Vaccine 37 (2019) 7081–7089

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

# Accelerating bioprocess development by analysis of all available data: A USP case study

Diego A. Suarez-Zuluaga [a,1], Daniel Borchert [b,c,1], Nicole N. Driessen [a], Wilfried A.M. Bakker [a], Yvonne E. Thomassen [a,*]

[a] Intravacc, Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven, the Netherlands
[b] Exputec GmbH, Mariahilferstraße 147/2/2D, 1150 Vienna, Austria
[c] Vienna University of Technology, Research Area Biochemical Engineering, Gumpendorferstrasse 1a, 1060 Vienna, Austria

ARTICLE INFO

*Article history:*
Available online 20 July 2019

*Keywords:*
Multivariate data analysis
Comprehensive data gathering
Cause and effect analysis
Inactivated poliovirus vaccine
Design of experiments

ABSTRACT

Bioprocess development generates extensive datasets from different unit operations and sources (e.g. time series, quality measurements). The development of such processes can be accelerated by evaluating all data generated during the experimental design. This can only be achieved by having a clearly defined data logging and analysis strategy. The latter is described in this manuscript. It consists in a combination of a feature based approach along with principal component analysis and partial least square regression.

Application of this combined strategy is illustrated by applying it in an upstream processing (USP) case study. Data from the development and optimization of an animal component free USP of Sabin inactivated poliovirus vaccine (sIPV) was evaluated. During process development, 26 bioreactor runs at scales ranging from 2.3 to 16 L were performed. Several operational parameters were varied, and data was routinely analyzed following a design of experiments (DoE) methodology.

With the strategy described here, it became possible to scrutinize all data from the 26 runs in a single data study. This included the DoE response parameters, all data generated by the bioreactor control systems, all offline data, and its derived calculations. This resulted in a more detailed, reliable and exact view on the most important parameters affecting bioreactor performance.

In this case study, the strategy was applied for the analysis of previously produced data. Further development will use this data analysis methodology for continuous enhancing and accelerating process development, intensified DoE and integrated process modelling.


## 1. Introduction

Data analysis and mathematical modeling [1] have been previously applied to Sabin inactivated polio vaccine (sIPV) manufacturing. This was done to increase process understanding [2], optimize [3] and support scale-up [4,5]. On those studies, the performed experiments and resulting models focused on well-known process parameters. Nevertheless, deviations in other process parameters were also identified. In the case study described here, we used an experimental dataset produced during upstream process (USP) development. This dataset was generated following a design of experiments (DoE) approach in which the evaluated parameters and selected responses were defined based on previous process knowledge. Even though only two parameters were varied, it was possible to gain knowledge on the production process. However, the variance in the response was not thoroughly explained by applying the standard DoE approach.

To increase process understanding, to establish a more comprehensive process optimization and to accelerate the process development, we performed data analysis by evaluating all the recorded data holistically. This entails that data available from the supervisory control and data acquisition (SCADA) system linked to the bioreactor (e.g. temperature, pH, and oxygen levels), the low frequency recorded offline measurements (e.g. cell, substrate, and product concentrations), and the initial and final one point measurements and calculations (e.g. harvest turbidity and cell growth rates), were gathered in one dataset for a comprehensive data analysis and evaluation.

For identification of the causes of variance in the selected response parameters and for holistic assessment of the available raw data, we performed a cause and effect analysis (CEA) [6] as developed by Borchert et al. [7]. This permitted to identify the

* Corresponding author.
*E-mail address:* yvonne.thomassen@intravacc.nl (Y.E. Thomassen).
[1] Diego A. Suarez-Zuluaga and Daniel Borchert contributed equally to this work.

https://doi.org/10.1016/j.vaccine.2019.07.026
0264-410X/

process parameters which affect the target response. The idea of the applied workflow is to use and increase the available process knowledge [8]. This allows to generate new hypotheses regarding the cause of the variance of the response.

This first step in the CEA consisted of aligning the datasets and performing data mining. Here, a combination of raw data analysis (RDA) [9] and feature based analysis (FBA) [10] was applied. This resulted not only in increased process knowledge but also in a reduction of the effort needed to perform information mining [7]. Both tools (FBA and RDA) can also be independently used to perform a CEA [9,11], and their combination leads to an effective and strong data analysis tool.

The carried out data mining step resulted in a gapless data matrix, which was used as basis for the final multivariate data analysis (MVDA). Prior to the MVDA, standard univariate data analysis approaches were applied to test data validity. Fig. 1 depicts the followed methodology.

As mentioned above, the USP development of a polio vaccine production process was selected to apply the aforementioned data process methodology. Many efforts have been made to optimize inactivated polio vaccine production contributing to global polio eradication aims. Optimization efforts have focused both on reduction of biosafety risks by the use of for instance attenuated Sabin strains [5,12] or genetically engineered strains [13] and reduction of vaccine manufacturing cost by process optimization [3]. In these process optimization studies, a wealth of data is frequently generated, yet conclusions are usually drawn based on a selection of process parameters that were chosen for the specific study. Hidden information can be found in data of, for instance, bioreactor controllers. Data used here, was generated in such an optimization project in which poliovirus was produced on Vero cells grown in animal component free media [14].

By applying the workflow from Fig. 1, we identified the most critical time series variables to control the response variance (product yield). This new process information was evaluated and further investigated. This resulted in the identification of deviations within the critical time series variables which potentially caused the observed differences. These events were mined from the time series data. This led to the generation of a data matrix that contained only time independent variables which definitely influenced the response. Finally, the complete cause for the observed variance in the response was identified using a straightforward MVDA approach.

By performing this case study, we describe a data analysis methodology which allows to identify the cause for the variance of a certain response. This approach uses well-known statistical tools [15–17] and comprises a holistic evaluation of all available raw on– and offline measurements. Hence, the main advantage of the described methodology is that it uses all available data to identify the significant variables and uses them to generate a model which describe the variation on the selected response.

The results of the case study show that, even though only virus production process parameters were evaluated during the DoE, cell culture parameters also influenced product yield. This additional knowledge leads to consider both, cell culture and virus production, as equally important for increasing yields during sIPV production.

## 2. Materials and methods

### 2.1. Experimental conditions

Vero cells (obtained from WHO (10-87) originally derived from ATCC (CCL-81)), growing adherent to Cytodex 1 (GE Healthcare) microcarriers, were used to produce Sabin type 2 poliovirus (P712 Ch2ab-$KP_2$). In total, 26 bioreactors were run in batch mode in an animal component free medium environment essentially as described previously [14]. The process was split into two phases. First, a cell culture phase (pH 7.2 ± 0.1; DO 50 ± 0.1%; T 37 ± 0.1 °C) with VP-SFM medium (Thermo Fisher Scientific). And second, a virus production phase (pH 7.4 ± 0.1; DO 25 ± 0.1%; T 32.5 ± 0.1 °C) with M199 medium (Thermo Fisher Scientific). The key performance indicator, also selected as response variable (D-antigen concentration [18] further referred as product yield), was measured once at the end of the upstream process (USP) [13].

Studies were carried out in glass bioreactors (Applikon or Sartorius) with working volumes ranging from 2.3 L to 16 L. Temperature, pH, DO, gas flows and stirring speeds were logged using the bioreactor control system (Bioexpert and MFCS for Applikon and Sartorius controllers, respectively). We assume that the used controllers monitor the data similar and are therefore not taken into account for the performed data analysis. DO was controlled via a cascade. It consisted on using air flow; and if needed, oxygen, carbon dioxide and increased stirring speed (max. 130 rpm) to reach the desired DO set-point. Time of infection (TOI; when virus is added to the Vero cell culture) and multiplicity of infection (MOI, ratio viral particles:cells) were varied. TOI between 96 and 120 h and MOI between 0.001 and 0.01.

### 2.2. Software

Two commercial available software tools were used to perform the analysis. SIMCA version 13.0.3.0 (Umetrics AB, Umea, Sweden)

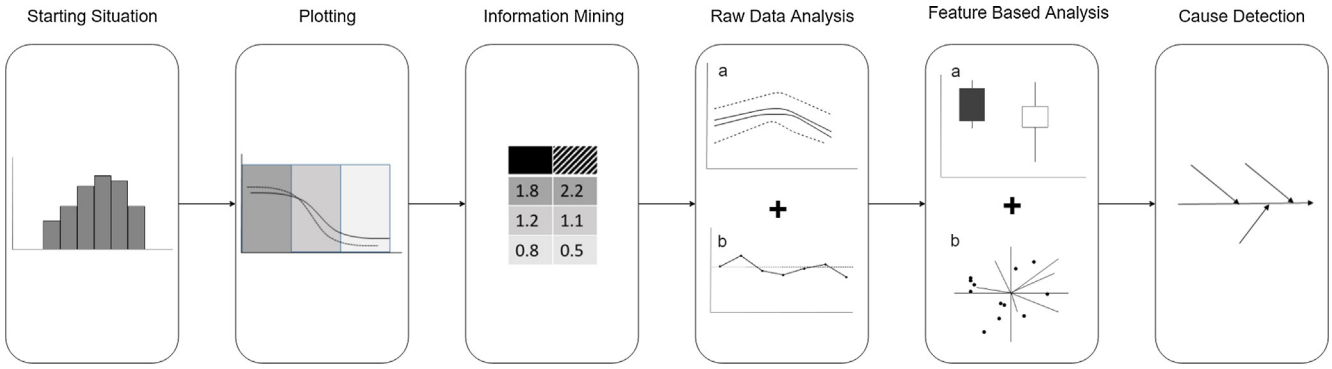


**Fig. 1.** Schematic workflow of the applied cause and effects analysis. At the very beginning, unknown variance of a selected response was identified (Starting Situation). Subsequent plotting of all available time series variables (Plotting) leads to the identification of differences within certain variables ranges. Information, of the detected deviations, were extracted from time series data (Information Mining). Next, uni- and multivariate analysis on time series data (Raw Data Analysis, a and b, respectively) results in further process understanding and detection of the most important variables effecting the selected response. The following uni- and multivariate analysis cycle (Feature Based Analysis, a and b, respectively) used just one point measurements and extracted information for the identification of the variables significantly effecting the response variable (Cause Detection).

was used for Raw Data Analysis (RDA). inCyght® Fermentation Expert version 2018.04 (Exputec GmbH, Vienna, Austria) was used to perform Feature Based Analysis (FBA). The performed uni- and multivariate statistical analysis tools were already implemented in the used software tools. Prior to data analysis with SIMCA and inCyght®, data was preprocessed with MS Excel 2016 (Microsoft, Redmond, WA) and Python 3.3 (Python Software foundation, https://www.python.org/).

### 2.3. Statistical methods

Principal component analysis (PCA) and partial least square regression (PLS) were used as statistical evaluation tools. These standard methods were already available in the software tools. A 7-fold cross validation was applied to select the best PLS subset [7]. Here, the subset with the lowest root mean square error of cross validation (RMSCV) was considered as the best [19]. Robust slope was calculated using the integrated robust linear regression function within inCyght® version 2018.04, applying the bisquare weighted function.

## 3. Results and discussion

Below, the application of a data analysis methodology which can be used to accelerate processes development is described. In this case study, the focus is on the USP of polio vaccine production [12]. Although different poliovirus production processes were examined previously [1,2], the entire product yield variance was not explained by the resulting models. By using this holistic approach, (Fig. 1) the complete dataset originated from 26 bioreactors was evaluated using state of the art statistical tools. This determined the process parameters which affect product yield and generated a hypothesis that could be tested in an experimental setup.

### 3.1. Plotting and information mining 1

All available data from different sources was collected. This resulted in a dataset consisting of 29 parameters (17 time series and 12 one-point measurements/set-points). At the end of the data mining, and as starting point for the methodology applied here, an unaligned dataset with 26 batches was obtained.

To compare the data, the first step is to align the dataset. This means that start/end times were evaluated, and process phases were defined. Data was aligned based on the main bioreactor operational phases, cell growth and virus production. As indicated by the featured based approach for data analysis [7], the next step consisted on extracting features that describe differences within certain time series.

Overlay plots for each variable and process phase were created to determine the existing variations of certain variables and extract information that described them. Focus was put on variables that showed a clear difference in the recorded data (for instance oxygen flow, Fig. 2) and/or those variables which are known to explain different process performance (e.g. metabolites concentration). In Fig. 2, it can be observed how the oxygen flow starts to decrease at different time points for each batch. In the example, two features were extracted, the length of time until the flow started to decrease and the slope of the decrement. These characteristics, defined as features, describe the moment and speed at which cells need less oxygen to keep the dissolved oxygen (DO) at its set-point. Or in other words, a description of how fast cells were dying during virus replication. Further, additional differences were detected within other time series variables. Other extracted features were for instance: the slope at which air flow decreases during cell growth and the slopes for the consumption/production of the measured metabolites (glucose, glutamate, glutamine, ammonia and lactate).

This information mining step could result in a high number of new features created per batch. To identify relevant differences in variables, score plots (Fig. 3) were generated for this purpose.

### 3.2. Raw data analysis and information mining 2

Score plots allow graphical determination of whether a variable has a positive or negative influence over a response in a given moment of time. Fig. 3 displays the score plot for dissolved oxygen (DO) in relation to response product yield. During the cell culture phase (Fig. 3a) a net positive or negative effect cannot be observed. However, during the virus production phase (Fig. 3b) it indicates a

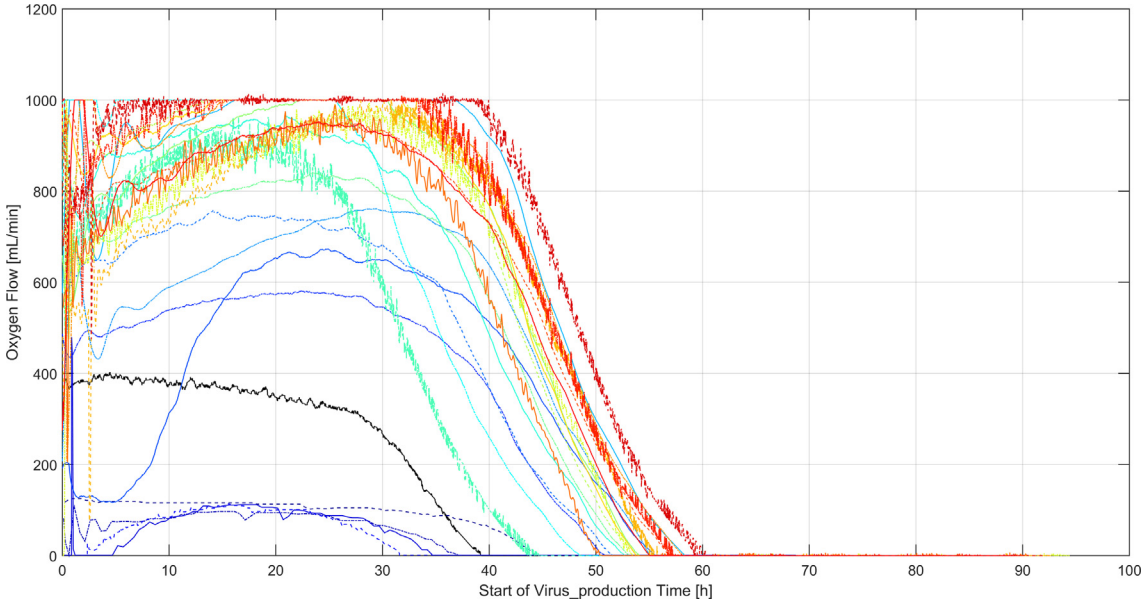


**Fig. 2.** Oxygen flow during the virus production phase. Each line represents a single batch. The line can be described by two features/information: the length of time until the oxygen flow starts to decrease and the slope of the decrement.

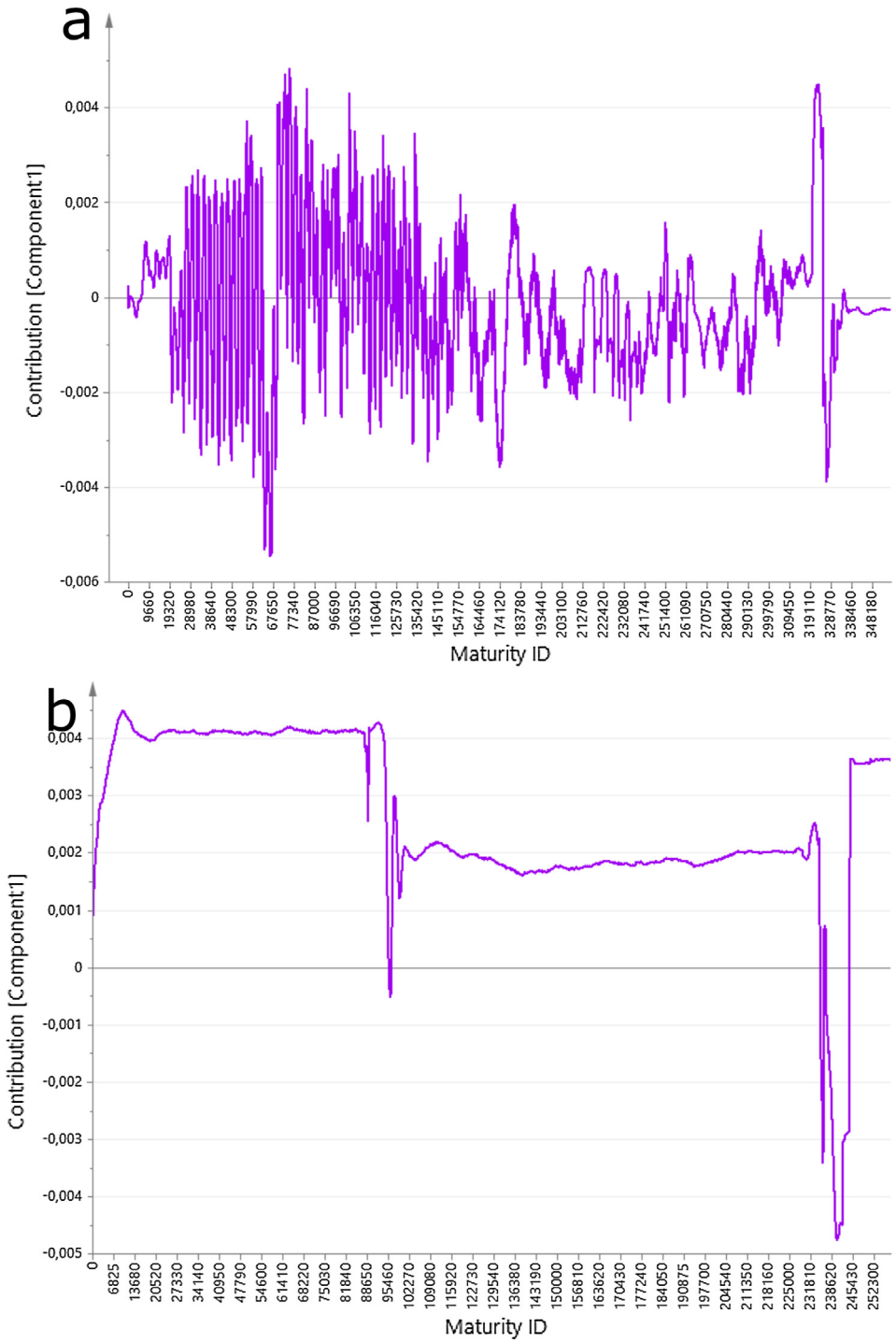


**Fig. 3.** Score plot for Dissolved oxygen in relation to response (product yield). "a" and "b" correspond to the cell culture and virus production phases respectively. In panel "a" it can be seen that this variable does not seem to have a positive or a negative effect, while in panel "b" a positive effect is indicated. Features were extracted to describe this phenomenon and it resulted in data reduction.

positive effect. By creating these plots for all-time series data, significant differences on time series variables were detected and features that describe these differences were extracted. This included slope and duration of carbon dioxide flow decrease during the first hours of the cell growth phase, standard deviation and median DO value during cell growth and virus production phases, the median of air flow rate at beginning of virus production phase and the median of pH value during virus production phase.

Information was mined from the time series and the data matrix was reduced to features. However, not all features are descriptors of time series variation, but they are also one-point measurements (e.g. product yield or amount of substrate added), calculations (e.g. growth rate, specific productivity) and design of experiments defined factors (e.g. MOI, TOI). Next, each variable was checked for completeness and outliers and variables without any variation were removed (e.g. DO, temperature and pH

set-points). Hence, a gapless features table required for MVDA was obtained.

### 3.3. Feature based analysis 1: Principal component analysis

In order to gain multivariate process understanding and to reduce noise a principal component analysis (PCA) was performed. PCA is used to identify collinearity and not used for explaining the entire process variance. Fig. 4 (Biplot) presents the result of the analysis for the latest obtained dataset. 25 principal components are required to describe 100% or the process variance. The first two components describe 37% of the process variance (principal component 1 and 2 with 19% and 18% respectively) and were used to identify collinear features (i.e. features that point towards the same direction). Such features were assessed and only those which can be controlled were further considered. Doing so, reduced the dataset and prevented collinearity issues on the following steps. One example is "cell density start viable" (x≈-0.18, y≈0.07 in Fig. 4) and "robust slope ammonium" (x≈-0.08, y≈0.02 in Fig. 4). The first one is a feature that can be controlled while the second is an outcome of selected process settings. Consequently, "robust slope ammonium" was not further considered. By performing this analysis, 12 out of 33 features were eliminated from the data matrix.

### 3.4. Feature based analysis 2: Partial least squares

A partial least squares (PLS) regression was performed to determine which features have a significant effect to the selected response (product yield). PLS results can be used to generate hypothesis to be studied in follow up experiments. However, the resulting model must be critically evaluated, as its validity is dependent on the quality of the data used and might include features which are not controllable.

Fig. 5 shows the coefficient plot of the PLS results, this is the graphical representation of the resulting model (Equation (1)) which explains 74% of the variance on the target response (product yield), the $Q^2$ of this model is 35%. A $R^2$ to $Q^2$ difference of more than 30% means overfitting within the model [20]. Which indicates that the model needs to be reduced. On the x-axis features that are part of the model are depicted. The blue bars represent the relative importance of each feature while the error bar indicates the standard error. If the latter includes zero, then the variable is non-significantly effecting the response. Therefore, it can be observed that eight features have a significant impact on the target response.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

***Equation (1). Multilinear regression formula.*** *Y is the response variable, which can be described by the intercept ($\beta_0$) and the significant variables ($\beta_n X_n$), where $\beta_n$ is the slope and $X_n$ the values the certain variable n of one of the significant variables out of PLS regression. $\epsilon$ represents the variance of the residual error of all other factors not accounted for. We assume that this error is normal distributed with 0 variance and therefore neglected from the analysis.*

## 4. Combining PLS results with process knowledge

Results from the previous step could be considered the main result of the presented methodology. As this is the model that best explains the variance on the target response. However, overfitting is present and not all features that have a statistical impact on the product yield are controllable. Or, they may not be controllable within certain economic, time and/or technological constraints. Hence, the next step in this data evaluation consisted of creating a model which contains only controllable features. Only then, a hypothesis, that can be tested in an experimental design, can be generated. Table 1 summarizes the characteristics of the significant features.

Once the uncontrollable features were removed, PLS was performed again. Results from the second PLS (Fig. 6) are similar to those previously obtained. Cell density at TOI, multiplicity of infection and cell growth rate again appear as most significant. The new model explains 40% of the variance on the product yield (compared to 74% of the previous model). But, as some features were removed,

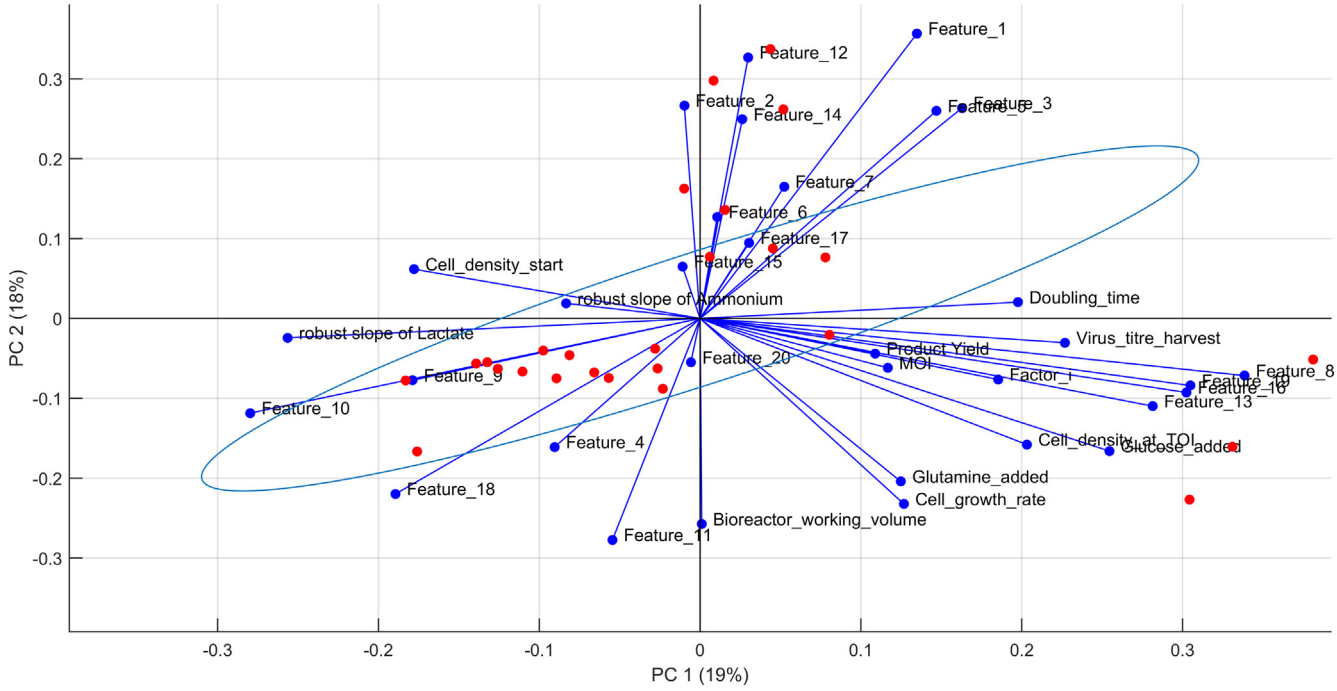


**Fig. 4.** Principal components analysis. PC1 and PC2 denote principal component 1 and 2 respectively. Dots at the end of the lines correspond to the features. Other dots correspond to the batches. The circle shows the 95% confidence interval of the population mean.

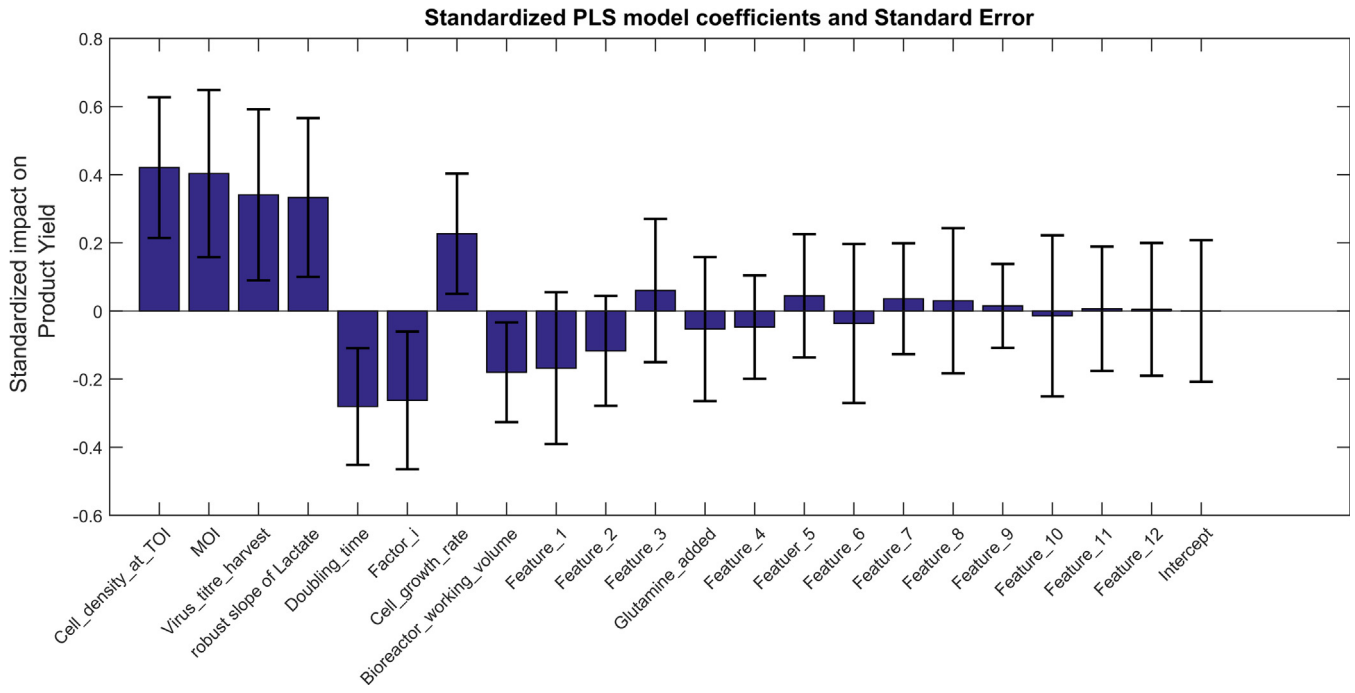


**Fig. 5.** Original Partial least squares. This plot indicates the relation of the variables with the product yield. Variables are ranked in order of statistical significance (high to low). The direction of the blue bar indicates the direction of the dependency. The black error bar represents the standard error of certain variables, if this does not include zero, then the variable is considered as statistically significant influencing the response variable. The first eight variables are seen as significant.

**Table 1**
Features with significant impact on product yield. Ranked from the largest to the lowest impact.

| Feature | Characteristic |
|---|---|
| Cell density at time of infection (TOI) | Controllable |
| Multiplicity of infection (MOI) | Controllable |
| Virus titer | Result. Cannot be directly controlled |
| Lactate slope | Result. Cannot be directly controlled |
| Factor i | Out of scope |
| Cell growth rate | Is a result. But, indirectly controlled. |
| Doubling time | Correlated with cell growth rate. Not further considered. |
| Bioreactor volume | Scale up effect. 2.3L vs 10L. Out of scope |

this reduction was expected. An increased explanation of the variance could potentially be obtained by including categorical variables on the data analysis strategy (such as medium lot number, day of production or operator). The $Q^2$ value of the updated model also decreases to 14%. The difference of $R^2$ to $Q^2$ is here less than 30%, which can be seen as no present overfitting in the model.

### 4.1. Cause detection

A PLS model which includes only controllable items was described in the previous section. There, cell growth rate was considered as an indirectly controllable feature. Therefore, this variable was investigated and those parameters that affected it were identified. To do this, cell growth rate was selected as response parameter and the methodology was re-applied using only data from the cell growth phase. An obtained PLS model which explains 71% of the variance on the cell growth rate (data not shown) was obtained. It indicates that glucose and glutamine addition during cell growth phase, as well as median of DO are the controllable parameters with significantly influence the cell growth rate. If the results of this cell growth rate model are combined with the product yield PLS (Fig. 6), then a comprehensive cause and effect diagram can be created (Fig. 7).

The new cause and effect diagram (Fig. 7) shows the parameters that, based on all recorded data, significantly influenced the product yield (poliovirus D-antigen). On the right side (Fig. 7b) of the diagram are virus production related parameters (MOI and cell density at TOI). These parameters correspond to those evaluated in the experimental design from which this data was generated (see materials and methods section). Thus, confirming significant influence of the parameters that were selected for the DoE optimization studies. Fig. 7a on the other hand, shows cell growth related parameters. These were not considered in the performed experiments and identifying them as influential factors opens the way to new product yield improvement experiments. Hence in this way, the diagram provides the basis for hypotheses which can be tested experimentally

Previous findings [21–23] confirm the results displayed in Fig. 7a. They show that glucose, glutamine and oxygen influence Vero cell growth. Huang et al. (2006) [24] additionally reported the glutamine influenced cell growth. Similar behavior was reported by Cruz et al. (1999) [25] by using baby hamster kidney cells as mammalian host. Including this knowledge in combination with our results, it can be stated that glutamine addition might be beneficial. Nevertheless, its addition might result in ammonia compounds production which could negatively affect growth.

The result in Fig. 7b was not described by Ursache et al. (2015) [1]. Nevertheless, other studies [12,26,27] identified a correlation of TOI and MOI with product yield from Vero cells and Baculovirus. This confirms our findings and highlights the importance of well controllable MOI and TOI during virus production processes.

As mentioned above, set-point variables (DO, pH and temperature) were not considered within the performed CEA. This is because they were not varied and consequently could not explain any of the observed variance. Future experiments should also look at their influence on the selected response parameter. Additional work should also be performed to confirm the hypotheses gener-

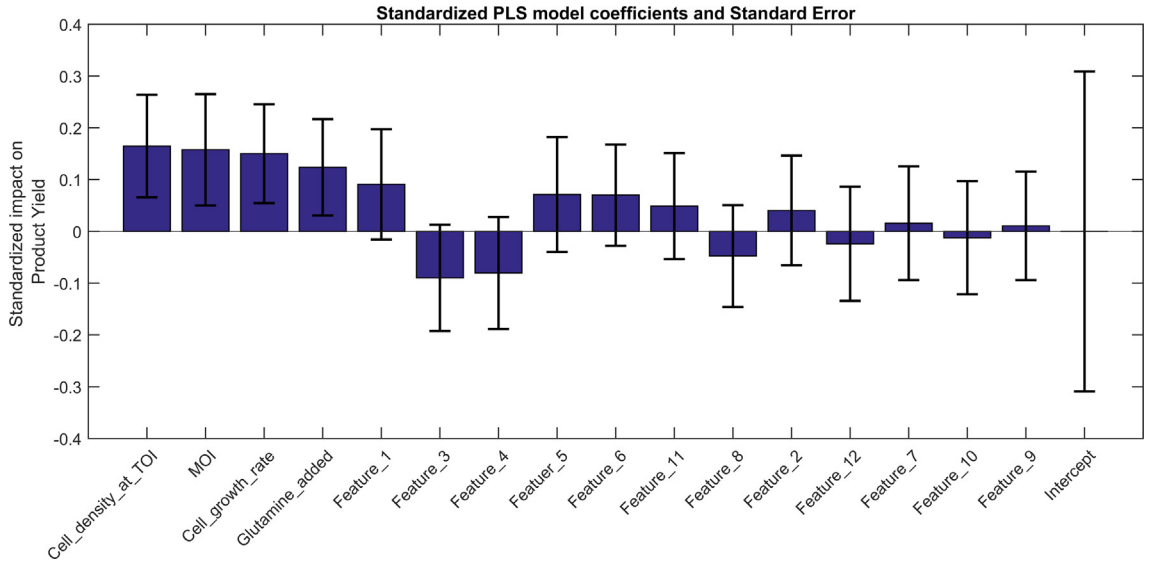


**Fig. 6.** Partial least squares obtained when combining initial results with process knowledge. The blue bar indicates the direction of the dependency, and the black error bar represents the standard error of certain variable. If the error bar includes zero, the variable is considered as not significantly influence the response variable.

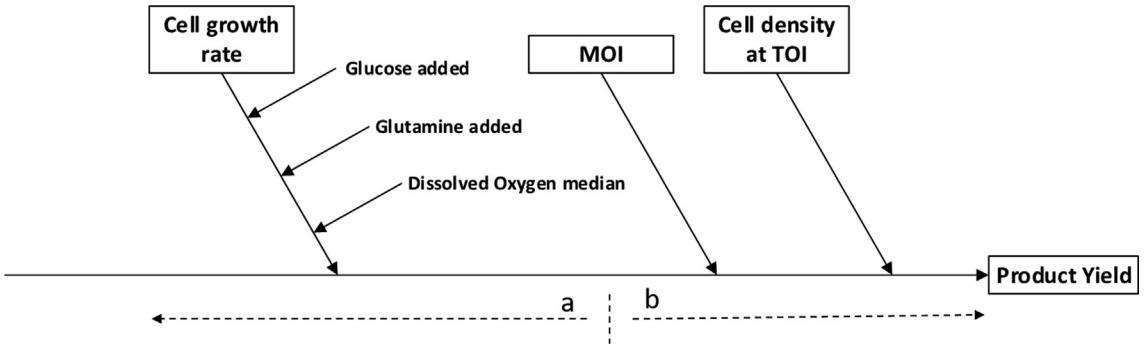


**Fig. 7.** Case study result. Main parameters influencing product yield, directly and indirectly. Proximity to end of the horizontal arrow represents higher influence on the response variable. All parameters are above the line, this means that they have a positive influence on their response. "a" and "b" indicate which parameters influence cell culture and virus culture phase respectively.

ated and to obtain further theoretical understanding of why some parameters do or do not influence the target response variable. Finally, this data analysis strategy was completed after performing a DoE. In the future, this analysis should be performed before. This will allow to create an experimental design which considers the most relevant variables. However, it must be noticed that availability of an historical dataset is a requirement. In other words, data including variance is essential for performing a comprehensive analysis; which subsequently will help to perform a successful DoE.

### *4.2. Comparison to previous performed MVDA (Importance of process knowledge)*

In order to detect the parameters effecting the selected response variable, state of the art statistical tools were applied. Thomassen et al. (2010) [2] also performed a MVDA on historical polio production data using similar methods (PCA and PLS). That study analyzed production data originated from 350L and 750L bioreactors and identified outliers and the source of their variation. However, a model describing the response variance from the USP was not derived. The herein demonstrated case study by applying the CEA approach suggested by Borchert et al. (2018) focused more on process knowledge and reasonable information mining from time series data than the previous MVDA. This knowledge and data structure allows to perform a comprehensive data analysis using well known and simple statistical methods. Obtaining, as a result, the identification of the parameters which significantly influence the select response.

Standard statistical methods were developed to deal with a high number of observations (n) over a small number of parameter (p). This was done to facilitate calculation before the computer era. Additionally, in case of generating a regression model with data from matrix with less n than p, the change to over fit for model increases drastically [28]. We demonstrated that this well know problem (n << p), as we have it, can be solved with standard tools in the statistical manner by including process knowledge. From a statistical point of view, more complex algorithms like shrinkage methods like, Least Absolute shrinkage and Selection operator (LASSO) or ridge regression, are preferred tools to handle a n << p data matrix [29]. Those methods are used in order to find the best subset by including a tuning of parameters to minimize the residual sum of squares [30,31]. Although the fast application of such algorithms can be seen as positive, the underlying calculation might be complex and difficult to understand without deep statistical background.

We demonstrated, that it is possible to use well known and simple statistical algorithms (for instance PCA and PLS regression) for a CEA. We reduced the dimension by previously applying a standard PCA and identified significant parameters affecting a response by applying a standard PLS. Our main finding was that process knowledge is crucial for comprehensive information mining and that a well gathered data set is mandatory for a successful analysis.

## 5. Conclusion and outlook

The aim of this study was to illustrate the benefits of a CEA analysis by using available raw data. The results of the applied workflow show that by using all available raw data to detect hidden process knowledge new hypotheses for accelerating bioprocess development could be generated. We were able to gather all the available data (high frequent records from SCADA, low frequent offline measurements as well as one-point measurements) in one data table. Then, values which described the difference in the bioreactors performance were extracted by creating overlay and score plots. Such plotting and information mining was not previously possible because of the limitation in using all available recorded data comprehensively. The holistic database facilitates identification of process deviations within time series, accelerating process analysis. The thereby resulting dataset, including only time independent one point measurements, can then be evaluated with a combination of simple uni- and multivariate data analysis tools. This leads to the identification of new hypotheses. This level of process understanding permits to use the created hypothesis to perform targeted experiments and to accelerate process development by testing only those variables that effectively have an effect on the response.

On the presented case study, only 40% of the variance on the selected response was identified and hypotheses that explain the remaining variance were created. The generated process knowledge can be used, for example, for:

- Model based experimental design: The use of mechanistic models in combination with the herein presented statistical analysis can be used to design additional experiments. The aim should be, to detect the remaining hidden reason for the response variance and to create models that explain it.
- Design of experiments and beyond: Well know experimental design methods as standard design of experiments (DoE) or augmented DoEs can be planned and performed including the new process knowledge. Advanced DoE, like Bayesian experimental design, will be able to use the results from the MVDA as well the information from previously performed DoEs.
- Transferring the gained know how: Interlink the knowledge from this USP case study with the subsequent performed downstream process (DSP). The holistic assessment of USP and DSP to the final product can be used to refine the present process over unit operations. This refinement, will accelerate future process development of new products.

The herein resulted data matrix and process knowledge can be holistically used for any of the above-mentioned next steps in process development. Finally, independent of the method, it will result in a more rational design of new experiments for process optimization and acceleration of future development.

## Funding

This project has received funding from the Dutch Ministry of Economic Affairs under PPP-Allowance under the TKI-programme Life Sciences & Health.

## Declaration of Competing Interest

The authors declare no conflict of interest.

## References

[1] Ursache RV, Thomassen YE, van Eikenhorst G, Verheijen PJT, Bakker WAM. Mathematical model of adherent Vero cell growth and poliovirus production in animal component free medium. Bioprocess Biosyst Eng 2015;38:543–55. https://doi.org/10.1007/s00449-014-1294-2.
[2] Thomassen YE, van Sprang ENM, van der Pol LA, Bakker WAM. Multivariate data analysis on historical IPV production data for better process understanding and future improvements. Biotechnol Bioeng 2010;107:96–104. https://doi.org/10.1002/bit.22788.
[3] Thomassen YE, Bakker WAM. sIPV process development for costs reduction. Vaccine 2015;33:4307–12. https://doi.org/10.1016/j.vaccine.2015.03.076.
[4] Thomassen YE, Van't Oever AG, Vinke M, Spiekstra A, Wijffels RH, van der Pol LA, et al. Scale-down of the inactivated polio vaccine production process. Biotechnol Bioeng 2013;110:1354–65. https://doi.org/10.1002/bit.24798.
[5] Bakker WAM, Thomassen YE, Van't Oever AG, Westdijk J, van Oijen MGCT, Sundermann LC, et al. Inactivated polio vaccine development for technology transfer using attenuated Sabin poliovirus strains to shift from Salk-IPV to Sabin-IPV. Vaccine 2011;29:7188–96. https://doi.org/10.1016/j.vaccine.2011.05.079.
[6] Rathore AS, Kumar D, Kateja N. Role of raw materials in biopharmaceutical manufacturing: risk analysis and fingerprinting. Curr Opin Biotechnol 2018;53:99–105. https://doi.org/10.1016/j.copbio.2017.12.022.
[7] Borchert D, Suarez-Zuluaga DA, Sagmeister P, Thomassen YE, Herwig C. Comparison of data science workflows for root cause analysis of bioprocesses. Bioprocess Biosyst Eng 2018. https://doi.org/10.1007/s00449-018-2029-6.
[8] Charaniya S, Hu W-S, Karypis G. Mining bioprocess data: opportunities and challenges. Trends Biotechnol 2008;26:690–9. https://doi.org/10.1016/j.tibtech.2008.09.003.
[9] Kirdar AO, Green KD, Rathore AS. Application of Multivariate Data Analysis for Identification and Successful Resolution of a Root Cause for a Bioprocessing Application. Biotechnol Prog 2008;24:720–6. https://doi.org/10.1021/bp0704384.
[10] Sagmeister P, Wechselberger P, Herwig C. Information Processing: Rate-Based Investigation of Cell Physiological Changes along Design Space Development. PDA J Pharm Sci Technol 2012;66:526–41. https://doi.org/10.5731/pdajpst.2012.00889.
[11] Golabgir A, Gutierrez JM, Hefzi H, Li S, Palsson BO, Herwig C, et al. Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. Biotechnol Adv 2016;34:621–33. https://doi.org/10.1016/j.biotechadv.2016.02.011.
[12] Thomassen YE, Van't Oever AG, van Oijen MGCT, Wijffels RH, van der Pol LA, Bakker WAM. Next generation inactivated polio vaccine manufacturing to support post polio-eradication biosafety goals. PLoS ONE 2013;8:. https://doi.org/10.1371/journal.pone.0083374e83374.
[13] Knowlson S, Burlison J, Giles E, Fox H, Macadam AJ, Minor PD. New strains intended for the production of inactivated polio vaccine at low-containment after eradication. PLoS Pathog 2015;11:. https://doi.org/10.1371/journal.ppat.1005316e1005316.
[14] Thomassen YE, Rubingh O, Wijffels RH, van der Pol LA, Bakker WAM. Improved poliovirus d-antigen yields by application of different Vero cell cultivation methods. Vaccine 2014;32:2782–8. https://doi.org/10.1016/j.vaccine.2014.02.022.
[15] Sokolov M, Morbidelli M, Butté A, Souquet J, Broly H. Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target. Biotechnol J 2018;13:1700461. https://doi.org/10.1002/biot.201700461.
[16] Kozma B, Salgó A, Gergely S. Comparison of multivariate data analysis techniques to improve glucose concentration prediction in mammalian cell cultivations by Raman spectroscopy. J Pharm Biomed Anal 2018;158:269–79. https://doi.org/10.1016/j.jpba.2018.06.005.
[17] Sawatzki A, Hans S, Narayanan H, Haby B, Krausch N, Sokolov M, et al. Accelerated bioprocess development of endopolygalacturonase-production with saccharomyces cerevisiae using multivariate prediction in a 48 mini-bioreactor automated platform. Bioengineering 2018;5:101. https://doi.org/10.3390/bioengineering5040101.
[18] ten Have R, Thomassen YE, Hamzink MRJ, Bakker WAM, Nijst OEM, Kersten G, et al. Development of a fast ELISA for quantifying polio D-antigen in in-process samples. Biologicals 2012;40:84–7. https://doi.org/10.1016/j.biologicals.2011.11.004.
[19] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab Syst 2001;58:109–30.
[20] Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR models - strategies and importance. Int J Drug Des Discov 2011;2:511–9.
[21] Petiot E, Guedon E, Blanchard F, Gény C, Pinton H, Marc A. Kinetic characterization of vero cell metabolism in a serum-free batch culture process. Biotechnol Bioeng 2010;107:143–53. https://doi.org/10.1002/bit.22783.

[22] Quesney S, Marc A, Gerdil C, Gimenez C, Marvel J, Richard Y, et al. Kinetics and metabolic specificities of Vero cells in bioreactor cultures with serum-free medium. Cytotechnology 2003;42:1–11. https://doi.org/10.1023/A:1026185615650.
[23] Nahapetian AT, Thomas JN, Thilly WG. Optimization of environment for high density Vero cell culture: effect of dissolved oxygen and nutrient supply on cell growth and changes in metabolites. J Cell Sci 1986;81:65–103.
[24] Huang H, Yi X, Zhang Y. Improvement of Vero cell growth in glutamate-based culture by supplementing ammoniagenic compounds. Process Biochem 2006;41:2386–92. https://doi.org/10.1016/j.procbio.2006.06.018.
[25] Cruz HJ, Ferreira AS, Freitas CM, Moreira JL, Carrondo MJT. Metabolic responses to different glucose and glutamine levels in baby hamster kidney cell culture. Appl Microbiol Biotechnol 1999;51:579–85. https://doi.org/10.1007/s002530051435.
[26] Power JF, Reid S, Radford KM, Greenfield PF, Nielsen LK. Modeling and optimization of the baculovirus expression vector system in batch suspension culture. Biotechnol Bioeng 1994;44:710–9. https://doi.org/10.1002/bit.260440607.
[27] Wong KT, Peter CH, Greenfield PF, Reid S, Nielsen LK. Low multiplicity infection of insect cells with a recombinant baculovirus: The cell yield concept. Biotechnol Bioeng 1996;49:659–66. https://doi.org/10.1002/(SICI)1097-0290(19960320)49:6<659::AID-BIT7>3.0.CO;2-N.
[28] Friedman J, Hastie T, Tibshirani R. High-Dimensional Problems: p >> N. The elements of statistical learning. Berlin: Springer series in statistics Springer; 2017. p. 649–99.
[29] Multicollinearity Least. Absolute shrinkage and selection operator, elastic net, ridge, adaptive lasso, fused lasso. Int J Statist Appl 2015:6.
[30] Kumamaru H, Schneeweiss S, Glynn RJ, Setoguchi S, Gagne JJ. Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. Emerg Themes Epidemiol 2016;13. https://doi.org/10.1186/s12982-016-0047-x.
[31] Friedman J, Hastie T, Tibshirani R. Shrinkage Methods. The elements of statistical learning. Berlin: Springer series in statistics Springer; 2017. p. 61–73.