systematic design on existing properties (such as the skewness of clusters) to optimise the coverage of characteristics.

Evaluation of developers' own tools using internally generated synthetic datasets is inherently biased, therefore external and independent testing is a prerequisite for software credibility in the clinical and biomanufacturing communities. Benchmarking datasets are used in independent studies to compare software performance, however, existing studies performed have solely relied on experimental data toolsets and have not used synthetic datasets (21,22). This may be related to a limited amount of synthetic datasets available within public flow cytometry repositories for the community to use (23). Software benchmarking studies holds similarities to other quality assurance methods, such as those applied in proficiency testing defined in ISO 13528:2005 (24), and similar statistical methods can be used to evaluate software output performances in an external and independent manner. When using real datasets as the test material, determination of the software performance is achieved through comparison of software results against an estimate of the true value. This value is assigned through a choice between 1) formulation, 2) cellular certified reference materials (of which very few exist for flow cytometry) (25), 3) manually gated analysis from one expert, 4) consensus manual analysis results from a group of experts, or 5) consensus values from participant results. Possible bias from the results of experts or participants reduces the robustness of the test. Synthetic datasets can be used adjacent to certified reference materials with potential benefit.

In this research we propose the use of synthetic datasets for benchmarking unsupervised learning automated flow cytometry data analysis software of which there are a large array of options available to the data analyst (26). We define a description of the data characteristics of flow cytometry data and demonstrate two methods to generate highly controlled, systematically designed synthetic datasets with different degrees of separation between clusters, and different levels of skew. We illustrate the use of our synthetic datasets using an exemplar software, SPADE3 (27), and present results that allow robust calculations of performance metrics not possible with real cell data. This work starts to explore the role of synthetic datasets as digital reference materials and standards, and the potential regulatory implications as the biomedical and biomanufacturing fields move increasingly towards using automated systems, machine learning and artificial intelligence techniques.

# 2 Materials and Methods

## 2.1 Target characteristics for synthetic flow cytometry datasets

We identified certain commonly recognised data characteristics or potential statistical attributes of flow cytometry data and put forward a strategy to control and modify these characteristics to create systematic scenarios for testing software (Table I). In this research we targeted the separation / overlap and the skew properties in our simulation studies because these had not been addressed in previous work and/or the designs had not been approached in a systematic way. In order to focus on these properties, non-target characteristics such as the number of clusters, number of datapoints, and number of dimensions were kept constant, and noise was excluded in our simulations.

## 2.2 Description of the Separation Index

The separation index (SI) is used throughout this research to define the distance between clusters. The SI measures the magnitude of the gap between a pair of clusters based on the upper and lower percentiles of the two clusters (28). In the one-dimensional example (Figure 1), the SI can be summarised as Eq. 1:

$$SI = \frac{L_2(\alpha/2) - U_1(\alpha/2)}{U_2(\alpha/2) - L_1(\alpha/2)} \quad \text{Eq. 1}$$

where $L_i(\alpha/2)$ and $U_i(\alpha/2)$ are the sample lower and upper $(\alpha/2)$ quantiles of cluster $i$. The interpretation of the SI is relatively straight forward, the range is [-0.999, +0.999] with values approaching +1 indicating increasing separation, SI of 0 indicating clusters touching, and SI approaching -1 indicating total overlap. In practice, our working range for the SI was [-0.3, +0.3]. These limits were defined because at a SI of +0.3 clusters were already very well separated, and at a SI of -0.3 clusters appeared well overlapped or merged.

## 2.3 Hardware and software

Dataset generation and analysis was run on a 64-bit Windows 10 operating system with a 3.00 GHz processor and 64 GB of RAM. Computational tools used are listed in Table II. Throughout this paper, we use regular type to refer to software or computing environments, *italics* for packages, and `monospace` font to designate functions.

## 2.4 Synthetic datasets

The concept of creating artificial, computer-generated flow cytometry datasets is essentially random number generation, with numbers typically drawn from a normal distribution. Other probability distributions are available e.g. binomial, exponential, Poisson, Student's t, etc. If flow cytometry data are considered as mixtures of subpopulations of a heterogenous sample, then the generation of synthetic data is a process of creating a mixture of random clusters.

### 2.4.1 Separation dataset generation

We designed a library of two-cluster synthetic datasets in two dimensions with 1,000 datapoints per cluster as an exemplar size of cell populations in real flow cytometry data, with different degrees of separation between neighbouring clusters ranging from well-separated to merged. The datasets were prepared using the R *clusterGeneration* package, with SI values ranging from [-0.3, +0.3] at 0.05 intervals. Nine random normally distributed cluster replicates were generated at each SI value. Covariance matrices were randomly generated from eigenvalues between 1 and 5 to give a variability in the diameter and shape of clusters that is similar to those seen in real flow cytometry data. These parameters produced clusters with known separation, but which were random in their elliptical shape attribute. Datasets were converted to FCS3.1 format using the R package *flowCore*.

### 2.4.2 Skew dataset generation

We designed a library of two-cluster synthetic datasets in two dimensions with 1,000 datapoints per cluster, with different levels of skew and skew-direction pairs. Single skew clusters were prepared with the package *sn*, of which the α parameter regulates asymmetry (29). Likewise random cluster replicates were generated at each skew direction (left and right) along the x-axis in addition to each level of skew input α values between 2.5 to 10, at intervals of 2.5. Applying the skewing α parameter causes the diameter of the elliptical cluster to reduce along the x-axis. To compensate for this, clusters were elongated to obtain a pre-skew diameter using the package *rescale*. The skewness of the clusters before and after rescaling were identical (measured using the package *psych*) determined by the asymmetry around the mean remaining unchanged (Figure 2A). Two clusters were joined together, and one cluster shifted further away from the other through vector arithmetic operations in R (Figure 2B). The distance between two clusters was measured with the *clusterGeneration* package, datasets with a SI value between -0.25 and -0.15 were selected for further processing. Files were converted to FCS3.1 standard using *flowCore* and visualised within FlowJo software.

## 2.5 Real datasets

All material was obtained with the approval of and in accordance with the respective Ethics Committees of Loughborough University and LGC, and under jurisdiction of the Human Tissue Authority.

### 2.5.1 Real cell PBMC dataset 1

Fresh whole blood from healthy donors (Cambridge Bioscience, UK) was processed using Ficoll-Paque (Fisher) to isolate the buffy coat layer containing peripheral blood mononuclear

cells (PBMCs). Cells were single-stained separately with CD4-PerCP-Vio700, CD45RO-APC-Vio770, and CCR7-VioBlue (all from Miltenyi Biotech). Data were acquired using BD FACSCantoII cytometer equipped with 3 lasers (405nm/ 30mW, 488nm/ 20mW, 633nm/ 17mW). 100,000 cell events were collected.

### 2.5.2 Real cell PBMC dataset 2

PBMCs (LGC, UK) were stained with CD3-BB515, CD4-BB700, CD45RA-BV786 (all from BD Biosciences), and live/dead fixable aqua dead cell stain (Invitrogen). Data were acquired using a BD LSRFortessa cell analyser equipped with four lasers (355nm /20mW, 405nm/ 50mW, 488nm/ 50mW, 640nm/ 40mW). 200,000 cell events were collected. Single-stained beads and fluorescence-minus-one controls were used to calculate compensation.

## 2.6 SPADE3 analysis of synthetic datasets

SPADE3 was run within Matlab R2019a. Each FCS file was run separately. User input parameters that were selected were: overlapping markers used for SPADE tree = CH1, CH2; ignore compensation; no transformation. All other settings were left as default values (local density neighbourhood size = 5, local density approximation factor = 1.5, maximum allowable cells = 50,000, outlier density = 1, target density = 20,000 cells, algorithm = K-means, number of desired clusters = 100).

## 2.7 Statistics and performance metrics

Methods used for statistical analysis included the mean, standard deviation, coefficient of variation, and metrics derived from the confusion matrix as shown in Table III.

The performances of software runs were calculated using Eq. 2 and Eq. 3:

$$Absolute\ difference\ to\ reference\ count = |A - B| \quad \text{Eq. 2}$$

$$Population\ percentage\ difference\ (\%) = \frac{|A - B|}{Total\ population} \times 100 \quad \text{Eq. 3}$$

where *A* is the software output count, and *B* is the reference value defined here as the known count of cluster 1 (1,000 events) which was designed inherently in the dataset.

As in binary classification (30), here a true positive (a 'hit') is defined as the correct SPADE3 assignment of a target cell to its reference target population set during cluster generation. Events in cluster 1 of the synthetic datasets were arbitrarily selected as the 'target' cases. SPADE3 assignment of a non-target cell to its non-target population is a true negative. Misclassification of a non-target cell to a target population is a false positive, and misclassification of a target cell to a non-target population is a false negative (a 'miss'). The evaluation metrics calculated from the confusion matrix include the accuracy, precision, recall and F1 measure, and are defined in Eq. 4 to Eq. 7 (31). We compared the individual cell assignments to a cluster predicted from SPADE3 with the reference cell assignments, using the R package *caret*.

$$Accuracy = \frac{True\ positive + True\ negative}{Total\ population} \quad \text{Eq. 4}$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad \text{Eq. 5}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad \text{Eq. 6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Eq. 7}$$

266

# 3 Discussion

In flow cytometry, data typically contain cell populations that are positive or negative for a marker of interest. The distance between the positive and negative cell populations is variable, ranging from well resolved to merged. Multiple factors affect this separation, including but not limited to biological attributes (the level of marker expression, affinity and avidity of antibody binding, the number of antibody bound per cell) and assay variables (antibody panel design and concentrations used in the staining process, fluorophore brightness and dye stability, sensitivity and resolution of the detectors). This separation directly impacts the accuracy and precision of manual data analysis, with significantly higher technical variation seen in poorly resolved populations compared with clearly defined cell populations in human peripheral blood (32).

## 3.1 Distance between clusters

To simulate flow cytometry data with different distances between a positive and a negative population we generated a normally distributed synthetic two-cluster dataset with different degrees of separation between clusters. For comparison, we measured the SI between two clusters in a real cell dataset. The PBMCs dataset 1 contained negatively and positively stained populations in each fluorescent channel. These subpopulations were separated using the automated cell population identification software SPADE3 (27). Then the magnitude of the gap between pairs of real cell clusters along each channel was measured using the `sepIndex` function in the *clusterGeneration* package. We observed similar SI values between the real-world positive and negatively stained cell populations and the synthetic clusters, within the range of -0.3 and +0.2 (Figure 3). These results show our method defined here for generating synthetic flow cytometry datasets is able to successfully simulate the distance parameters seen between clusters in a real example of flow cytometry data.

Individual cluster statistics can potentially vary by a small amount, however at this stage of the research this was not a focus of this study.

### 3.2 Clusters with non-normal distributions

The synthetic datasets generated in Section 3.1 contain clusters following a normal distribution, visualised as symmetrical bell curves for univariate data or symmetrical circles and ellipses in scatterplots for multivariate data. However, real flow cytometry data consist of cell population clusters that follow a normal distribution as well as those that display non-normal distributions. The exact distribution along a marker channel is difficult to predict and may depend on the state of the cell along a differentiation pathway. For example, a stable haematopoietic stem cell population may display a normal distribution of CD34+ expression that transitions to a non-normal distribution during cell differentiation as CD34 expression decreases (33).

Non-normal data are characterised by asymmetry around the sample mean. These cell populations can display positive (right) skew or negative (left) skew. The skewness can be estimated using the adjusted Fisher-Pearson coefficient of skewness ($p$ value) (34), where a normal distribution has a skewness value of $p = 0$, a positive skewness value indicates a tail pointing to the right, and a negative skewness value indicates a tail pointing to the left. The further away the value is from 0, the greater the skew and typically the longer the tail.

There are different strategies to generate synthetic flow cytometry datasets with skewed cell populations. One method is to create multiple Gaussian distributions that can then be merged together to form an overall distribution with the desired skew. This strategy has been used previously to create synthetic data with non-convex shapes (15,16). This method may require many rounds of trial and error. To avoid this shortcoming, here we developed and tested a

different method using the *sn* R package to generate random clusters with multivariate skew-normal distributions (29).

The multivariate skew normal distribution extends the class of normal distribution (defined through a mean vector and covariance matrix) by the addition of a skew parameter.

A comparison of the skewed clusters generated from computer simulation against real cell populations from the PBMCs dataset 2 demonstrates that the synthetic and real cases are comparable (Figure 4). This result shows that the simulated data we generated is a realistic model of both positive and negative skew observed in real flow cytometry cell populations, and therefore has biological relevance. Thus, the synthetic dataset can be reliably used to gain understanding on how automated software responds to skewed flow cytometry data, with the additional benefit of the ability to systematically control the strength of the skew as well as the absolute cell number.

The strategy we devised to create a dataset with multiple skewed clusters was to generate individual clusters in parts then combine them together to form one whole dataset. The gap between the clusters can be controlled by shifting one cluster closer or further away to the other through vector arithmetic operations, with the SI being measured after the clusters were combined. With skewed clusters, a new level of complexity is introduced compared to normally distributed clusters, because assuming the skew is introduced only in one parameter, each cluster can be left-skewed or right-skewed. Thus, the possible permutations of pairs of skewed clusters in a two-cluster dataset increases from one to three. In this paper we refer to these combinations as: head-to-head, head-to-tail (this is the same orientation as tail-to-head), and tail-to-tail (Figure 5).

### 3.3 Application of synthetic datasets to a cell population identification software

To demonstrate the efficacy of the synthetic datasets, they were passed through an exemplar software, SPADE, in order to illustrate how synthetic data can reveal limitations of automated software, and can provide a deeper understanding of the inner workings of 'black box' algorithms in a way that real cell datasets are unable to.

SPADE, named for spanning-tree progression analysis for density-normalised events, is a widely applied software package that uses automated down-sampling, clustering and minimum spanning tree construction to aid analysis of high-dimensional flow cytometry data (35). There are two versions of SPADE with different algorithms. The original SPADE1 applies a stochastic down-sampling algorithm paired with an agglomerative hierarchical clustering algorithm that produces different outputs when run on the same data. This specific issue of reproducibility in SPADE1 was subsequently resolved in SPADE3 by removing the stochastic algorithms and replacing them with deterministic ones (27). In addition, a tree partitioning function was introduced to assist interpretation of the outputs.

Here, we used the Matlab implementation of SPADE3 to process our synthetic datasets using default parameters (as described in section 2.6). We used the auto tree partitioning tool to split the spanning tree into two populations, then compared the population number to our known reference value, which was 1,000 cells per cluster, or 50% of total cells events, for both the separation and skew datasets.

For the separation dataset, the absolute difference in cell count of each cluster between the software output and the reference value was calculated for each SI condition. The results show that the accuracy and precision of SPADE3 decreased as the SI decreased from +0.3 to -0.3, with performance deteriorating noticeably at a SI value of -0.2 and below (Figure 6).

These results were to be expected, because defining the boundary between one cluster and another becomes progressively more difficult as clusters get closer together.

The benefit of applying the synthetic datasets to test software such as SPADE3 was the ability to quantify for the first time the SI value where the software began to lose performance. The high level of control in designing the gap between cell populations within the synthetic datasets would have been very difficult to achieve with real cell data. Furthermore, since the absolute counts and frequencies of each cell population was known in the synthetic dataset, the evaluation of the software was based on robust absolute traceable figures, and did not rely on comparison with a manually gated reference subpopulation count, which has already been shown to be operator dependent and potentially biased (36,37).

In the design of the skew dataset, a constant SI value of -0.2 between clusters was chosen because it fell in the critical region where the SPADE3 software began to deteriorate. The skew dataset was processed through SPADE3, then the difference in cell population percentage of the cluster between the SPADE3 output and the reference value was calculated for each skew condition and cluster pair orientation. We found that, for each cluster pair orientation, increasing the level of skewness in the clusters had no effect on the accuracy and precision of SPADE3. However, at each level of skewness, SPADE3 was able to partition the two clusters with improved performance when the orientation was tail-to-tail, followed by head-to-tail and finally head-to-head (Figure 7).

This pattern of performance appeared to correlate with the density of points between the two clusters. The skew dataset was planned with skewness and skew orientation among the variable design factors, and the separation between clusters as constant factors. The systematic way this skew dataset was designed allowed for the pattern of behaviour of

SPADE3 to become apparent. This finding suggests the SPADE3 algorithm is well suited to analysis of skewed data albeit with a performance bias and sensitivity depending on skewed cluster orientation. This may not be the case for other algorithms that use different clustering techniques, in particular those that use a Gaussian mixture model-based clustering approach. Further work to investigate this in a software comparison study is warranted.

### 3.4 Assessment of software performance based on synthetic data

One of the benefits of synthetic data is that, as well as 'true' population counts and frequencies, an estimate of the true membership of a cell to its cluster is known *a priori*. This is not the case with real cell data, where membership of a cell to a population is estimated by an analyst performing manual gating. Here, we demonstrate the evaluation of robust performance metrics of SPADE3 runs on synthetic datasets using confusion matrix analysis.

Each cell event in the synthetic dataset was pre-assigned a cluster membership on generation. These cluster memberships were withheld for the SPADE3 analyses. After running the datasets through SPADE3, the software predictions of cluster memberships for all 2,000 cell events were compared with the reference cluster memberships using the R *caret* package. Events in cluster 1 were arbitrarily assigned as positive cases.

The results from the SPADE3 analysis of the synthetic separation dataset (Table IV) showed a classification accuracy (Eq. 4) greater than 90% with SI values of -0.1 or greater. Accuracy fell to 86% and 43% at SI values of -0.2 and -0.3 respectively (Figure 8A).

The same pattern appeared with precision, also called positive predictive value (Eq. 5) with values greater than 90% at SI values of -0.1 or larger, then falling to 81% and 50% at SI values of -0.2 and -0.3 respectively (Figure 8B).

The recall metric that measures the rate of true positives identified (Eq. 6) gave perfect scores of 100% at SI of +0.3, scored greater than 90% at SI values of -0.2 or greater, and deteriorated to 46% with SI of -0.3 (Figure 8C).

The F1 score (Eq. 7) was calculated from the precision and accuracy to give the overall accuracy of SPADE3. The F1 score remained above 90% for SI values greater than -0.1, then fell to 85% at a SI value of -0.2 and reduced further to 47% at a SI value of -0.3 (Figure 8D).

The results from the classification analysis reinforce the finding that SPADE3 performs strongly when clusters are well-separated, but the accuracy decreases as clusters approach a SI value of -0.2 and falls to below 50% when processing overlapping clusters with SI values of -0.3. Our application of this synthetic flow cytometry dataset in this instance has helped to illustrate good performance characteristics and ranges, but also limitations of SPADE3 with respect to cluster separation both with normal and skewed cluster probability distributions.

# 4 Conclusion

In this article we have introduced a systematic method of designing and generating synthetic flow cytometry datasets, with specific focus on control of the distance between clusters and the probability distributions of events within clusters. We applied our computer-generated flow cytometry datasets to an automated data analysis software, SPADE3, and have shown that the synthetic datasets are capable of critically assessing the quality of the software outputs and hence the software performance. In addition, we have given an example of quantifying performance assessment using synthetic datasets that is robust compared with using real-world datasets.

The systematic approach we have implemented to produce flow cytometry datasets is straight forward to execute computationally, but would be complicated to achieve experimentally due to uncontrollable external sources of variation within real cell datasets, thus synthetic datasets here overcome the limitations of acquiring real datasets. The synthetic datasets have the same range of data properties as their biological equivalents and can serve as credible substitutes for real flow cytometry datasets for the testing of automated cell population identification software.

It is noted here that in most cases algorithms that underpin flow cytometry analysis software have been previously published, but it is (at times) inherently opaque implementation of algorithms in executable code that whilst allows understanding of inputs and outputs, does not allow a full understanding of the data transfer functions. Application of our synthetic datasets to an automated cell population identification software such as SPADE3 can

therefore help users understand how the underlying ‘black box’ algorithm works. Here we have identified the regions where SPADE3 begins to lose performance, specifically where two clusters are located at a SI value of -0.2 or less. Our results suggest that SPADE3 is not specifically affected by the probability distributions of data, but is more sensitive to the relative density of data points between two clusters. Findings such as these can provide guidance to users on software selection when having to contend with large array of potential software solutions (26) (i.e. in this exemplar whether SPADE3 would be an appropriate tool of choice for automated analysis of real data containing heavy overlapping of clusters), and then help to understand the validity of their automated analysis outputs.

A further benefit of synthetic datasets was apparent when we assessed the performance of SPADE3 using the metrics calculated from the confusion matrix. This classification analysis relied on comparison of software predictions with known ‘true’ conditions hence an absolute analysis. In the synthetic dataset, an estimate of the true assignment of cells to clusters was designed into the data, making this analysis relatively rapid and robust. With real data, the assignment of cells to subpopulations must first be manually determined (often with potential difficulty and error), then the dataset labelled, before a confusion matrix can be calculated. These additional steps are time-consuming and prone to error. The variability observed in manually gated datasets means either the analysis from a single expert must be taken as the best estimate of the ‘true value’, or a pooled manual analysis from a group of experts is used. The first option risks bias, and the second is dependent on the precision and accuracy of the group. In both instances, it is difficult for the final analysis to be as robust compared with synthetic datasets.

There are also a few disadvantages with synthetic datasets. The datasets are built on assumptions of real data. These approximations are based on mathematical models and have limits. Although the aim is to create synthetic datasets that are as realistic as possible, there may be features missing as a function of boundary conditions and design assumptions.

Further investigations on flow cytometry synthetic datasets will follow two main directions. The first is the extension of the work to generate datasets with controls on other flow cytometry data properties identified in Table I, and more complex datasets with multiple controlled factors. We are currently developing and optimising synthetic datasets with rare cell populations with and without skewed distributions, and noise characteristics. The second area is the comparison of software performance across multiple platforms when challenged with synthetic datasets targeting flow cytometry automated data analysis software that employ various clustering algorithms such as K-means, hierarchical, partition, density-based, model-based, spectral clustering and self-organising maps.

Besides being a benchmarking tool for software developers, possible further applications of synthetic datasets include their use as educational and training tools for manual gating, as part of external quality assessment (EQA) and proficiency testing schemes. In addition, as cell identification and quantification in medical diagnostics and cell therapy/ regenerative medicines manufacturing fields move increasingly towards automated machine learning and artificial intelligence techniques, it is likely that synthetic datasets will have important regulatory applications as digital reference materials and standards, as well as potential regulatory implications.

## Acknowledgements

The authors would like to thank Lydia Beeken and Dr Ben Diffey from the Centre for Biological Engineering, Loughborough University, and Dr Shiqiu Xiong at LGC, for their help with acquiring real flow cytometry datasets.

## Conflict of Interest Declaration

The authors declare that they have no competing interests.

# References


(1) Phillips, W.; Medcalf, N.; Dalgarno, K.; Makatoris, H.; Sharples, S.; Srai, J.; Hourd, P.; Kapletia, D. *Redistributed Manufacturing in Healthcare: Creating New Value through Disruptive Innovation*; 2019.

(2) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (Oct), 2825–2830.

(3) Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A. M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016**, 3234–3243. https://doi.org/10.1109/CVPR.2016.352.

(4) Wrenninge, M.; Unger, J. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. *arXiv Preprint.* **2018**. https://arxiv.org/abs/1810.08705 (accessed 2020-11-20)

(5) Tremblay, J.; To, T.; Birchfield, S. Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* **2018**, 2038–2041. https://doi.org/10.1109/CVPRW.2018.00275.

(6) Hagiwara, A.; Warntjes, M.; Hori, M.; Andica, C.; Nakazawa, M.; Kumamaru, K. K.; Abe, O.; Aoki, S. SyMRI of the Brain: Rapid Quantification of Relaxation Rates and Proton Density, with Synthetic MRI, Automatic Brain Segmentation, and Myelin Measurement. *Invest. Radiol.* **2017**, *52* (10), 647–657. https://doi.org/10.1097/RLI.0000000000000365.

(7) Ratanaprasatporn, L.; Chikarmane, S. A.; Giess, C. S. Strengths and Weaknesses of Synthetic Mammography in Screening. *Radiographics* **2017**, *37* (7), 1913–1927. https://doi.org/10.1148/rg.2017170032.

(8) Niazi, M. K. K.; Parwani, A. V.; Gurcan, M. N. Digital Pathology and Artificial Intelligence. *Lancet Oncol.* **2019**, *20* (5), e253–e261. https://doi.org/10.1016/S1470-2045(19)30154-8.

(9) Perez, L.; Wang, J. The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. *arXiv Preprint.* **2017**. https://arxiv.org/abs/1712.04621 (accessed 2021-01-05).

(10) Arvaniti, E.; Claassen, M. Sensitive Detection of Rare Disease-Associated Cell Subsets via Representation Learning. *Nat. Commun.* **2017**, *8* (1), 1–10. https://doi.org/10.1038/ncomms14825.

(11) The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC. *Off. J. Eur. Union* **2016**, 1–88.

(12) Sugar, I. P.; Sealfon, S. C. Misty Mountain Clustering: Application to Fast Unsupervised Flow Cytometry Gating. *BMC Bioinformatics* **2010**, *11* (1), 1–14.

https://doi.org/10.1186/1471-2105-11-502.

(13) Samusik, N.; Good, Z.; Spitzer, M. H.; Davis, K. L.; Nolan, G. P. Automated Mapping of Phenotype Space with Single-Cell Data. *Nat. Methods* **2016**, *13* (6), 493–496. https://doi.org/10.1038/nmeth.3863.

(14) Naim, I.; Datta, S.; Rebhahn, J.; Cavenaugh, J. S.; Mosmann, T. R.; Sharma, G. SWIFT-Scalable Clustering for Automated Identification of Rare Cell Populations in Large, High-Dimensional Flow Cytometry Datasets, Part 1: Algorithm Design. *Cytom. Part A* **2014**, *85* (5), 408–421. https://doi.org/10.1002/cyto.a.22446.

(15) Pyne, S.; Hu, X.; Wang, K.; Rossin, E.; Lin, T.-I.; Maier, L. M.; Baecher-Allan, C.; McLachlan, G. J.; Tamayo, P.; Hafler, D. A.; De Jager, P. L.; Mesirov, J. P. Automated High-Dimensional Flow Cytometric Data Analysis. *Proc. Natl. Acad. Sci.* **2009**, *106* (21), 8519–8524. https://doi.org/10.1073/pnas.0903028106.

(16) Ge, Y.; Sealfon, S. C. Flowpeaks: A Fast Unsupervised Clustering for Flow Cytometry Data via K-Means and Density Peak Finding. *Bioinformatics* **2012**, *28* (15), 2052–2058. https://doi.org/10.1093/bioinformatics/bts300.

(17) Zare, H.; Shooshtari, P.; Gupta, A.; Brinkman, R. R. Data Reduction for Spectral Clustering to Analyze High Throughput Flow Cytometry Data. *BMC Bioinformatics* **2010**, *11* (1), 1–16. https://doi.org/10.1186/1471-2105-11-403.

(18) Bendall, S. C.; Davis, K. L.; Amir, E. A. D.; Tadmor, M. D.; Simonds, E. F.; Chen, T. J.; Shenfeld, D. K.; Nolan, G. P.; Pe'Er, D. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **2014**, *157* (3), 714–725. https://doi.org/10.1016/j.cell.2014.04.005.

(19) Bigos, M. Separation Index: An Easy-to-Use Metric for Evaluation of Different Configurations on the Same Flow Cytometer. *Curr. Protoc. Cytom.* **2007**, *40* (1), 1–21. https://doi.org/10.1002/0471142956.cy0121s40.

(20) Telford, W. G.; Babin, S. A.; Khorev, S. V.; Rowe, S. H. Green Fiber Lasers: An Alternative to Traditional DPSS Green Lasers for Flow Cytometry. *Cytom. Part A* **2009**, *75* (12), 1031–1039. https://doi.org/10.1002/cyto.a.20790.

(21) Aghaeepour, N.; Finak, G.; Hoos, H.; Mosmann, T. R.; Brinkman, R.; Gottardo, R.; Scheuermann, R. H.; Gottardo, R.; Scheuermann, R. H. Critical Assessment of Automated Flow Cytometry Data Analysis Techniques. *Nat. Methods* **2013**, *10* (3), 228–238. https://doi.org/10.1038/nmeth.2365.

(22) Weber, L. M.; Robinson, M. D. Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. *Cytom. Part A* **2016**, *89* (12), 1084–1096. https://doi.org/10.1002/cyto.a.23030.

(23) Spidlen, J.; Breuer, K.; Rosenberg, C.; Kotecha, N.; Brinkman, R. R. FlowRepository: A Resource of Annotated Flow Cytometry Datasets Associated with Peer-Reviewed Publications. *Cytom. Part A* **2012**, *81* (9), 727–731. https://doi.org/10.1002/cyto.a.22106.

(24) International Organization for Standardization. *ISO 13528:2005 Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison*; Geneva, 2005.

(25) Wang, L.; Abbasi, F.; Ornatsky, O.; Cole, K. D.; Misakian, M.; Gaigalas, A. K.; He, H. J.; Marti, G. E.; Tanner, S.; Stebbings, R. Human CD4 + Lymphocytes for Antigen

Quantification: Characterization Using Conventional Flow Cytometry and Mass Cytometry. *Cytom. Part A* **2012**, *81 A* (7), 567–575. https://doi.org/10.1002/cyto.a.22060.

(26) Cheung, M.; Campbell, J. J.; Whitby, L.; Thomas, R. J.; Braybrook, J.; Petzing, J. N. Current Trends in Flow Cytometry Automated Data Analysis Software. *Cytom. Part A* **2021**, *99* (10), 1007–1021. https://doi.org/10.1002/cyto.a.24320.

(27) Qiu, P. Toward Deterministic and Semiautomated SPADE Analysis. *Cytom. Part A* **2017**, *91* (3), 281–289. https://doi.org/10.1002/cyto.a.23068.

(28) Qiu, W.; Joe, H. Separation Index and Partial Membership for Clustering. *Comput. Stat. Data Anal.* **2006**, *50* (3), 585–603. https://doi.org/10.1016/j.csda.2004.09.009.

(29) Azzalini, A.; Capitanio, A. Statistical Applications of the Multivariate Skew Normal Distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61* (3), 579–602. https://doi.org/10.1111/1467-9868.00194.

(30) Fleiss, J. L.; Levin, B.; Paik, M. C. *Statistical Methods for Rates and Proportions*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2003.

(31) Tharwat, A. Classification Assessment Methods. *Appl. Comput. Informatics* **2021**, *17* (1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003.

(32) Burel, J. G.; Qian, Y.; Lindestam Arlehamn, C.; Weiskopf, D.; Zapardiel-Gonzalo, J.; Taplitz, R.; Gilman, R. H.; Saito, M.; de Silva, A. D.; Vijayanand, P.; Scheuermann, R. H.; Sette, A.; Peters, B. An Integrated Workflow to Assess Technical and Biological Variability of Cell Population Frequencies in Human Peripheral Blood by Flow Cytometry. *J. Immunol.* **2017**, *198* (4), 1748–1758. https://doi.org/10.4049/jimmunol.1601750.

(33) Salati, S.; Zini, R.; Bianchi, E.; Testa, A.; Mavilio, F.; Manfredini, R.; Ferrari, S. Role of CD34 Antigen in Myeloid Differentiation of Human Hematopoietic Progenitor Cells. *Stem Cells* **2008**, *26* (4), 950–959. https://doi.org/10.1634/stemcells.2007-0597.

(34) Joanes, D. N.; Gill, C. A. Comparing Measures of Sample Skewness and Kurtosis. *J. R. Stat. Soc. Ser. D (The Stat.)* **1998**, *47* (1), 183–189. https://doi.org/10.1111/1467-9884.00122.

(35) Qiu, P.; Simonds, E. F.; Bendall, S. C.; Gibbs, K. D.; Bruggner, R. V; Linderman, M. D.; Sachs, K.; Nolan, G. P.; Plevritis, S. K. Extracting a Cellular Hierarchy from High-Dimensional Cytometry Data with SPADE. *Nat. Biotechnol.* **2011**, *29* (10), 886–893. https://doi.org/10.1038/nbt.1991.

(36) Grant, R.; Coopman, K.; Medcalf, N.; Silva-Gomes, S.; Campbell, J. J.; Kara, B.; Braybrook, J.; Petzing, J. Understanding the Contribution of Operator Measurement Variability within Flow Cytometry Data Analysis for Quality Control of Cell and Gene Therapy Manufacturing. *Measurement* **2020**, *150*, 106998. https://doi.org/10.1016/j.measurement.2019.106998.

(37) Grant, R.; Coopman, K.; Medcalf, N.; Silva-Gomes, S.; Campbell, J. J.; Kara, B.; Braybrook, J.; Petzing, J. N. Quantifying Operator Subjectivity within Flow Cytometry Data Analysis as a Source of Measurement Uncertainty and the Impact of Experience on Results. *PDA J. Pharm. Sci. Technol.* **2021**, *75* (1), 33–47. https://doi.org/10.5731/pdajpst.2019.011213.

629

# Tables

Table I. Characteristics of flow cytometry datasets

| Characteristic | Description |
|---|---|
| Number of clusters | Number of cell subpopulations in a sample |
| Number of datapoints | Number of cell events acquired from a sample |
| Number of dimensions | Number of parameters recorded in the experiment, e.g. forward scatter, side scatter, fluorescent markers |
| Separation | The gap between negatively and positively stained cell populations |
| Overlap | Poorly resolved populations that appear merged |
| Placement | Projection direction of one cluster to another in space |
| Distribution | The shape of the cell population, as modelled on probability distributions e.g. Gaussian, Student's t, exponential, Chi-squared |
| Spread | The variance of the cell population |
| Skew and kurtosis | The level of asymmetry around the mean of the cell cluster |
| Orientation | The direction of the asymmetry |
| Elongation | Stretched out populations with long tails |
| Noise | Events that are excluded from analysis e.g. outliers, dead cells, debris, doublets, false events detected in the region of interest. |

Table II. Toolset used in this research for the generation of synthetic datasets, automated cell population identification, and performance evaluation. R packages in italics.

| Tool | Version | Purpose in this research |
|---|---|---|
| R | 3.5.1 | Programming |
| RStudio IDE | 1.2 | Programming environment |
| Matlab | R2019a | Environment for SPADE analysis |
| FlowJo | 10.6 | Flow cytometry data analysis and visualisation |
| SPADE | 3 | Automated analysis of synthetic datasets |
| *caret* | 6.0-82 | Calculate performance metrics, confusion matrix |
| *clusterGeneration* | 1.3.4 | Generate synthetic clusters |
| *flowCore* | 1.48.1 | Manipulate flow cytometry data |
| *psych* | 1.8.12 | Measure skew |
| *scales* | 1.1.0 | Scale functions for visualisation |
| *sn* | 1.5-3 | Build and manipulate probability distributions of the skew-normal family |
| *tidyverse* | 1.3.0 | Data manipulation, analysis and visualisation |

639 Table III. Confusion Matrix

| | | Reference | |
|---|---|---|---|
| | | **Target** | **Non-target** |
| **Predicted** | **Target** | True positive | False positive |
| | **Non-target** | False negative | True negative |

640

641

642 Table IV. Performance metrics.

| **Separation index** | **Mean Accuracy (%)** | **Mean Precision (%)** | **Mean Recall (%)** | **Mean F1 (%)** |
|---|---|---|---|---|
| -0.3 | 43.0 | 49.8 | 45.8 | 47.4 |
| -0.2 | 86.1 | 80.5 | 91.6 | 85.1 |
| -0.1 | 93.9 | 93.5 | 94.4 | 93.9 |
| 0 | 97.0 | 97.0 | 97.2 | 97.0 |
| 0.1 | 98.6 | 98.7 | 98.5 | 98.6 |
| 0.2 | 99.7 | 99.8 | 99.7 | 99.7 |
| 0.3 | 99.98 | 99.97 | 100.00 | 99.98 |

643

644

# Figure captions

Figure 1. One-dimensional example of the separation index (SI) that measures the magnitude of the gap between two clusters. Vertical lines indicate the lower and upper quantiles of the clusters. The difference between U1 and L2 (numerator) is divided by the difference between L1 and U2 (denominator) to calculate the SI value. This method is robust against outliers in between the two clusters that may affect the SI. Figure adapted from (28).

Figure 2. Workflow for skew dataset generation. Top panels show scatterplots, bottom panels show density estimates. A) A cluster with a normal distribution is generated, then skew is added through the alpha parameter in the R package *sn*, then the cluster is rescaled. B) Two clusters are combined, then the distances between them can be varied through vector arithmetic operations.

Figure 3. Comparison of synthetic and real clusters with representative separation index values ranging from -0.3, 0.0 to +0.2, showing overlapping, touching and well-separated clusters, respectively. Top panel shows synthetic data generated from R package *clusterGeneration*, bottom panel shows real peripheral blood mononuclear cells (PBMCs) data after automated population detection and partition with SPADE3 software followed by separation index calculation.

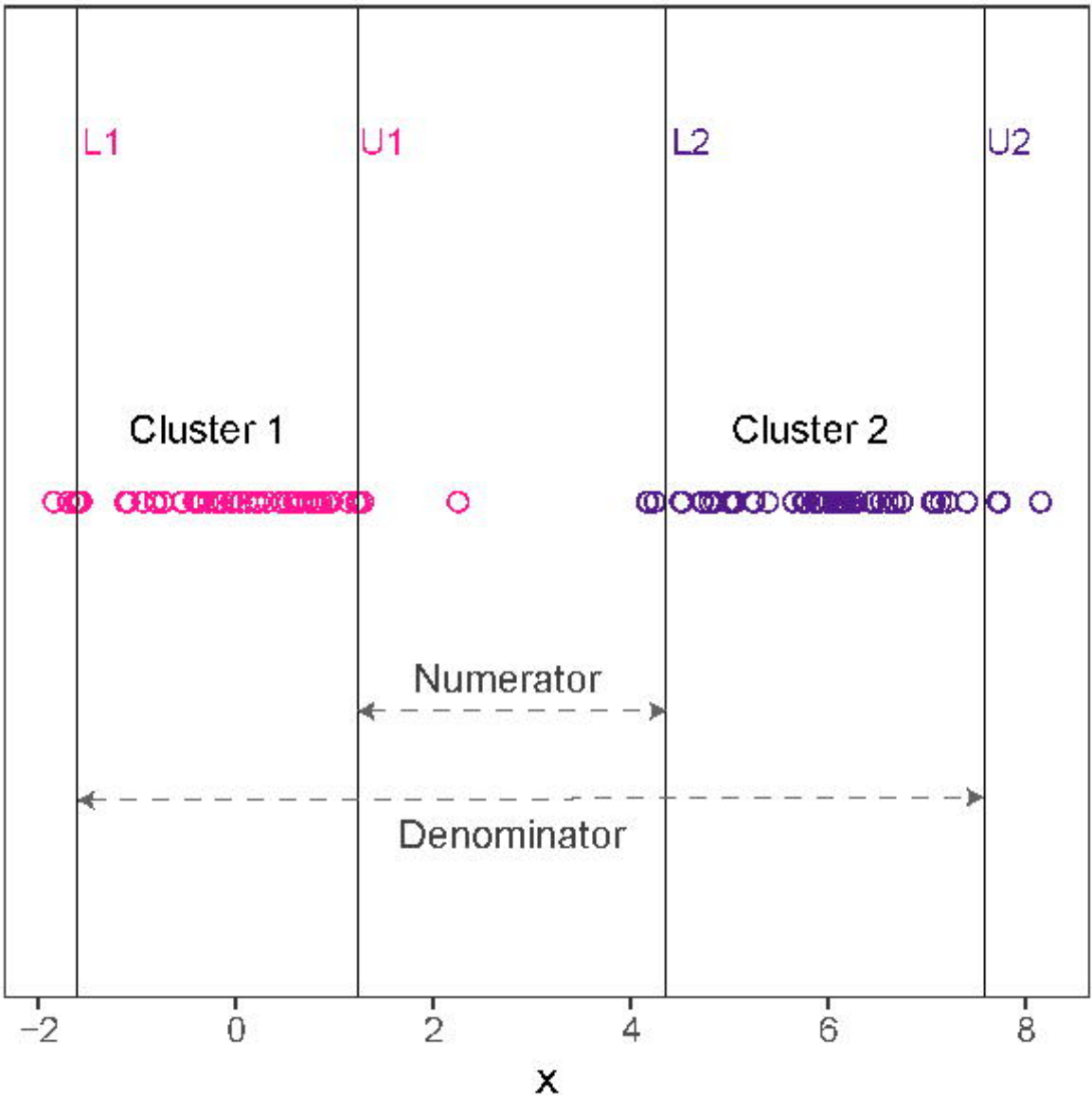
Figure 4. Comparison of synthetic (A) and real (B) flow cytometry data with skewed distributions. Both left-skewed and right-skewed synthetic clusters can be generated that mimic real data. Asymmetry around the mean is clearly shown with contour plots (top) and histograms (bottom).

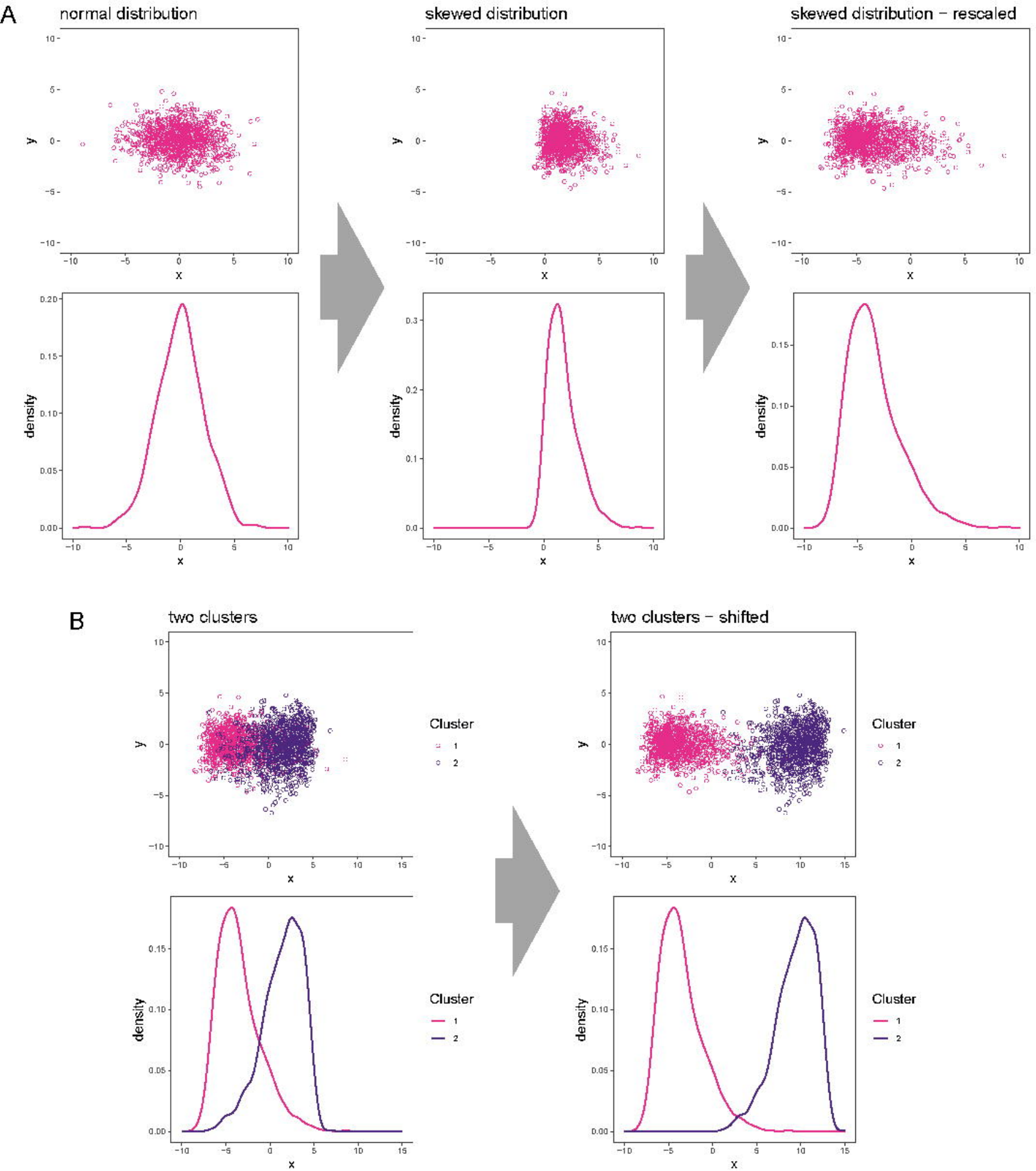
Figure 5. Generation of synthetic two-cluster skewed datasets. Three combinations of cluster pairs are shown here; head-to-head, head-to-tail and tail-to-tail.

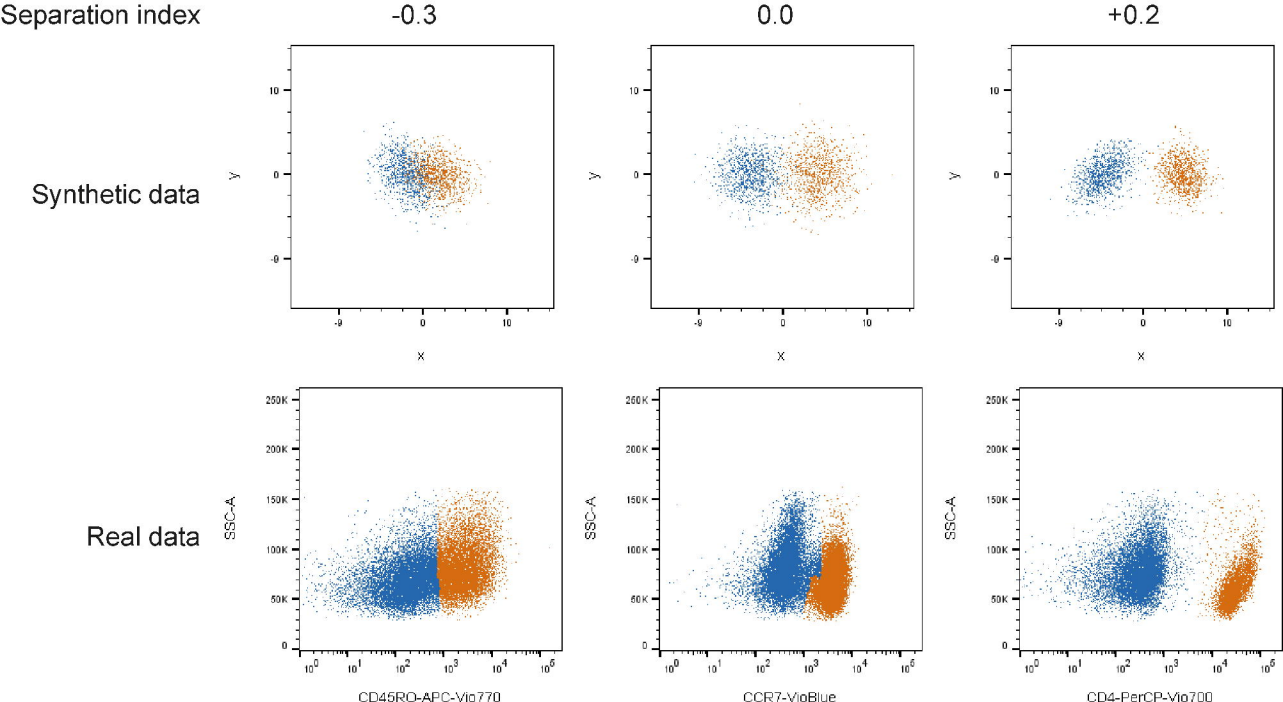
Figure 6. The synthetic two-cluster separation dataset was run through SPADE3, then the absolute difference between the SPADE3 cell count to the reference value of cluster 1 was calculated. Result demonstrates the accuracy and repeatability of SPADE3 deteriorates as the distance between clusters decreases. Data represents mean ±1SD.

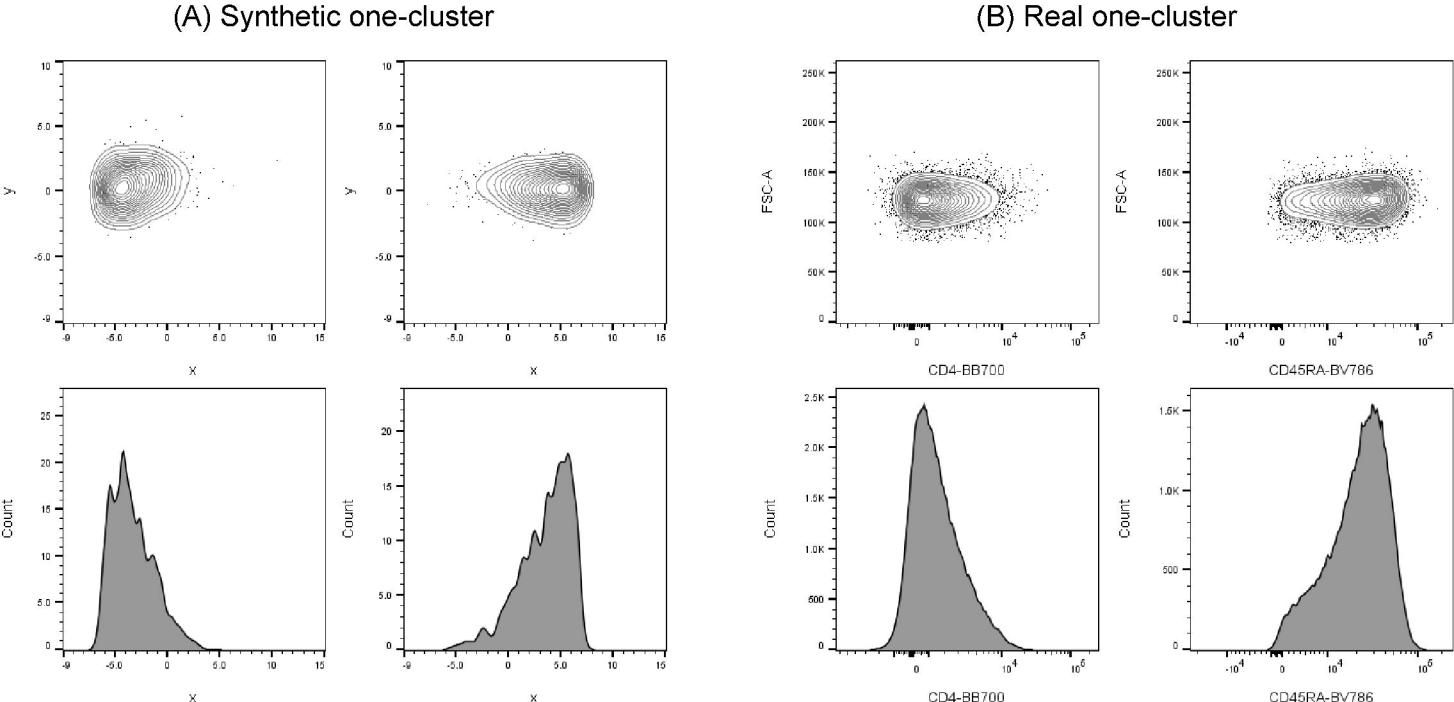
Figure 7. The synthetic two-cluster skewed dataset was run through SPADE3, then the gap between the SPADE3 cell population percentages and the reference cell population percentages was calculated. Increasing the cluster skewness in the datasets did not affect SPADE3 performance. However, the accuracy and repeatability of SPADE3 improved as direction of the cluster pairs changed from head-to-head to head-to-tail, and then to tail-to-tail. All clusters had separation index of -0.2. Data represents mean ±1SD.

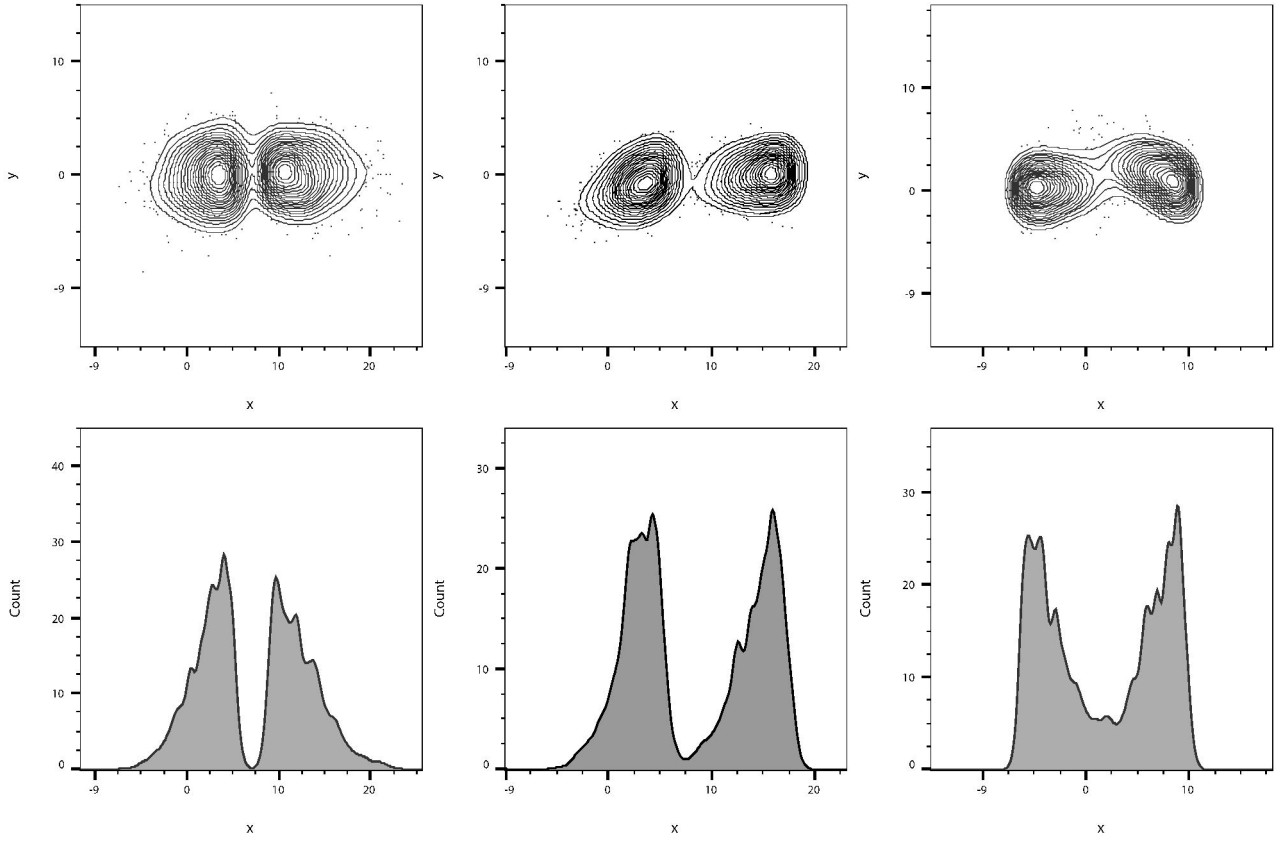
Figure 8. Performance metrics for SPADE3 analysis of Separation dataset.

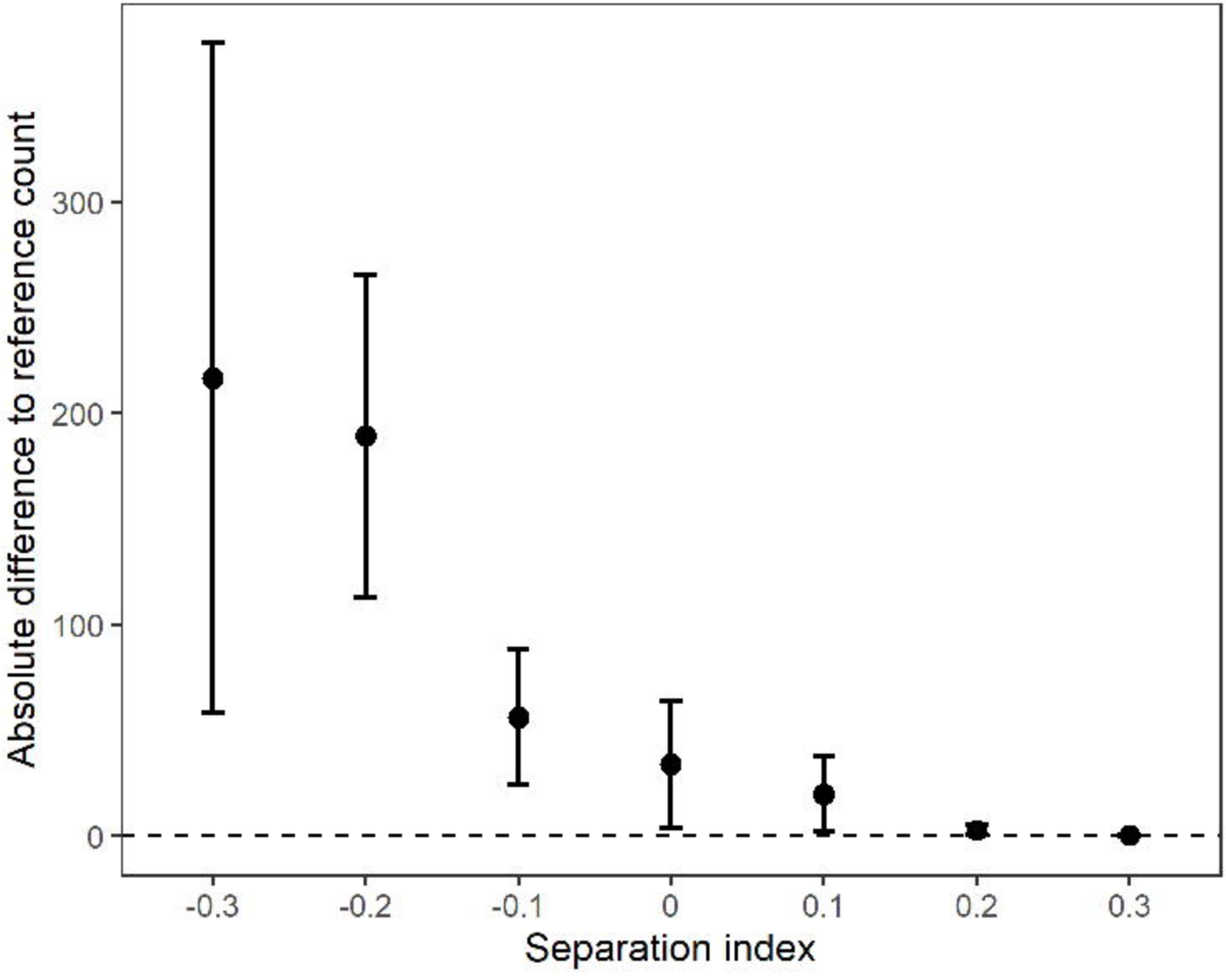

L1
U1
L2
U2
Cluster 1
Cluster 2
Numerator
Denominator
−2
0
2
4
6
8
x

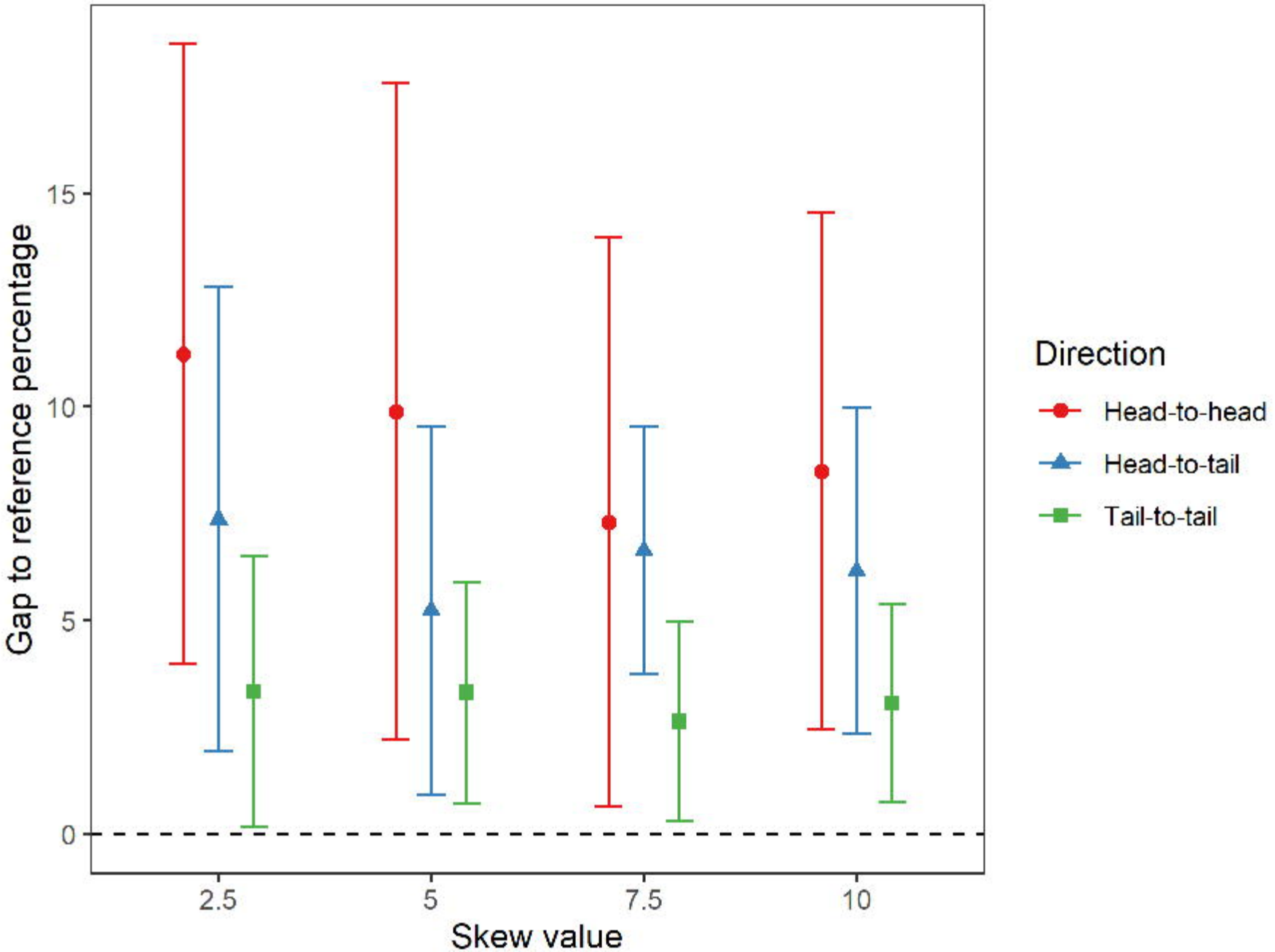

A
normal distribution
skewed distribution
skewed distribution − rescaled
y
x
density
B
two clusters
two clusters − shifted
Cluster
1
2

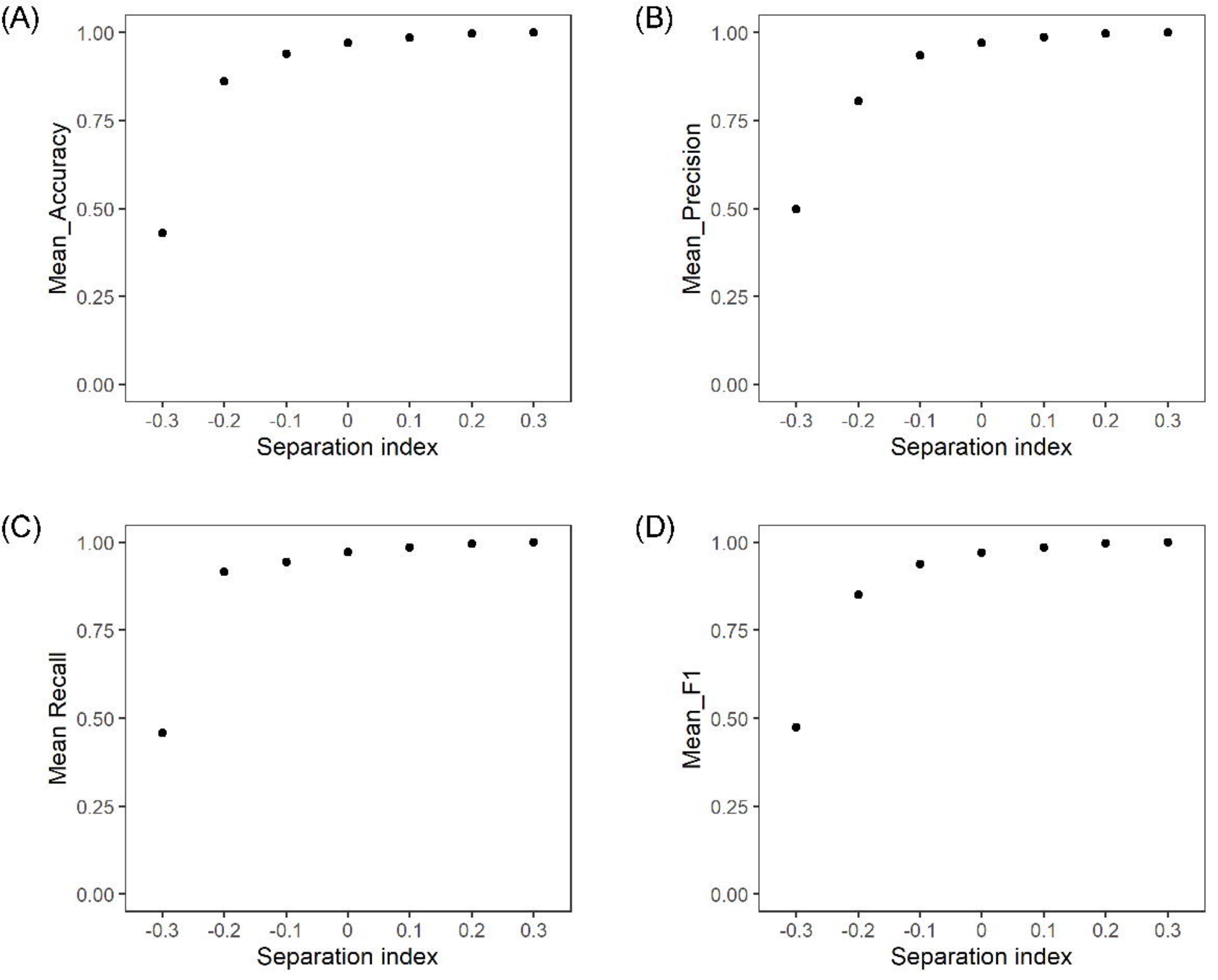

Separation index
-0.3
0.0
+0.2
Synthetic data
Real data
y
x
SSC-A
CD45RO-APC-Vio770
CCR7-VioBlue
CD4-PerCP-Vio700


(A) Synthetic one-cluster
(B) Real one-cluster
y
x
Count
FSC-A
CD4-BB700
CD45RA-BV786


y
x
Count
10
0
-9
20
30
40


Absolute difference to reference count
300
200
100
0
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Separation index


Gap to reference percentage
15
10
5
0
2.5
5
7.5
10
Skew value
Direction
Head-to-head
Head-to-tail
Tail-to-tail


(A)
Mean_Accuracy
1.00
0.75
0.50
0.25
0.00
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Separation index
(B)
Mean_Precision
1.00
0.75
0.50
0.25
0.00
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Separation index
(C)
Mean Recall
1.00
0.75
0.50
0.25
0.00
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Separation index
(D)
Mean_F1
1.00
0.75
0.50
0.25
0.00
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
Separation index

PDA Journal
of Pharmaceutical Science and Technology