Multivariate Data Analysis Methodology to Solve Data Challenges Related to Scale-Up Model Validation and Missing Data on a Micro-Bioreactor System

Stephen Goldrick, Viktor Sandner, Matthew Cheeks, Richard Turner, Suzanne S. Farid, Graham McCreath, and Jarka Glassey*

Multivariate data analysis (MVDA) is a highly valuable and significantly underutilized resource in biomanufacturing. It offers the opportunity to enhance understanding and leverage useful information from complex high-dimensional data sets, recorded throughout all stages of therapeutic drug manufacture. To help standardize the application and promote this resource within the biopharmaceutical industry, this paper outlines a novel MVDA methodology describing the necessary steps for efficient and effective data analysis. The MVDA methodology is followed to solve two case studies: a "small data" and a "big data" challenge. In the "small data" example, a large-scale data set is compared to data from a scale-down model. This methodology enables a new quantitative metric for equivalence to be established by combining a two one-sided test with principal component analysis. In the "big data" example, this methodology enables accurate predictions of critical missing data essential to a cloning study performed in the ambr15 system. These predictions are generated by exploiting the underlying relationship between the off-line missing values and the on-line measurements through the generation of a partial least squares model. In summary, the proposed MVDA methodology highlights the importance of data pre-processing, restructuring, and visualization during data analytics to solve complex biopharmaceutical challenges.

1. Introduction

"Big data" refers to complex data sets that are too difficult to analyze using traditional data analytic techniques and are classified by the 6 V's. The extreme Volume of data, the wide Variety of structured and unstructured data types, the reliability or Veracity within the data, the Velocity at which the data is produced or analyzed, the Variability of the data over time, and the Value within the data.[1,2] There is no consensus on the size of data classified as "Big data" with tremendous differences in the size of data sets recorded across different industrial sectors. For example, in the retail sector: Walmart collects an estimated 2.5 petabytes of data every hour from its customer transactions[3] whereas the genome human project required analyzing up to 6 trillion base pairs equaling ≈6 terabytes of data.[4] Data recorded within the biomanufacturing environment is on a much smaller scale, typically in the gigabyte range. However, this data are highly complex. Consider the data recorded on a micro-bioreactor system, this includes meta-information containing cell line references, initial conditions related to set-points, and inoculation concentrations in addition to off-line measurements recorded on multiple different analytical devices with varying time delays. Furthermore, the on-line data contains information of up to 50 variables recorded every second for up to 48 different vessels. Therefore, the variety, veracity, velocity, variability, and potential value within these biomanufacturing data sets warrants the

S. Goldrick, S. S. Farid
The Advanced Centre for Biochemical Engineering
Department of Biochemical Engineering
University College London
Gower Street, London WC1E 6BT, UK
E-mail: s.goldrick@ucl.ac.uk

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/biot.201800684

DOI: 10.1002/biot.201800684

S. Goldrick, M. Cheeks, R. Turner
Cell Sciences, Biopharmaceutical Development
MedImmune
Cambridge CB1 6GH, UK

V. Sandner, G. McCreath
FUJIFILM Diosynth Biotechnologies
Process Design and Data Science
Belasis Ave, Stockton-on-Tees, Billingham TS23 1LH, UK

J. Glassey
School of Engineering
Newcastle University
Newcastle upon Tyne NE1 7RU, UK

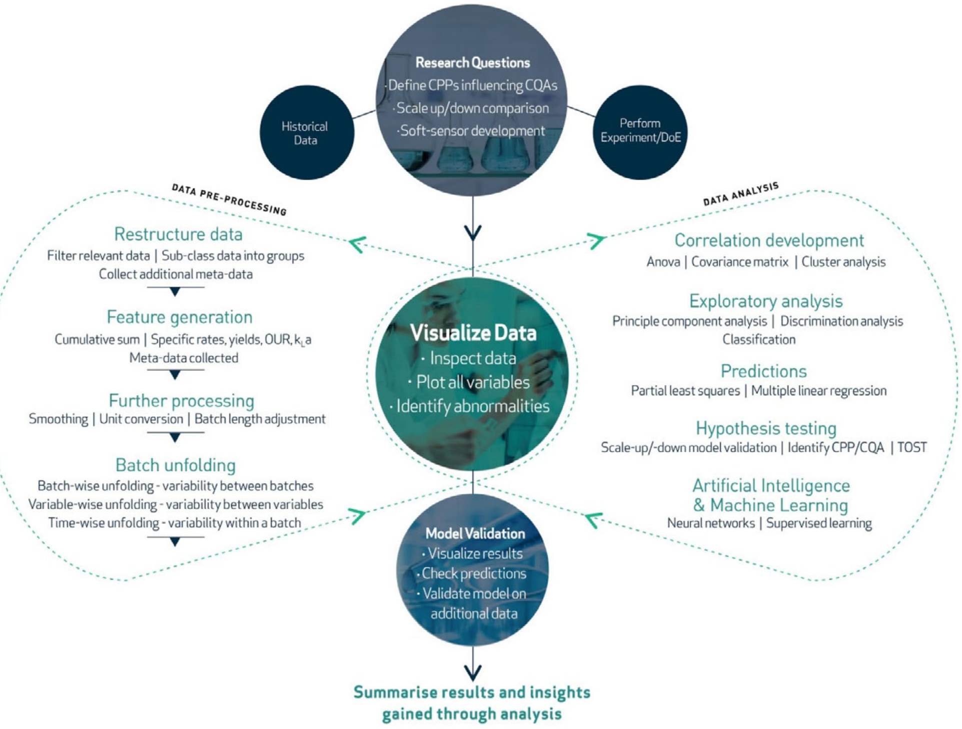

Historical Data
Research Questions
·Define CPPs influencing CQAs
·Scale up/down comparison
·Soft-sensor development
Perform Experiment/DoE
DATA PRE-PROCESSING
Restructure data
Filter relevant data | Sub-class data into groups
Collect additional meta-data
Feature generation
Cumulative sum | Specific rates, yields, OUR, kLa
Meta-data collected
Further processing
Smoothing | Unit conversion | Batch length adjustment
Batch unfolding
Batch-wise unfolding - variability between batches
Variable-wise unfolding - variability between variables
Time-wise unfolding - variability within a batch
Visualize Data
· Inspect data
· Plot all variables
· Identify abnormalities
DATA ANALYSIS
Correlation development
Anova | Covariance matrix | Cluster analysis
Exploratory analysis
Principle component analysis | Discrimination analysis
Classification
Predictions
Partial least squares | Multiple linear regression
Hypothesis testing
Scale-up/-down model validation | Identify CPP/CQA | TOST
Artificial Intelligence & Machine Learning
Neural networks | Supervised learning
Model Validation
· Visualize results
· Check predictions
· Validate model on additional data
Summarise results and insights gained through analysis

Figure 1. MVDA methodology outlining necessary steps for data restructuring, pre-processing, and visualization necessary to implement advanced data analytics on complex biopharmaceutical data sets.

classification of "Big Data." Data analytics and the appropriate management of these small and large complex data sets represent a major challenge for the majority of industries including biopharmaceuticals. The increase in sensor technology,[5] combined with recent advancements in parallelization and development of disposable bioreactors for scale-down purposes[6–8] has resulted in the production of a wide variety of complex data sets to interpret. The sheer volume and variety of the available data remains a significant unmet challenge within biopharmaceutical manufacturing.

With the continued accumulation of these challenging biopharmaceutical data sets, the industry as a whole is leveraging more advanced statistics and multivariate data analysis (MVDA) to understand better the hidden relationships and interactions between their critical process parameters (CPPs) and critical quality attributes (CQAs). The application of MVDA to the biopharmaceutical sector is not a new concept and has been successfully applied for the last 60 years.[9–12] The pioneering work by Wold et al. (1987),[13] Nomikos and MacGregor (1994),[14] and Eriksson (1999)[15] has extended the fundamental principles of principal component analysis (PCA) and partial least squares (PLS) to enable direct comparison of batch-to-batch fermentation systems and the analysis of within batch variability. These techniques provide the necessary mathematical basis for on-line monitoring,[16] root cause analysis,[17] missing data algorithms,[18,19] and real-time control.[20–22] More recent advancements include the development of a MVDA tool kit to simplify the scale-up from ambr15 experiments to pilot-scale (300 L), resulting in shorter process development timelines and reduced costs.[23] Furthermore, MVDA underpins the FDA's quality by design (QbD)[24] and process analytical technology (PAT)[25] initiatives that provide the necessary framework and guidance for a risk-based approach to drug development through better process understanding and control. Other sectors are going one step further and shifting focus toward "Industry 4.0" through the digitization and automation of their processes.[26] These technological innovations promise to revolutionize the biopharmaceutical sector enabling flexible, smart, and better controlled processes. A core component of Industry 4.0 is data analytics and the need to standardize the pre-processing methods of these complex biopharmaceutical data sets is therefore paramount. Previous MVDA methodologies have presented high-level generalizations of data pre-processing and model building,[27,28] with other literature defining various decision trees suitable for specific MVDA techniques such as multilinear regression, factor, discrimination, and cluster analysis.[29] However, neither of these methodologies have emphasized the importance of data visualization which is a core component of our proposed MVDA methodology shown in **Figure 1**. Furthermore, this methodology was specifically developed to focus on the challenges of analyzing biopharmaceutical data and outlines the primary steps necessary for data consolidation and analysis, exemplified in two case studies.

One of those studies includes an example from the "small data" world: Often during scale-up, a single large-scale run ($n = 1$) is compared with multiple small-scale runs ($n = 8$). Plotting daily measured samples between two groups and comparing them visually to ensure they are within certain statistical tolerance ranges are the most widely used methods to gauge comparability across scales. However, the imbalance of runs between small- and large-scale poses a complex challenge. We have addressed this problem, and followed our proposed MVDA methodology to develop a multivariate equivalence test involving time-series data that allows for more objective comparisons.

The second study features a "big data" problem. Here, data at critical time-points are missing due to a hardware failure of the viable cell concentration (VCC) analyzer. These missing VCC values were essential criteria in the ranking of the top clones during a micro-bioreactor clone screening carried out in the ambr15 system ($n = 1 \times 24$). Repeating this whole experiment would result in significant delays to the time-sensitive project milestones. Using our methodology, we were able to effectively recover these missing values through an MVDA approach. Our novel modeling approach took full advantage of on-line measurements and additional off-line measurements recorded during this period, enabling accurate predictions of the missing VCC using a PLS model.

In summary, we created a MVDA methodology, shown in Figure 1, outlining the necessary steps to analyze big and small data problems specific to biomanufacturing data sets. The methodology was demonstrated on two complicated data bioprocessing problems, and can be used by other researchers in solving their own challenges more efficiently.

2. Experimental Section

2.1. MVDA Methodology

This section describes the systematic MVDA workflow that was created through an in-depth investigation into the primary requirements of biopharmaceutical data analysis. The MVDA workflow is outlined in Figure 1 and describes the necessary steps to help improve data analytics and advanced modeling implementation. This methodology was built on previously described "Best-Practices" for MVDA with a particular focus on the importance of visualization and model validation. Prior to any data analysis, the primary research question needs to defined: What process insights or understanding can be gained from this data? Can this data identify any correlations with product quality or productivity? Can the comparability between the scale-up/-down models be concluded? Could any previous or historical experiments help answer this research question? If no suitable data were available, experimental work has to be carried out.

Once the data has been generated, the initial step is to visualize and inspect the data. The importance of this step may not seem critical, but is in fact hugely important, as the researcher can quickly ascertain whether the data contains the correct variables and/or number of batches to effectively investigate the research question. All time-series variables are recommended to be plotted on x-y charts and all end-point or single-point measurements on bar charts. This should be repeated for all groups or categories such as different projects, scales, batches, or cell lines. Restructuring the data frequently occurred during preliminary investigations and this enabled any abnormalities, patterns, gaps, or outliers to be quickly spotted and removed or rectified. Often coined a "sanity-check," this step verified that variables recorded for different groups were recorded across similar timeframes and had the same units. Often, the data set was explored and filtered at the same time during visualization. Many irrelevant or redundant variables that would not be part of the analysis were removed during this iterative process. In a regulated environment all data modifications should be governed by the ALCOA principles ensuring the data are Attributable, Legible, Contemporaneously recorded, Original or a true copy, and Accurate.[30] This ensures data integrity and regulatory standards are maintained during analysis.

Typically, data sets are enriched using feature generation leveraging additional information through the generation of meaningful feature vectors. These can include the cumulative sum, specific productivity, or calculated variables such as oxygen transfer rate (OTR) or oxygen mass transfer rate ($k_L a$).[31] Off-line data recorded at slightly different times each day could be categorized into daily off-line measurements to simplify subsequent analysis. The final pre-processing steps require smoothing, unit conversions, and interpolation. To handle missing data during the analysis of two groups, there are two options. The first is to estimate the data through interpolation or more advanced missing data algorithms and the second option is to remove the data in each group to enable comparability.

After boiling down the data set to its essence, it was again visualized to ensure all the previous pre-processing, feature generation, and restructuring had been correctly carried out and no obvious errors could be observed. Restructuring the data in this fashion significantly simplified the implementation and evaluation of any advanced modeling techniques. When the data set includes batches, as is often the case in bioprocesses, there are three algorithms suitable for unfolding the data: i) time-wise unfolding, suitable for analyzing variability throughout the batch time, ii) batch-wise unfolding to investigate variability between batches, and iii) variable-wise unfolding, allowing differences among variables to be identified.[16,32] The previously described visualization, pre-processing, and data restructuring steps were sequential and iterative operations which should be automated through the development of algorithms in preferred software, for example, Matlab, R, or Python.

Once the data have been correctly restructured and preprocessed, there are multiple MVDA techniques that can help visualize the research solution. These include, but are not limited to: correlation development, exploratory analysis, predictive modeling, hypothesis testing, and artificial intelligence/machine learning solutions. If none of these techniques can answer the research question, the data itself may not have the required depth and quality. This can be a result of key process variables not recorded, range of variables not varied enough to have a measurable impact on performance, or too much noise present to statistically validate a research question. In this case, new laboratory experiments have to be designed, which could be further optimized through a Design of Experiment (DoE).

2.2. Cell Line and Culture Propagation

The first case study involved cell culture data at Fujifilm Diosynth Biotechnologies, recorded at the facility in RTP, NC, USA and analyzed in Billingham, UK. The second case study involved cell culture data recorded by AstraZeneca at their facility in Cambridge, UK. The cell lines utilized in these experiments were recombinant Chinese hamster ovary (CHO) expressing high levels of a protein. The cells were cultured in defined CHO media and maintained at 37 °C under 5% carbon dioxide, shaken at a constant RPM, and passaged 2–3 times per week for propagation and scale-up.

2.3. Bioreactor Systems

Scale-up experiments were performed at the 2 and 200 L scale in standard glass and stainless steel bioreactors, respectively. Two clone screening experiments were conducted on the ambr15 micro-bioreactors (TAP Biosystems, Greenville, DE) with 24 single vessels and operated between 11 and 15 mL working volume.

2.4. Cell Culture Process

Initial seeding density for all culture stations was between $1–10 \times 10^5$ cells mL^{-1}. Nutrient feeding followed a standard fed-batch protocol. The dissolved oxygen set point was set between 30% and 60% and maintained by gassing with air and oxygen. The agitator speed was systematically ramped up for all experiments to maintain the dissolved oxygen set point. Temperature and pH set points of all experiments were maintained between 35–37 °C and 6.8–7.2 °C, respectively. For each bioreactor, antifoam was added as required. Daily at-line samples were analyzed for metabolic profiles and VCC for all experiments.

2.5. Software

Missing data and predictive regression algorithms were performed using Matlab 2018b (The MathWorks,Inc., Natick, MA). R, an open source software, was used for TOST and PCA calculations (R Foundation for Statistical Computing, Version 3.4.4, Vienna, Austria).

2.6. Case Studies—Application of MVDA Methodology

2.6.1. Case Study 1—Small Data: Quantify Equivalence with a Limited Number of Batches

The first case study features a mammalian cell culture process that was scaled up from 2 to 200 L and operated at similar process conditions. Historically, comparability is determined using a variety of statistical techniques. The most popular ones are the two-sample *t*-test, statistical tolerance intervals, and the two one-sided test (TOST). The *t*-test fails to provide a measure of equivalence, since the absence of an evidence to declare inequivalence does not equate equivalence. Additionally, tolerance intervals provide a statistical measure of the upper and lower bounds where a certain portion of the data is expected to lie but does not quantify the comparability of new data. The TOST is specially designed to assess comparability of data and is the preferred statistical equivalence testing method within the pharmaceutical sector. For example, the TOST correctly identified a statistical difference between the performance of a tablet dissolution test carried out in a development laboratory compared to a contract manufacture organization whereas the two-sample *t*-test incorrectly declared these two methods equal.[33]

The MVDA methodology (Figure 1) was implemented to add more depth to the current practice of equivalence testing by turning an otherwise univariate TOST into a multivariate one by utilizing the full wealth of the time-series data and increasing the *N* numbers and confidence of the equivalence testing. The TOST is therefore well suited to quantify equivalence between the two scales. The test calculates the confidence interval between their means and standard deviations while taking into consideration the number of batches at each scale.[33] The methodology to employ the TOST on time-profiles is described in great detail by diCesare et al.[34] All variables used in the analysis can be seen in Figure 3.

Visualization and Filtering: First, the data set was inspected visually to generate leads on how to solve the challenge at hand. Typically, the analysis starts by drilling down from the large data set into a more compact version that contains all necessary observations for the required analysis, without redundant or irrelevant information. Variables with large proportions of missing values were removed and variables not available in both scales were removed to ensure consistent data sets. Finally, scale-dependent process variables such as RPM were by default not equivalent during scale-up and were excluded from this analysis. More information on selection of variables for scale-up can be found in refs. [8] and [35].

Feature Generation and Data Processing: Features such as specific productivity, integrated VCC, and lactic acid production were calculated and appended to the data set. A Savitzky-Golay/pchip interpolation or smoothing step was implemented on noisy variables. All units were changed to a consistent format (i.e., g L^{-1} instead of mM, and L instead of mL), to ensure comparability across runs. Finally, the consolidated data set was repeatedly visualized with scatter plots to ensure all filtering and data pre-processing operations were carried out as expected. The data set contained data from day 2–12, had 17 variables, of which 3 were derived variables (IVCC, qLac, qP, see ref. [36] for their calculation) and were labeled with identifiers with one group containing the eight batches at 2 L scale and the second group containing one batch at the 200 L scale.

Batch-Wise Unfolding Per Day: The consolidated data set was subdivided for every measurement day, resulting in a batch-wise unfolded matrix with rows representing different batches and columns representing variables, this was calculated for each day. The unfolded subset had dimensions equal to a 9×17 matrix and was utilized for the next operation.

Principal Component Analysis: Iteratively for every measurement day, every variable's data was mean-centered and scaled by dividing by the standard deviation before calculating loadings and scores with PCA using R's FactoMineR package version 1.34. PCA calculates the weight of particular variables in each principal component (PC) and both their cumulative and individual variance explained in percent were used to calculate the new metric of equivalence. Although more PCs could be calculated, the default threshold was set to 5 PCs in this study. Using more than 5 PCs did not improve the analysis because only a small amount of variance was attributed to additional PCs. Meta-information on the quality of the PCA (how much variance was explained by the PCs and which variables contributed to individual PCs) were stored for later re-combination with the outcomes of the TOST. This procedure was repeated for all measurement days and the scores with the matrix dimensions 9 × 5 were progressed to the next step of the analysis.

Multivariate Time-Series Two One-Sided Test: Each of the 5 PCs score matrices was used iteratively as input in R's TOSTER package version 0.3 (https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf). TOST is used when the equivalence of variables (or lack thereof), belonging to two different groups (here 2 and 200 L), is being tested. Typical inputs into the *TOSTtwo* function are: a variable's mean and its standard deviation in both groups, the sample size per variable per group (the variables in this case are the PC's scores within one of the two groups), Cohen's d, and an alpha level for the statistical quantification of the test. For more details, please refer to the github repository: (https://github.com/Lakens/TOSTER/blob/master/R/TOSTtwo.R). Some inputs into the TOST require special consideration, such as the 200 L group with only one value per variable per day available for analysis. The standard deviation was calculated from the PCA scores from the 2 L group and also used for the 200 L group. The mean of the PCA scores was taken as the mean for the 2 L group, while the single PCA score value from the 200 L group was taken as the mean for the 200 L group. Furthermore, in order to represent a point value without uncertainty spread, an artificially high value of $n = 10^6$ was well suited without compromising the integrity of the TOST results.

Finally, TOST requires an input of statistical confidence metrics: alpha represents the level of significance of the test, taken here as 0.025, and the standard deviation multiplier (SD×) relates to the desired level of comparability between the groups, taken here as 2σ. The standard deviations are utilized to calculate the Cohen's d margin, one of the most important metrics that will support or refute equivalence in the TOST.[37] Cohen's d is a measure of the distance between two means, measured here in standard deviations[38] (see Supporting Information Section).

Quantification of Individual and Global Equivalence: Testing equivalence of the variance between two scales with PCA scores (instead of measurement variables as input) allows calculation of variable contribution (e.g., pH, LAC...) toward equivalence between two scales. For any given variable, only the variable's percentile weights in an equivalent PC (the scores that passed the TOST) were summed up, while the scores from PCs, which did not pass the TOST, were not. Therefore, only variables that contain the most variance in the process were considered, and their individual contributions were reflected as a percentage. In other words, the comparability between variables across the different scales can now be quantified in percent, which is defined here as the variable's individual equivalence (IEQ) score. More details on the calculation of the IEQ is provided in Supporting Information Section. Furthermore, these individual time-series equivalence scores were averaged to generate a single global equivalence value, defining the overall level of equivalence between the two groups throughout the duration of the cell culture run. A result matrix was created which holds all permutations enabling the M-TOST results to be plotted. Combinations of 25 different alpha and SD × conditions, 17 variables, and 10 days of process information, sum up to a total of 4250 results, which provide a new level of detail to the previously rather small data set. The sensitivity of the method can be gauged when different alpha and SD × values are used as input for the TOST, which can be reviewed as CSV file in the Supporting Information Section.

2.6.2. Case Study 2—Big Data: Predictive Modeling of Missing Data in High-Throughput Ambr15

The first ambr15 clone screening experiment contained missing data in VCC from day 6–10. The second screening experiment had no missing data and served to validate this missing data algorithm. The MVDA methodology outlined in Figure 1 is followed to pre-process the data. All the relevant on-line and off-line data were visualized, collated, and restructured similarly to the previous case study. The noise associated with the time-series measurements was minimized through a Savitzky–Golay smoothing algorithm. To handle the sparse matrices that are generated by feed additions and gas flow rate variables, their cumulative sum was calculated. These newly created feature vectors are easier to interpret and analyze. Furthermore, the OTR was calculated here as a feature vector. To predict the missing VCC values a time-wise unfolded algorithm was applied to the data set.

Characterization of Missing Data: There are two distinct features which characterize missing data: the missing data pattern referring to the configuration of missing data within the data set and the missing data mechanism referring to relationships between the measured variables and the probability of missing data.[39,40] The five primary missing data patterns are: 1) univariate pattern, data is missing from single variable; 2) multivariate pattern, data is missing from multiple variables; 3) monotone pattern, all data from a specific point till the end of an experiment is missing, that is, as a result of probe failure during an experiment; 4) general pattern, data is missing at random with no particular pattern; 5) planned missing pattern, not all data is collected due to experiment design, that is, to reduce burden of analytics. The three mechanisms describing the cause of the missing data are: 1) missing at random (MAR) where there is a systemic relationship between one or more of the measured variables and the probability of missing data; 2) missing completely at random (MCAR) indicates no relationship between any of the measured variables and the probability of missing data; 3) missing not at random (MNAR) defines where the probability of missing data depends on the missing data itself. Depending on the missing data pattern and mechanism, there are multiple missing data algorithms to apply which are outlined in detail in refs. [39,40].

Missing Data Algorithms within Bioprocessing: There are two primary strategies to deal with missing data, these include ignorance (discarding incomplete data sets) and imputation.[40] When the missing data contains important information essential for subsequent analysis, predicting these missing values through imputation is required.

The available imputation techniques are based on univariate, multivariate, or Bayesian statistics.

Simple univariate techniques are advantageous due to their simplicity and infer missing data through linear interpolation methods or replacing with simple approximations such as mean or nearest neighbor. These predictions are typically only valid for linear variables or to predict small sections of missing data.

More advanced techniques to predict missing data include interpolation using an inverse distance weight (IDW) method. Nancy et al.[41] extended this technique to predict some missing time-series clinical trail observations through a concept of tolerance rough set (TR) and particle size optimization (PSO). Bayesian methods can predict missing data by estimating a probability distribution of the missing data and benefit by enabling uncertainty and confidence levels to be selected for the predicted data measurements. These methods were successfully applied to predict missing data within gene expression profile data.[42] The most routinely applied multivariate techniques rely on PCA or PLS and exploit the correlations between the missing data and measured variables. These techniques are well-established methods within the biopharmaceutical industry[43–45] and perform well with moderate amounts of missing data (up to 20% of the measurements[46]). This paper focuses on the generation of a PLS model based on its proven ability within industry to enable on-line predictions of off-line variables[47,48] and missing data.[18,19] Additionally two linear interpolation methods (linear and cubic splines) were applied to compare against the PLS model predictions.

The PLS model was developed as described by Geladi and Kowalski 1986.[49] An individual PLS model was developed for each cell culture run. Initially, all the J on-line and off-line variables were interpolated to a fixed length equal to K. This process simplifies the generation of the model and subsequent analysis, however this procedure does not account for any potential non-linearities between each off-line measurement. Each interpolated variable ($1 \times K$) was concatenated to form a 2D data ($J \times K$) matrix. The PLS model was built using all available on-line and off-line process variables (summarized in Figure 4B) and selected four latent variables based on minimizing the root mean squared error (RMSE) of the calibration data set. Additionally a cross validation procedure was implemented to ensure the model was not overfitted to the data. The four latent variables captured 81% of the total variance in the X-block and 99% in the Y-block. Taking additional latent variables unnecessarily increased the complexity of the PLS model with only a marginal decrease in the RMSE. A similar procedure was carried for the second ambr15 except the calibration data set used the time series data from days 0 to 6 and 10 to 14 with the validation data set using the known VCC values recorded on days 6–10. The linear regression and cubic splines were generated as a function of time and were compared against the PLS model predictions.

Cell Line Ranking Algorithm: To quantify the accuracy of the different missing data algorithms the newly generated VCC predictions of the second ambr15 were used to rank the cell lines in ascending order based on their maximum VCD achieved. This was then compared to the actual ranking of the cell lines using the known VCC values. The similarities between the different rankings were assessed by the Kendal Ranking coefficient (τ).

$$\tau = \frac{(\text{Number of agreements in order}) - \text{ number of disagreements in order}}{(n)\,(n-1)\,/2} \tag{1}$$

where n is the number of cell lines to be ranked. The Kendall rank correlation coefficient evaluates the similarity between different ranking algorithms applied to the same set of objects. A value equal to 1 represents a perfect positive relationship, 0 represents no relationship, and –1 represents a perfect inverse relationship between the ranking lists. Typically, additional metrics are also used for lead clone selection which includes but are not limited to productivity (titre and specific growth rates), product inhibition (lactate and ammonia), and product quality (aggregation and fragment concentrations) considerations.[50]

3. Results and Discussion

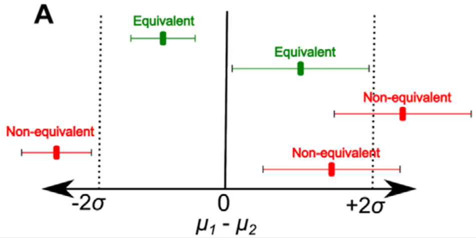
3.1. Case Study 1—Small Data

Typically, scale-up studies focus on comparable oxygen transfer rates, mixing times, and geometry parameters.[35,51–53] However, biopharmaceutical facilities are not normally designed according to a fixed scale-up criteria but are dependent on separate process development and optimization strategies at each scale, which can differ for products, processes, or facilities.[54] Therefore, a metric to quantify equivalence of process performance between the two scales is required. Routinely equivalence can be judged by either overlaying time-series plots to visually judge the similarity, or with a series of univariate TOST conducted on end-point key performance indicators (KPIs). The first method lacks objectivity, while the second method reduces the whole data set to a comparison of just a few endpoints. We applied the MVDA methodology (Figure 1) to add more depth to the current practice of equivalence testing by turning an otherwise univariate TOST into a multivariate one that utilizes the entire time-series data set. The TOST is an ideal equivalence test as the aim of this study is not to show differences, but to conclude similarity.[55]

Here, TOST was combined with the previously calculated PCA scores of either the 2 or 200 L scale group. Our extension was its combination with another well-established algorithm, PCA. PCA was selected as it is one of the most commonly used MVDA techniques within the biopharmaceutical sector and can help visualize, interpret, and quantify the variance of complex data sets using a reduced dimensional space. PCA was implemented here to calculate the global equivalence between the two groups. In brief, the application of a multivariate time-series TOST enables a quick and clear picture of a comparison of the variance in these variables that are most and least equivalent, defined by a single percentage metric.

Figure 2A gives an overview of the TOST equivalence determination implemented for case study 1. Figure 2B outlines a summary of the TOST equivalence statistical test results from day 2–12 for each principal component. The global equivalence, averaging all TPVs, was 81%. This simple metric summarizes the overall equivalence between the 200 and 2 L systems taking into account all the time-series data recorded by the 17 variables used in the test. The variables of both systems are found to be equivalent throughout the duration of the runs as shown in Figure 2B.

According to the multivariate TOST, the five variables with the largest average level of individual equivalence (IEQ) were VCC, BGApH, BGApCO2, Titre, and pO2, while the five least equivalent variables were GLU, GLN, LAC, pH, and qp (**Figure 3**). The



B

PC reference	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 12
$PC-1$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$PC-2$	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓
$PC-3$	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
$PC-4$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$PC-5$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TPV	82%	86%	87%	77%	76%	74%	72%	78%	75%	98%

Figure 2. A) Outline of TOST determination with statistical equivalence defined by the difference in means and confidence limits of the two groups lying within predefined ranges. Non-equivalence is determined by the difference in means and confidence limits of the two groups lying outside the predefined ranges. In this case study, the predefined ranges were taken as two standard deviations (±2σ). B) A summary of the TOST equivalence statistical result based on comparing the scores of the 200 and 2 L systems recorded from day 2–12 for each principal component (PC). Total passed variance (TPV) sums up all variance that was explained by the PCs that passed the TOST.

biggest differences were related to metabolic markers, especially amino acids, such as glutamate, glutamine, lactic acid, and ammonia, but also qLac and qP.

Two versions of pH measurements were included in the data set one from the off-line Blood Gas Analyzer (BGApH), and the other from the on-line pH probe (pH). As can be seen from Figure 3, pH measurements between the two methods are significant different. The pH measured by the (on-line) probe had some of the lowest equivalence scores, while the bench-top BGApH scored on the opposite spectrum. The on-line pH probe may be more inaccurate as these probes require frequent recalibration during the process because of drift,[56] which might be related to harsher conditions during equipment sterilization. The on-line pH probe is typically recalibrated taking the off-line pH probe measurement as the reference and therefore the off-line method is more accurate. However, both of these measurements should be included in the equivalence test as the poor performance of the on-line pH ensures this deviation is further investigated. Possible differences could be the result of different pH probes used at the different scales or due to the position of the probe in the 200 L system. Additional evaluation of these pH deviations should be carried out to ensure process behavior between the scales is not affected.

In the case of the variables VCC and Titre, both are highly equivalent in both scales, which indicate an overall good control of KPIs. The variable BGApCO2 was not specifically controlled at either scale but monitored nevertheless. A large variance of pCO2 in all processes was considered normal and enabled the pCO2 to have a high IEQ value.

Comparing our analytical findings with the consolidated data, the same conclusions can be drawn, but in a more objective way. Our new metric allowed us to quantify equivalence at particular statistical confidence levels. The scaled-up 200 L process reached a total equivalence score of 81% at a statistical confidence level of alpha = 0.025 and SD × of 2σ. Testing other statistical confidence levels in an in-silico DoE, where alpha and SD × were varied, we found that varying SD × had a stronger impact on overall equivalence than alpha.

Varying alpha between 0.005 and 0.1 while keeping the SD × level constant, the following average TPV were obtained: 38–63% at SD × 1σ, 63–78% at SD × 1.5σ, 77–84% at SD × 2σ, 81–85% at SD × 2.5σ, and 85–88% at SD × 3σ. The variance in TPV decreased as SD × was increased, while in general higher SD × values result in easier passing of the equivalence test, leading to a higher score of equivalence. Industrial use of SD × 2σ, SD × 2.5σ, and SD × 3σ is common, although more precise values for multiplying SD could be derived.[33] In contrast to other approaches, where a minimum effect size has to be calculated[57] or an equivalence limit defined,[58] this work utilizes scores and loadings of reduced dimensionality to determine equivalence by TOST. Judging by the results presented here, we found that using SD × was an appropriate approach. Only little guidance can be given regarding the choice of alpha and SD ×,[59] and thus the recommendations of other researchers[60] to report the justification of level chosen should be followed. The choice should be based on the level of statistical confidence required, which will be guided by the difficulty of establishing equivalence at a given level. SD × 2σ and alpha 0.025 were used as the statistical scale-down model equivalence level in this research.

An important test of equivalence is comparability between product quality at two scales—for this case, a simple TOST may be sufficient. This contribution's multivariate TOST focus is on operating conditions and did not consider end point measurements. However, our approach is to quantify the level of equivalence and variable contribution in percent, resulting in an overall equivalence score of multiple variables between two groups. In addition to providing a metric, it helps to quickly identify variables which are most and least equivalent. This approach could be applied also in different settings, for example, for other upstream processes, such as microbial, attachment cell cultures, tissues cultures, or downstream operations such as chromatography.

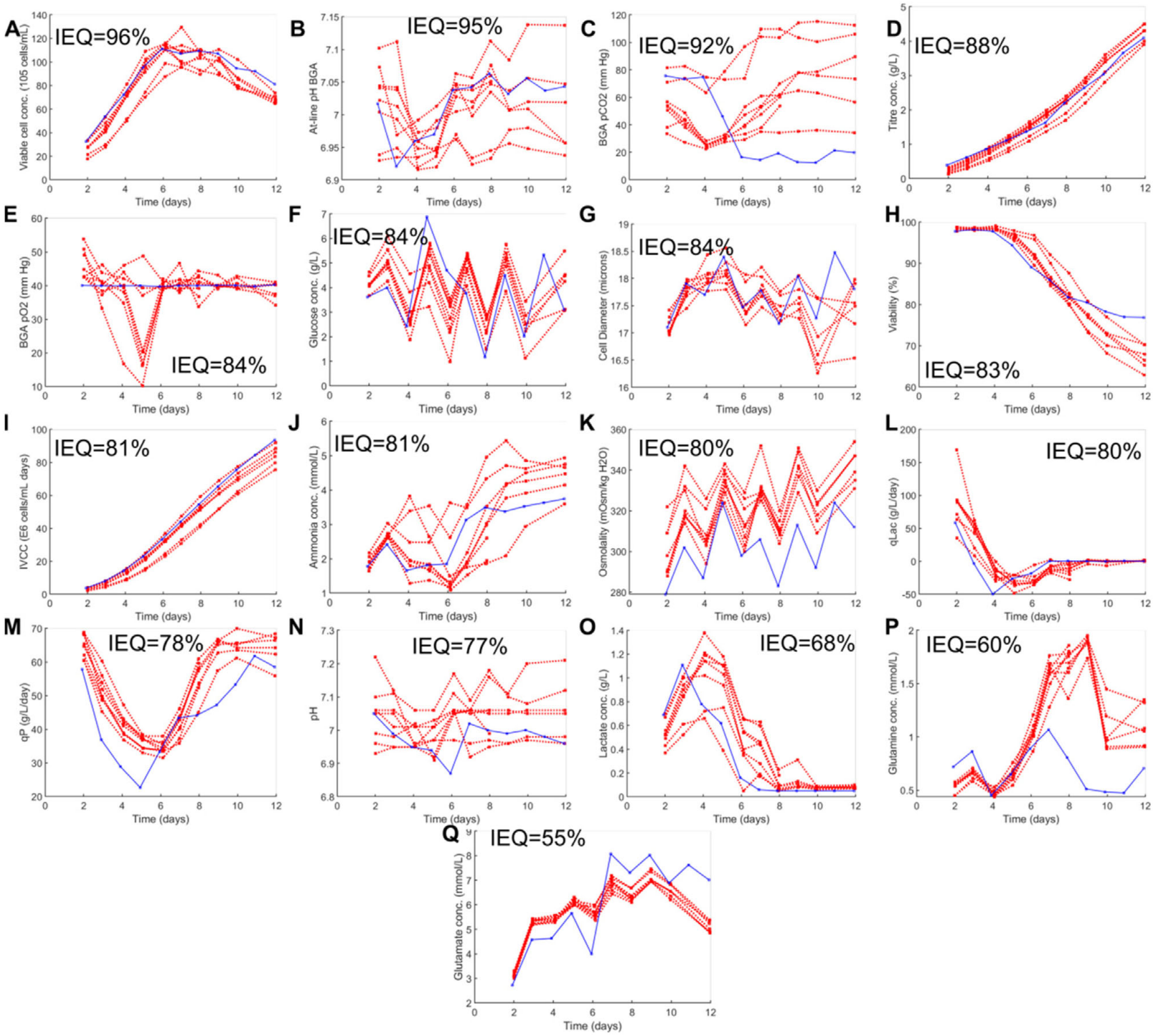
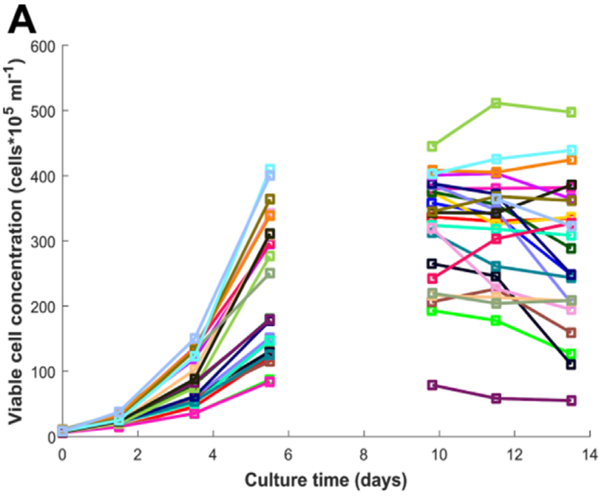


Figure 3. Consolidated time-series data set. Dashed red lines represent the 2 L runs and the solid blue line represents the 200 L run. Percentages describe the average individual equivalence (IEQ) of each variable over time in descending order.

3.2. Case Study 2—Big Data

High-throughput screening experiments enable rapid and efficient selection of highly productive lead clones.[61] This cell line screening strategy ensures a highly robust lead clone is selected that demonstrates high protein expression levels with the necessary phenotypic and product quality attributes while maintaining consistent growth performance metrics across predefined bioprocess conditions.[62] However, the caveat of operating high-throughput experiments relates to any technical problems resulting in rapid and simultaneous issues to all cell culture runs.[63] **Figure** 4A demonstrates the significant challenge of handling missing data during a cell line selection protocol carried out in the high-throughput ambr15 system. A technical fault with the Vi-Cell automated cell viability analyzer resulted in no VCC measurement recorded on day 8. The VCC measurement is defined as a CPP and is a vital selection criteria for optimum cell line selection. Therefore, these missing VCC time-points compromise the entire run. Previous missing data algorithms specific to bioprocessing utilize only the available off-line data and single point measurements of on-line data to infer these missing values,[64] however the novel approach taken here utilizes both high frequency on-line and low frequency off-line data to predict these missing values. This approach is advantageous as it exploits all available data recorded during this time period of missing data. Figure 4B highlights the available off-line and on-line variables that were utilized to infer the missing VCC values in addition to the necessary pre-processing steps required for each variable.

To investigate whether any correlations existed between the available on-line and off-line variables and the missing VCC



B

Variable	Smooth	Cumulative	Feature Generation
Off-line measurements			
Lact	✗	✗	✗
Gluc	✗	✗	✗
Osmo	✗	✗	✗
Titre	✗	✗	✗
On-line measurements			
Temp	✓	✗	✗
pH	✓	✗	✗
NFD	✗	✗	✗
F_{Gluc}	✗	✓	✗
$F_{a/b}$	✗	✓	✗
V	✓	✗	✓(OTR)
DO_2	✓	✗	✓(OTR)
CO_2	✓	✗	✓(OTR)
$F_{air}, F_{CO_2}, F_{O_2}$	✓	✓	✓(OTR)
RPM	✓	✗	✓(OTR)

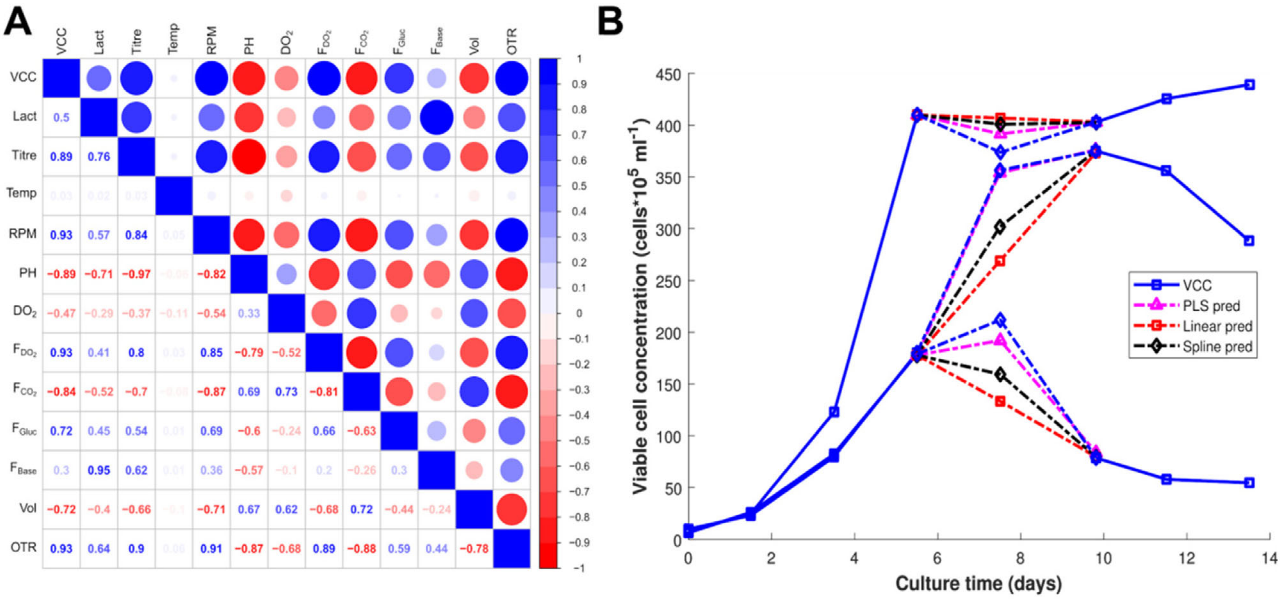
Figure 4. A) Viable cell concentration (VCC) measurements highlighting missing values on day 8 due to technical fault. B) Summary of pre-processing requirements for all available on-line and off-line measurements that were utilized to generate the PLS model.

variables, a correlation matrix was generated (**Figure 5**A). This matrix highlights the interdependence between some of the on-line and off-line variables and the VCC values in particular the gas flow rates (F_{DO2}), RPM, OTR, and titre values which have a correlation coefficient (R^2) close to 0.9. The strengths of these correlations demonstrate the significant potential of these variables to enable accurate predictions of the missing VCC values. The missing data is characterized by a monotone pattern as only data from a single variable is missing and is classified by MCAR mechanism. This pattern and mechanism of missing data is highly suited for established MVDA techniques to exploit the correlations shown in Figure 5A between available measurements and the missing data. PLS was selected to predict the missing data based on its versatility and proven ability to handle data containing both high frequency data (recorded every second) and off-line data (recorded every 24 h). The PLS model utilized the pre-processed data outlined in Figure 4B. It was trained using data from days 0 to 6 and days 10 to 14 and was used to predict the missing VCC values during days 6–10. The true values of the missing VCC are irretrievable, and therefore the accuracy of this PLS model to predict these measurements cannot be assessed. In order to validate this approach, a second data set without missing values was modified. On day 6–10, VCC values were removed and a PLS model was created as previously described. This enabled the accuracy of the predictions to be evaluated. Other simpler methods, including linear regression and cubic splines were also implemented to compare against the PLS model predictions.

A subset of these predictions for the second ambr15 are highlighted in Figure 5B. The linear interpolation method poorly predicts the missing VCCs and fails to capture the nonlinearities inherent to the cell growth profile. The cubic spline gives a better fit enabling a natural continuation between the missing VCC measurements, however any significant deviations of the VCC on day 8 are not captured by these simple interpolation methods. The more advanced PLS model takes advantage of the strong correlations between the other available process measurements resulting in accurate predictions of the VCC measurements. The PLS model captures the significant nonlinearities associated with this variable at this critical time-point. The performance of each missing data algorithm is quantified through the calculation of the RMSE between the predicted and experimental VCC values in each of the 24 cell culture runs. The PLS model has the lowest RMSE measurement, thus demonstrating the superiority of this technique to accurately predict the missing VCC values.

Previous work has demonstrated the ability of four different imputation methods (deletion, mean, nearest neighbor, and maximum likelihood) to predict missing data for time-series data,[65] they concluded mean imputation yielded the most inaccurate predictions and the maximum likelihood method was the best performer. A problem with these imputation models is their inability to predict previous non-observed behavior. Therefore, they may not predict the observed nonlinear growth patterns observed in Figure 5B as the cells shift from exponential stage to stationary stage.[41] Mante et al.[64] demonstrated the ability of simple polynomial, logarithmic regression, and mean imputation techniques to successfully predict missing time-series titre data suitable for secondary analyses. Therefore, these methods can be useful. However, the more advanced MVDA methodologies including PLS enable better predictions by leveraging the strong correlations of the other variables recorded during the period of missing data. The advantage of the PLS algorithm is that it focuses on maximizing the relationship between the input data and the response, observed in Figure 5A, which improves the overall predictions of the missing VCC values.

Figure 5A highlights a number of strong positive correlations, for example, an R^2 value of 0.93 is shown between the viable cell concentration (VCC) and both the dissolved oxygen gas flowrates (F_{DO2}) and calculated OTR. Similarly positive correlations were observed by Casablancas et al.[66] and Fleischer.[67] These linear relationships demonstrate the importance of the oxygen for cellular growth and maintenance. The strength of these correlations enable exploitation for soft-sensor development or advanced control applications. This was demonstrated by Goldrick et al.[31] to



	VCC	Lact	Titre	Temp	RPM	PH	DO_2	F_{DO_2}	F_{CO_2}	F_{Gluc}	F_{Base}	Vol	OTR
VCC													
Lact	0.5												
Titre	0.89	0.76											
Temp	0.03	0.02	0.03										
RPM	0.93	0.57	0.84	0.05									
PH	−0.89	−0.71	−0.97	−0.06	−0.82								
DO_2	−0.47	−0.29	−0.37	−0.11	−0.54	0.33							
F_{DO_2}	0.93	0.41	0.8	0.03	0.85	−0.79	−0.52						
F_{CO_2}	−0.84	−0.52	−0.7	−0.06	−0.87	0.69	0.73	−0.81					
F_{Gluc}	0.72	0.45	0.54	0.01	0.69	−0.6	−0.24	0.66	−0.63				
F_{Base}	0.3	0.95	0.62	0.01	0.36	−0.57	−0.1	0.2	−0.26	0.3			
Vol	−0.72	−0.4	−0.66	−0.1	−0.71	0.67	0.62	−0.68	0.72	−0.44	−0.24		
OTR	0.93	0.64	0.9	0.06	0.91	−0.87	−0.68	0.89	−0.88	0.59	0.44	−0.78	

C

Prediction Method	R^2 (-)	Cross validation-RMSE (cells*10^5 mL^{-1})	Kendall rank coefficient (-)
Linear regression	0.62	23±16	0.88
Spline (Cubic)	0.67	20± 15	0.88
PLS	0.90	7±5	0.93

Figure 5. A) Correlation between the offline variables (VCC, Lac, and Titre) and the available on-line measurements (Temp, RPM, PH, DO_2, F_{DO2}, F_{Gluc}, F_{Base},Vol, and OTR). Circle size indicates strength of correlation and numbers indicate R^2 value, color-coded for proportional (blue), and inverse (red) correlation. B) Performance of predicted VCC values versus measured VCC values (blue squares), the legend shows different predictions: PLS predictions (triangles), linear predictions (squares), and spline fit (diamonds). C) Table highlighting the accuracy of the three missing data algorithms and their ability to rank each of the 24 cell culture runs in the correct order.

predict the glucose concentration on-line on a mammalian cell culture based off a strong correlation between the cumulative oxygen transfer rate and the cumulative glucose consumed. This soft-sensor was incorporated into a control algorithm enabling glucose concentration to be controlled to a fixed set-point. The correlations observed in this work could be further exploited to predict titre concertation, in particular the OTR, which had a R^2 value of 0.93. Previous linear correlations were also observed between titre and OUR during the cultivation of mammalian cell cultures.[68]

Furthermore, the predicted missing VCC values enabled the maximum VCC values to be utilized as a key performance indicator in this cell line ranking protocol. The accuracy of the missing data algorithms was compared to the true values recorded on the second ambr15 data set as summarized in Figure 5C. The ranking order of each missing data algorithm is shown in Figure S1, Supporting Information, and quantified using the nonparametric Kendall rank correlation coefficient (Equation (1)). The VCC predictions generated by the PLS algorithm resulted in the highest Kendall rank coefficient (0.93) and ranked the top nine cell lines correctly in comparison to the true rankings utilizing the complete VCC data set. Both the linear and spline regression methods had a similar Kendall rank coefficient equal to ≈0.88 and were able to effectively rank the top five cell lines except not in the correct order. Thus both the spline and linear missing data algorithms can give an approximate estimate of missing data measurements, however for more confident predictions the PLS model significantly outperforms both methods.

The proposed methodology is highly transferable across scales and processes. The PLS model generated in this work only requires a representative data set where both the on-line and corresponding off-line data are available to generate the necessary correlation enabling predictions of additional off-line variables. A limitation of the PLS model is the challenge of quantifying the contribution of the measured variables toward the prediction of the missing data. This involves the generation of the variable of importance (VIP) graph as described in ref. [17].

4. Conclusions

This contribution outlines a methodology for effective evaluation of complex multivariate biopharmaceutical data. This MVDA methodology outlines important data pre-processing, restructuring, and visualization steps. Visualizing data is a recurrent activity and took a central role in our MVDA methodology (Figure 1). To demonstrate the benefits of following this MVDA methodology, two challenging case studies relevant to real-world biopharmaceutical problems were presented.

The first case study highlights the application of the MVDA methodology to develop a new quantification of equivalence, when only a very small data set was available. The solution was to process the data into a more suitable structure, where time profiles were tested, instead of comparing only endpoints. In order to work with meaningful data, a consolidated data set was obtained after adding features, smoothing, filtering, and normalization steps. Data unfolding techniques were combined with robust statistical tools such as PCA and TOST, and finally a new equivalence metric was developed. This new metric could quantify the similarity between a number of scale-down runs and a single large-scale run at different levels of statistical significance.

In the second case study, the MVDA methodology was implemented to infer missing VCC values that were essential to a cloning study carried out in the ambr15 system. To quantify the accuracy of three different missing data algorithms, a second ambr15 system was used with the predicted VCC values compared against the experimentally recorded VCC values. The PLS model outperformed both the spline and linear interpolation methods with superior accuracy. The PLS model was able to exploit the existing correlations between the VCC values and other recorded on-line and off-line measurements. Furthermore, these predictions enabled the top nine cell lines to be ranked correctly based on their maximum VCC values. This work demonstrates the benefits of correctly implemented MVDA to help recover previously thought failed experiments resulting in significant labor, resource, and time savings.

In summary, two highly diverse and representative biopharmaceutical problems could be solved through the implementation of MVDA solutions. Our proposed methodology takes the reader through our own examples of data processing, data visualization, and data analysis. The presented case studies demonstrate the power of following a structured and methodical MVDA approach. We believe it will help researchers to find insights hidden within complex data sets faster and more efficiently.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

S.G. and V.S. contributed equally to this work. V.S. and S.G. conceived the study and wrote the manuscript in equal parts. V.S. developed the small data study and designed the M-TOST to compute equivalence between small and large-scale data sets. S.G. developed the big data study and designed the PLS algorithm to compute missing data with alternative variables. R.T. and M.C. contributed the data sets for the big data study. S.F., G.M., and J.G. helped in drafting the manuscript and supported the study. S.F. and J.G. supervised research activities. This research is associated with the joint UCL-Astrazeneca Centre of Excellence for predictive multivariate decision support tools in the bioprocessing sector and financial support for S.G. is gratefully acknowledged. Furthermore, support from EPSRC for S.G. is also greatly appreciated (EP/I033270/1). V.S. gratefully acknowledges the financial support from the EU-Horizon 2020 Marie Skłodowska-Curie Actions (MSCA) ITN project BIORAPID (no. 643056). The authors would like to thank Dr. Don Ornek, at that time Senior Principal Scientist at FUJIFILM Diosynth Biotechnologies at RTP, North Carolina, US, who contributed the data set for the small data study.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

cell culture, missing data, multivariate data analysis, scale-up/down, two one-sided test

Received: November 13, 2018
Revised: September 26, 2019
Published online: November 22, 2019

[1] A. Gandomi, M. Haider, *J. Inf. Manage.* **2015**, *35*, 137.
[2] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, G. Z. Yang, *IEEE J. Biomed. Health* **2015**, *19*, 1193.
[3] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, D. Barton, *Harvard Business Rev.* **2012**, *90*, 61.
[4] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, B. Vaughan, D. Preuss, R. Leinonen, M. Shumway, S. Sherry, *Nat. Methods* **2012**, *9*, 459.
[5] P. O'Mara, A. Farrell, J. Bones, K. Twomey, *Talanta* **2018**, *176*, 130.
[6] N. Bourguignon, C. Attallah, P. Karp, R. Booth, A. Peñaherrera, C. Payés, M. Oggero, M. S. Pérez, G. Helguera, B. Lerner, *Integr. Biol.* **2018**, *10*, 136.
[7] M. Micheletti, G. J. Lye, *Curr. Opin. Biotechnol.* **2006**, *17*, 611.
[8] V. Sandner, L. P. Pybus, G. McCreath, J. Glassey, *Biotechnol. J.* **2019**, *14*, 1700766.
[9] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (No. 519.9 A53). Wiley, New York **1962**.
[10] N. R. Bohidar, F. A. Restaino, J. B. Schwartz, *J. Pharm. Sci.* **1975**, *64*, 966.
[11] J. F. MacGregor, T. Kourti, *Chemom. Intell. Lab. Syst.* **1995**, *28*, 3.
[12] A. F. Silva, J. Vercruysse, C. Vervaet, J. P. Remon, J. A. Lopes, T. De Beer, M. C. Sarraguça, *J. Pharm. Sci.* **2019**, *108*, 439.
[13] S. Wold, P. Geladi, K. Esbensen, J. Öhman, *J. Chemom.* **1987**, *1*, 41.
[14] P. Nomikos, J. F. MacGregor, *AIChE J.* **1994**, *40*, 1361.
[15] L. Eriksson, *Introduction to Multi-and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*, Umetrics AB, Umeå, Sweden **1999**.
[16] J. M. Lee, C. K. Yoo, I. B. Lee, *J. Biotechnol.* **2004**, *110*, 119.
[17] S. Goldrick, W. Holmes, N. J. Bond, G. Lewis, M. Kuiper, R. Turner, S. S. Farid, *Biotechnol. Bioeng.* **2017**, *114*, 2222.
[18] F. Arteaga, A. Ferrer, *J. Chemom.* **2002**, *16*, 408.
[19] A. Folch-Fortuny, F. Arteaga, A. Ferrer, *Chemom. Intell. Lab. Syst.* **2015**, *146*, 77.
[20] H. Zhang, B. Lennox, *J. Process Control* **2004**, *14*, 41.
[21] M. J. Carrondo, P. M. Alves, N. Carinhas, J. Glassey, F. Hesse, O. W. Merten, M. Micheletti, T. Noll, R. Oliveira, U. Reichl, A. Staby, *Biotechnol. J.* **2012**, *7*, 1522.
[22] V. Konakovsky, C. Clemens, M. Müller, J. Bechmann, M. Berger, S. Schlatter, C. Herwig, *Bioengineering* **2016**, *3*, 5.
[23] M. Sokolov, J. Ritscher, N. MacKinnon, J. M. Bielser, D. Brühlmann, D. Rothenhäusler, G. Thanei, M. Soos, M. Stettler, J. Souquet, H. Broly, *Biotechnol. J.* **2018**, *13*, 1700461.
[24] FDA., *Final Report on Pharmaceutical cGMPs for the 21st Century–A Risk-Based Approach*, DHHS, Rockville, MD **2003**.

[25] FDA. *Guidance for Industry: PAT-A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*, DHHS, Rockville, MD **2004**.

[26] J. Branke, S. S. Farid, N. Shah, *Cell Gene Ther. Insights* **2016**, *2*, 263.

[27] M. Mellinger, *Chemom. Intell. Lab. Syst.* **1987**, *2*, 29.

[28] L. Mears, R. Nørregård, S. M. Stocks, M. O. Albæk, G. Sin, K. V. Gernaey, K. Villez, *Comput.-Aided Chem. Eng.* **2015**, *37*, 1667.

[29] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham, *Multivariate Data Analysis*, Pearson Education Limited, Edinburgh Gate, Harlow Essex **2014**, https://is.muni.cz/el/1423/podzim2017/PSY028/um/_Hair_-_Multivariate_data_analysis_7th_revised.pdf.

[30] FDA. *Data Integrity and Compliance with CGMP Guidance for Industry*. DHHS, Rockville, MD **2016**.

[31] S. Goldrick, K. Lee, C. Spencer, W. Holmes, M. Kuiper, R. Turner, S. S. Farid, *Biotechnol. J.* **2018**, *13*, 1700607.

[32] L. H. Chiang, R. Leardi, R. J. Pell, M. B. Seasholtz, *Chemom. Intell. Lab. Syst.* **2006**, *81*, 109.

[33] G. B. Limentani, M. C. Ringo, F. Ye, M. L. Bergquist, E. O. McSorley, *Anal. Chem.* **2005**, *77*, 221A.

[34] C. DiCesare, M. Yu, J. Yin, W. Zhou, C. Hwang, J. Tengtrakool, K. Konstantinov, *BioProcess Int.* **2016**, *14*, 18.

[35] T. Tajsoleiman, L. Mears, U. Krühne, K. V. Gernaey, S. Cornelissen, *Trends Biotechnol.* **2019**, *37*, 697.

[36] M. Brunner, J. Fricke, P. Kroll, C. Herwig, *Bioprocess Biosyst. Eng.* **2017**, *40*, 251.

[37] D. Lakens, TOST Equivalence Testing R Package (TOSTER) and Spreadsheet. R-Bloggers, https://www.r-bloggers.com/tost-equivalence-testing-r-package-toster-and-spreadsheet/ (accessed: December 2016).

[38] H. Cooper, L. V. Hedges, J. C. Valentine, *The Handbook of Research Synthesis and Meta-Analysis*, Russell Sage Foundation, New York **2009**.

[39] C. K. Enders, *Applied Missing Data Analysis*, Guilford Press, New York **2010**.

[40] R. J. Little, D. B. Rubin, *Statistical Analysis with Missing Data*, Vol. 793, Wiley, Hoboken, NJ **2019**

[41] J. Y. Nancy, N. H. Khanna, K. Arputharaj, *Comput. Statist. Data Anal.* **2017**, *112*, 63.

[42] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. I. Matsubara, S. Ishii, *Bioinformatics* **2003**, *19*, 2088.

[43] B. Van Snick, J. Dhondt, K. Pandelaere, J. Bertels, R. Mertens, D. Klingeleers, G. Di Pretoro, J. P. Remon, C. Vervaet, T. De Beer, V. Vanhoorne, *Int. J. Pharm.* **2018**, *549*, 415.

[44] J. A. Westerhuis, T. Kourti, J. F. MacGregor, *J. Chemom.* **1998**, *12*, 301.

[45] B. Lennox, G. A. Montague, H. G. Hiden, G. Kornfeld, P. R. Goulding, *Biotechnol. Bioeng.* **2001**, *74*, 125.

[46] P. R. Nelson, P. A. Taylor, J. F. MacGregor, *Chemom. Intell. Lab. Syst.* **1996**, *35*, 45.

[47] P. Facco, F. Doplicher, F. Bezzo, M. Barolo, *J. Process Control* **2009**, *19*, 520.

[48] R. Bakirov, B. Gabrys, D. Fay, *Comput. Chem. Eng.* **2017**, *96*, 42.

[49] P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1.

[50] H. Le, N. Vishwanathan, N. M. Jacob, M. Gadgil, W. S. Hu, *Biotechnol. Lett.* **2015**, *37*, 1553.

[51] B. H. Junker, *J. Biosci. Bioeng.* **2004**, *97*, 347.

[52] F. Garcia-Ochoa, E. Gomez, *Biotechnol. Adv.* **2009**, *27*, 153.

[53] Z. Xing, B. M. Kenty, Z. J. Li, S. S. Lee, *Biotechnol. Bioeng.* **2009**, *103*, 733.

[54] F. R. Schmidt *Appl. Microbiol. Biotechnol.* **2005**, *68*, 425.

[55] M. Meyners, *M. Food Qual.* **2012**, *26*, 231.

[56] E. K. Ritchie, E. B. Martin, A. Racher, C. Jaques, *J. Biotechnol.* **2017**, *251*, 160.

[57] C. A. Mara, R. A. Cribbie, *Commun. Stat.—Simul. C.* **2012**, *41*, 1928.

[58] C. Chen, N. Rathore, W. Ji, A. Germansderfer, *BioPharm. Internat.* **2010**, *23*, 2.

[59] EMA. *Statistical Methodology for the Comparative Assessment of Quality Attributes in Drug Development*, EMA, London **2017**.

[60] E. Walker, A. S. Nowacki, *J. Gen. Inter. Med.* **2011**, *26*, 192.

[61] J. J. Priola, N. Calzadilla, M. Baumann, N. Borth, C. G. Tate, M. J. Betenbaugh, *Biotechnol. J.* **2016**, *11*, 853.

[62] A. D. Bandaranayake, S. C. Almo, *FEBS Lett.* **2014**, *588*, 253.

[63] A. Chen, R. Chitta, D. Chang, A. Amanullah, *Biotechnol. Bioeng.* **2009**, *102*, 148.

[64] J. Mante, N. Gangadharan, D. J. Sewell, R. Turner, R. Field, S. G. Oliver, N. Slater, D. Dikicioglu, *Bioprocess Biosyst. Eng.* **2019**, *42*(4).

[65] W. F. Velicer, S. M. Colby, *Educ. Psychol. Meas.* **2005**, *65*, 596.

[66] A. Casablancas, X. Gámez, M. Lecina, C. Solà, J. J. Cairó, F. Gòdia, *J. Chem. Technol. Biotechnol.* **2013**, *88*, 1680.

[67] R. J. Fleischaker, A. J. Sinskey, *Eur. J. Appl. Microbiol. Biotechnol.* **1981**, *12*, 193.

[68] C. M. Kussow, W. Zhou, D. M. Gryte, W. S. Hu *Enzyme Microb. Technol.* **1995**, *17*, 779.