Dissertation

Data Science Methods to Decrease Experimental Efforts in Quality by Design Tasks

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Prof. Dr. Christoph Herwig

E166 Institut für Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften

eingereicht an der Technischen Universität Wien
Fakultät für Technische Chemie

von

Dipl.-Ing. Daniel Borchert, BSc.
Matrikelnummer: 01040416
Flurgasse 14, 7053 Hornstein

_____ _____

Wien, am eigenhändige Unterschrift

Daniel Borchert

Data Science Methods to Decrease Experimental Efforts in Quality by Design Tasks

Dissertation; September, 15th 2020
Reviewers: Prof. Dr. Norbert Kockmann and Prof. Dr. Peter Filzmoser
Supervisor: Prof. Dr. Christoph Herwig

TU Wien
Institute of Chemical, Environmental and Bioscience Engineering
Faculty of Technical Chemistry
Karlsplatz 13
1040 Vienna
Austria

Acknowledgment

First, I would like to thank Prof. Christoph Herwig for his advice and support during the last years. His motivation and his way of guiding me, not just on letting me develop my ideas but also understanding my ideas and his support in scientific writing and presenting, were essential for making this thesis possible. Without your support, this thesis and my personal development would not have been possible in that way.

Furthermore, I want to thank all the persons who worked with me to my scientific contribution. From the Exputec team I want therefore, thank to Patrick Sagmeister, Valentin Steinwandter and Thomas Zahel. From the Intravacc team I want therefore, special thanks to Diego Suarez for his support during the last years and to Yvonne Thomassen for input when it was needed.

A big thanks also to the current and former Exputec team and to Werum IT Solutions. Particularly to Christopher Taylor, Lukas Marschall, Barbara Pretzner, Petra Lubitz, Gergö Szita and Ulrich Tröller as well as the entire software development team who supported my development at Exputec and who made never thinks boring.

Personally, I want to thank my wife Verena for giving me the time on the weekends and at the nights to make this thesis happen. Also, I want to thank my entire family for their support.

Finally thank you to my son Jonathan, who always motivated me with his good mood and power over the day. This thesis is dedicated to you and thank you for being part of my life.

Content

Abstract

Digitization in the pharmaceutical industry, Pharma 4.0, is becoming more important every year. The first concepts in this area are already being implemented by several large companies and will be completed in the next few years. Such concepts should make it more comfortable in the future to collect data, exchange knowledge, and conduct holistic data analyses.

The expected benefits of such systems are highly appreciated and, according to the current state of the industry, also needed, especially in the Quality by Design (QbD), which is applied to guarantee consistent product quality. QbD is used to understand the source of the variance of particular quality attributes as well as to understand the linkages of its interaction in order to ensure robust product quality. However, it is difficult to apply the widely used and often applied QbD approach correctly because, from a current perspective, there is a lack of basic QbD interpretation and sufficient statistical knowledge of the process experts. Therefore, more and more experiments will often be conducted within the QbD study in order to get the wanted process understanding. However, mechanistic process modeling and data analysis approach should be used to get to this aim.

The goal of that thesis is to go back to the roots, such as using process knowledge from models and statistical analysis, by investing more time in data analysis rather than in experimental effort by applying existing data science methods useful within the QbD approach. Furthermore, we aim to show that a combination of existing methods and correct result interpretation lead again to the QbD benefits as initially anticipated.

As particular achievements of this thesis, it is demonstrated how a simple combination of existing root cause analysis approaches is used to improve the cause and effect analysis to gather additional and more profound process knowledge using historical data. This method provides new insights into the process by using all available process data within data analysis and useful result interpretation. Since robust product quality is one of the most relevant results of the QbD approach, risk assessment and interpretation is the core element of that concept. Improvements of the current risk assessment and interpretation approaches of the individual process parameters are demonstrated, to finally get a quantitative risk evaluation for each process parameter individually. In the end, a comprehensive case study is presented to demonstrate the implementation and benefits of the improved QbD approach. A holistic design space evaluation at the end of the QbD study is shown and lays the basis for a holistic process control strategy.

The overall goal to reduce the number of experiments by investing more time in useful data analysis is shown along with the first implementation example of the newly developed approach. It is shown that appropriate interactions into the current QbD approach have a significant influence on the experimental effort and a high potential to reduce the time to market and overall process costs, to finally reduce the patient risk since drugs can be faster available.

1. Introduction

1.1. Motivation and Why now?

Within the biopharmaceutical industry, the buzzwords Digitalization, Industry 4.0, and the Internet of Things (IoT) became more and more important within the past years. Many big players within that field, GSK [1], Rentschler Biopharma [2], are currently launching their first holistic data storage clouds, also called data lakes. This industrial revolution within the biopharmaceutical industry is also called Pharma 4.0. This technological progress enables access to all available data generated within the biopharmaceutical process, reduces the effort in data gathering, and can be used for comprehensive data analysis and process assessment. These infinite data access capabilities can be used to detect prior hidden information since it should be possible to include all available data into data analysis. Theoretically, that is feasible with such systems but practically not possible with the current data analysis methods and existing data science environment.

A typical process within the biopharmaceutical or chemical industry consists of several consecutive linked steps, the unit operations [3]. Each phase has a particular purpose and aim. While the first step, often called fermentation, aimed to produce the product of interest in a specific amount, the following steps, separate, isolate and further capture the product of interest from the supernatant. Finally, the purification of the product will be conducted to get to the desired quality. Thus such processes generate different amounts and types of data. Table 1 classifies them exemplarily.

Table 1: Overview of available data within biopharmaceutical process. *The rows indicate the different data types and the columns represent their characteristics.*

Type	Data Source	Property	Example	Purpose
Online data	Time or volume dependent	High frequent generated data	Sensor data, CPPs	Process control
Offline data	Time or volume dependent	Low frequent generated data	Bioprofile measurements like metabolites, CPPs, KPIs	Key performance indication. Sample measurements. Process control
Feature	One point measurement	Time independent	Product titer or starting volume, KPIs and CQAs	Process quality attributes
Image data	Image	Static	Photo of a SDS result, CQAs	Raw material tracking and for quantification
Meta data	Categorical	Time independent	Media components	Holistic process assessment and raw material tracking

Spectral data	3D data	Multidimensional visualization	Raman data	Combination of dimension for comprehensive process evaluation

Since all information could be potentially relevant for the process in terms of performance and product quality, all of them should be considered within a comprehensive data analysis study. The data analysis within the biopharmaceutical industry often focused on a particular unit operation. Therefore, a selective view of the data is often conducted and may cause that hidden interactions cannot be detected, in the holistic process context. The Pharma 4.0 thinking in combination with data science tools, recently discussed by Steinwandter et al. [4], elaborates on the current and needed data gathering and data analysis strategies within the biopharmaceutical industry. Within this review, the current applied data science methods in the bioprocess life cycle are discussed in more detail. Four categories are presented to be the main parts of the life cycle: process development and scale-up, process validation and characterization, routine manufacturing, and life cycling. In each of these categories, different mathematical and data science methods are required and applied. It will be compared which data sources are available and which results are expected within each step. There are the next hurdles which have to be overcome by the data scientists:

- Complex processes with splitting and pooling events
- Risk reduction
- Too few experiments to be able to apply statistical methods correctly
- Different process steps with different data formats

It can be summarized that due to the complexity of a bioprocess, a data scientist has to master many skills, and additionally, process understanding is necessary to evaluate a bioprocess holistically. Many systems on the market already allow the collection and further processing of process data, so the first systems are already available. However, they must still be brought together, and all data must be accessible in a holistic way [4].

Furthermore, Steinwandter et al. showed that based on the current applied systems in this industry, it is challenging to implement a Data Science platform. This challenge is because, that some open issues such as deployment strategy, code testing, and the implementation of virtual environments are still in the early stages and are not yet really applied in the biopharmaceutical industry [4]. Although there are still some hurdles in this industry sector before data science platforms can be used holistically, and their potential fully exploited, the first steps have already been taken and the potential has been recognized for a realization of these systems shortly.

This new mindset also initiates the first steps in the direction of seeing the big process picture and the holistic view of the process since it will be possible to consider all available data. Since unit operation are often depending on each other [3], a unit operation vise assessment is probably

not the best analysis method. The holistic process assessment is, therefore, recommended in future process assessment and should be conducted. This holistic process evaluation is especially necessary for process characterization. Here, the process should be comprehensively considered and evaluated from beginning to end. Process knowledge and risk assessment play a decisive role in this process, whereby we often talk about Quality by Design.

1.2. Quality by Design

Quality by Design (QbD) is defined by the international conference of harmonization (ICH) guideline Q8 as *a systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and process control, based on sound science and quality risk management* [5] mainly applied to achieve a consistent product quality [6]. In this context, process understanding, in combination with risk assessment, can be considered as the most crucial factor in implementing a QbD approach successfully [6]. For any data analysis process, knowledge is indispensable [7, 8], and for further process assessment, the individual risks of data and parameters need to evaluated [9]. As this method has been used in the biopharmaceutical industry for years, several companies have conducted a joint case study, the A-Mab case study from the CMC working group, to comprehensively define the main components and concept of a QbD approach [10]. The therein presented approach is further discussed in the next section.

1.3. Quality by Design approach

The systematic approach presented within the A-Mab case study from the CMC working group [10] is summarized within Figure 1.

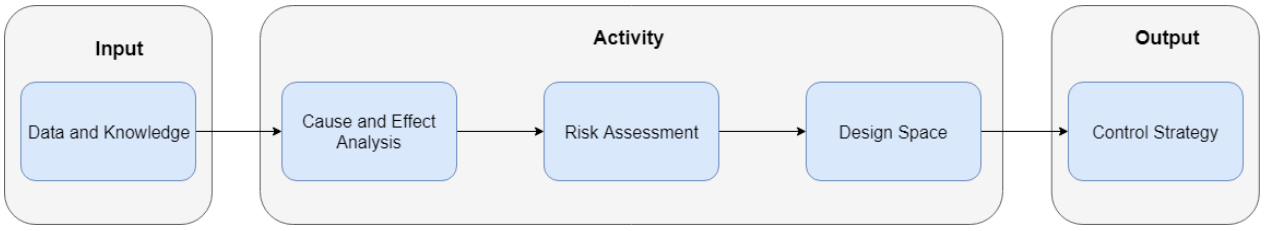

Figure 1:Quality by design approach interpretation based on the report of the A-Mab case study by the CMC working group. *Each main element of the QbD approach is shown as a grey box. The associated sub elements are therefore displayed in light blue boxes.*

Three main elements of this approach can be identified:

- Input
- Activity
- Output

All of these elements will be considered individually within this section.

1.3.1. Quality by Design Approach – Inputs

In the beginning, it is necessary to collect data. Ideally, all data should be collected and made available for evaluation. As already listed in Table 1, there are significant differences in the type of data. Steinwandter et al. [4] recently evaluated and summarized the application of process

data within the pharmaceutical industry and its importance. In this article, it can be seen that different data can be generated depending on the process step and that different analytical methods are used based on the available data. Within the process characterization and the QbD approach, mainly time-resolved data, categorical information, and one point measurements are used. Already based on those data sources, it is possible to select the correct procedures for each data type to evaluate the process with the current data science methods. Therefore, a first evaluation of the data at the beginning is of vital interest since needed methods can already be selected or implemented.

1.3.2. Quality by Design Approach – Activities

After all data are available, or it is known which data are needed, and the necessary analysis procedures are known, the data analysis can start. Figure 1 shows that the first step in the QbD data analysis approach is the Cause and Effect Analysis. This analysis procedure aims to identify the cause of a fluctuation of individual process relevant quality attributes [11]. The influencing factors identified in this way can be used to draw conclusions about the different relationships within the process and show how parameters can influence each other. The results generated here serve the process experts a more profound process knowledge and enable a better evaluation of potential interrelationships in the future.

In the next step, the risk within the process is evaluated based on the current process knowledge. This step is called Risk Assessment in Figure 1. The risk assessment is usually carried out for each previously defined critical quality attribute. The influence of each process parameter is assessed individually and usually within the normal operating ranges. Also, three factors, Severity, Occurrence, and Detectability, are evaluated individually to assess the parameter criticality [12].

Finally, the design space of the process will be evaluated, compare Figure 1, which describes the ratios of inputs on the CQAs. This evaluation can be a sophisticated mathematical operation or a combination of individual factors [5]. To investigate these relationships and influences, experiments, so-called Design of Experiments, are often carried out. The thereby resulted findings should provide a more profound holistic process understanding since interaction, and quadratic effects of the process parameter can be assessed from the conducted experiments.

1.3.3. Quality by Design Approach – Output

The output of the approach is the control strategy, as shown in Figure 1. The Control Strategy can be interpreted as the accepted range of variation of each process parameter within the entire process always to deliver a similar result [13].

1.4. Quality by Design Problem Statement

This approach is not new in itself, QbD has been lived and applied in the pharmaceutical industry for years, and with good reason [6]. The basic idea is that the QbD approach achieves its dedicated aim, compare section 1.2, by applying mathematical models, transferring process knowledge [8], and using data-driven process knowledge [14]. This interdisciplinary process evaluation leads, therefore, to [15]:

- profound process knowledge
- more targeted process optimization
- safe production of the product

Due to its widespread use and almost daily application, the routine has developed in the usage of this approach. It has been observed that the initial idea of the QbD approach is becoming more and more distant. More often, it is necessary to do more and more experiments to generate the necessary process knowledge, and this is, therefore, no longer an adequate interpretation of the QbD approach. This approach aims to design processes with the help of models and simulations, as well as the use of multivariate data analysis to gather process knowledge [14, 15]. One additional interpretation mistake within the current industry application is that each unit operation will be considered separately. However, we know that all the unit operations within the process chain affect each other [16].

Furthermore, it can be observed that the trend in data analysis goes more and more to sophisticated applications like neural networks and other machine learning algorithms in order to recycle historical process knowledge holistically and to consider all available data [17, 18]. Such a black-box model makes it additionally difficult for the process expert in applying and understanding such an application as well as to interpret the results correctly. Since it is often the case that process experts are not just trained on simple data analysis tools, such complex data analysis approaches can lead to a situation where even simple analytical approaches are not applied.

This inappropriate application ultimately leads to the fact that the expected added value of the QbD approach cannot be fulfilled and the following reasons must be given [15]:

1. Difficulty in the evaluation and analysis of large amounts of data.
2. There is a unit operation wise view on the process
3. Difficulty in creating a holistic process control strategy
4. Improper use and reuse of knowledge from other units
5. Lack of knowledge in the application and interpretation of multivariate data analysis

1.5. Thesis Goal: Focus More on Existing Data Analysis Than on Experiments Effort

The goal of this thesis is to ensure that data science methods are applied in a useful and sustainable way to reduce experimental effort. Especially in the QbD approach, these applications are more in demand than ever before. With this thesis, it is shown that the correct application of data science methods, their combination, and critical result interpretation creates an added value for process experts by revealing prior hidden process knowledge.

Furthermore, it is shown that even simple statistical methods, known by every process expert, are sufficient to derive an appropriate benefit from the analysis. Thus, the process knowledge can be deepened, and a holistic process evaluated can happen.

The herein presented manuscripts interact on different steps in the QbD approach, especially in the sections Cause and Effect Analysis and Risk Assessment. The last manuscript presents a

comprehensive case study to evaluate the impact of the connection and the limitations. Figure 2 depicts the covered areas of the herein presented manuscripts indicated by an arrow sign. The different color displays the type of the manuscript; a blue arrow represents a methodological evaluation of that topic, while orange arrows show the result of a case study, for this topic. For the present result of the case study, the beforehand developed methodology was applied.

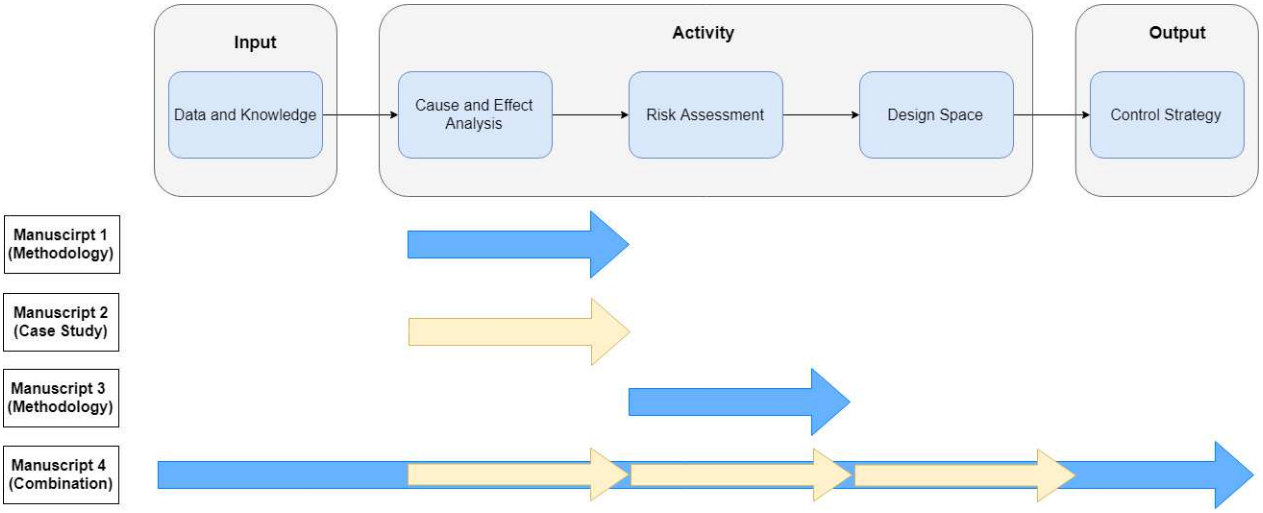

Figure 2: Overview of the topics covered by the individual manuscripts from the QbD approach. *The manuscripts discussed in this thesis are shown in each row. The columns represent the different tasks within the QbD approach. Each arrow shows the QbD tasks covered by the manuscript; a blue arrow represents a methodological evaluation of that topic, while orange arrows show the result of a case study for this topic.*

The first two manuscripts descript a new cause and effect analysis approach. The basic concept of the newly developed approach is explained in Manuscript 1. We present that the reasonable combination of the two commonly used tools for cause and effect analysis, the raw data analysis, and the feature-based approach, results in a comprehensive analysis workflow. It was additionally shown that sound data unfolding at a particular step of the analysis, in combination with suitable multivariate data analysis, returns all hidden information within the process data [19]. The Manuscript 2 demonstrates the result of the case study using upstream data from a virus production process. The entire conducted analysis of that case study was based on the methodology developed in Manuscript 1. The therein investigated process consists of one cell culture and one virus production phase. It was finally possible to identify that the cell culture phase of the investigated process remains as crucial. This observation has not been reported so far because it was assumed that this phase has no significant influence on the process. This result shows that a correct combination of simple procedures and their interpretation generates more process knowledge than before [20].

The next step in the QbD approach (Figure 2) highlights the evaluation of the Risk Assessment. A novel evaluation approach was developed and reported in Manuscript 3. The method developed there, clearly shows that the previously popular approach, the RPN approach, is no longer necessary and that a quantitative assessment of the risk of individual process parameters can be determined. The basis for this assessment is the translation of the factors Severity and Occurrence, obtained from the risk assessment, to model effect size and parameter distribution.

This translation further enables a linearization approach to ascertain the influences of the individual process parameters [21].

Lastly, Manuscript 4 shows the possible combination of the individual approaches from Manuscript 1 and 3, applied in the QbD approach. A case study based on real data shows the feasibility of the holistic concept. It was shown that a significant reduction of potentially critical process parameters can be expected and that the design space evaluation can be accelerated [22].

1.6. Quality by Design Tasks Covered Within This Thesis

This thesis can be seen as a guide, presenting how easy it is to gain more knowledge from existing data by using existing data science methods in combination with simple data analysis approaches.

The following table shows an overview of the manuscript presented in this thesis and their contribution to the individual elements within the QbD approach.

***Table 2: Overview of the individual manuscripts and their influence on the QbD approach**. The rows indicate four manuscripts involved in this thesis abbreviated with a short title, and the corresponding citation number is in brackets. Columns indicate the QbD tasks, and a blue field represented the covered task by the different manuscripts.*

	Holistic data gathering	Holistic process assessment	Individual process understanding	Multivariate data analysis and interpretation	Hypothesis creation and testing	Risk assessment evaluation	Quantitative risk evaluation
Cause and effect analysis methodology [19]	■		■	■	■		
Case study cause and effect analysis [20]	■		■	■	■		
Quantitative CPP assessment methodology [21]		■				■	■
Case study improved QbD approach [22]	■	■	■	■	■	■	■

2. Results

2.1. Authors Contribution

Table 3: Authors Contribution overview

Manuscript title	Authors	Contribution of Daniel Borchert (DBO)	Ref
Comparison of data science workflows for root cause analysis of bioprocesses	**Borchert**, Suarez-Zuluaga, Sagmeister, Thomassen, Herwig	DBO designed the methodology and did the analysis. DBO wrote the manuscript.	[19]
Accelerating bioprocess development by analysis of all available data: A USP case study	Suarez-Zuluaga*, **Borchert***, Driesen, Bakker, Thomassen	DBO did the raw data analysis and supported the feature-based analysis. DBO wrote the Introduction, Material and Methods, parts of the Result and the Conclusion section of the manuscript.	[20]
Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling	**Borchert**, Zahel, Thomassen, Herwig, Suarez-Zuluaga	DBO tuned the idea to transfer the FMEA to IPM. DBO developed the algorithm, set up and simulated the in-silico study and did the simulation on real data. DBO wrote the manuscript.	[21]
Risk Assessment and Integrated Process modeling – An Improved QbD approach for the Development of the Bioprocess Control Strategy	**Borchert**, Suarez-Zuluaga, Thomassen, Herwig	DBO developed the improved QbD approach, did the case study and wrote the manuscript	[22]
Data science tools and applications on the way to Pharma 4.0	Steinwandter, **Borchert**, Herwig	DBO wrote parts of the introduction and challenges. DBO discussed the current status within process developed and supported the discussion and conclusion sections	[4]

* Both authors contributed equally to this work.

2.2. Manuscript 1:
Comparison of Data Science Workflows for Root Cause Analysis of Bioprocesses

This manuscript has been published in *Bioprocess and Biosystems Engineering*.

Full citation

Borchert, D.; Suarez-Zuluaga, D.A.; Sagmeister, P.; Thomassen, Y.E.; Herwig, C. Comparison of data science workflows for root cause analysis of bioprocesses. *Bioprocess and Biosystems Engineering* **2019**, *42*, 245–256.
https://doi.org/10.1007/s00449-018-2029-6

2.3. Manuscript 2:

Accelerating Bioprocess Development by Analysis of All Available Data: A USP Case Study

This manuscript has been published in *Vaccine*, Special Issue: *Vaccine Technology VII*

Full citation:

Suarez-Zuluaga, D.A.; Borchert, D.; Driessen, N.N.; Bakker, W.A.M.; Thomassen, Y.E. Accelerating bioprocess development by analysis of all available data: A USP case study. *Vaccine* **2019**, *37*, 7081–7089.
https://doi.org/10.1016/j.vaccine.2019.07.026

2.4. Manuscript 3:

Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling

This manuscript has been published in *Bioengineering*.

Full citation:

Borchert, D.; Zahel, T.; Thomassen, Y.E.; Herwig, C.; Suarez-Zuluaga, D.A. Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling. *Bioengineering* **2019**, *6*, 114.
https://doi.org/10.3390/bioengineering6040114

2.6 Manuscript 5:

Data Science Tools and Applications on the Way to Pharma 4.0

This manuscript has been published in *Drug Discovery Today*

Full citation:

Steinwandter, V.; Borchert, D.; Herwig, C. Data science tools and applications on the way to Pharma 4.0. *Drug Discovery Today* **2019**, 24, https://doi.org/10.1016/j.drudis.2019.06.005

3. Summary and Conclusion

Within this thesis, an improved QbD approach was presented and shown by using real data. The main objective was to prove the hypothesis that the investment of more time in data analysis than before and the use of the newly generated knowledge to assess the process holistically can reduce experimental effort. Accordingly, the manuscripts included in this thesis, its contribution to the QbD approach, their achievements, deliverables, and solved problems, as described in section 1.4, are summarized in Table 4**.**

Table 4: Summary of the manuscript's contribution to QbD approach. *The rows indicate the four manuscripts involved in the QbD improvement, the name is an abbreviation of the manuscript title and the citation number is shown in brackets. The columns indicate the individual contribution, achievements, deliverables as well as the solved problem of the current QbD approach in more detail as shown in Table 2.*

	Contribution to QbD approach, from Table 2	**Achievements**	**Deliverable**	**Problem reference, section 1.4**
Cause and effect analysis methodology [19]	- Multivariate data analysis to increase and/or gather process knowledge - Identify CQA impacts - Hypotheses generation and testing	- Comprehensive data evaluation of all available data - Reasonable data unfolding to identify critical time series data and to prepare the data set for data analysis	- A simple workflow for cause and effect analysis including all available data - Combination of primary statistic approaches interpretable by any process expert	1, 5
Case study cause and effect analysis [20]	- Presentation of the simplicity of application in the approach	- User manual to analyze all existing data - Identification of additional crucial process phases	- Proof of concept of the cause and effect analysis methodology	1, 5

Quantitative CPP assessment methodology [21]	- Risk assessment interpretation - Factor evaluation for experimental planning	- Linearization of Severity and Occurrence to model effect size and parameter distribution	- Risk assessment evaluation tool to quantitative evaluate parameters risks	2, 4
Case study improved the QbD approach [22]	- Multivariate data analysis - Hypotheses generation and testing - Risk evaluation - Design space evaluation	- Connection of data science approaches to decrease drug time to market - Concept proof - Holistic process evaluation	- A holistic process assessment tool - Method proofs - Deeper process understanding tool	1,2,3,4,5

According, Table 2 and Table 4, this thesis can be seen as a methodological case study, where the integration of data science methods in combination with simple data analysis approaches into the QbD approach was investigated.

3.1. Achievements

In this thesis, a methodological workflow is presented, which shows how the state of the art data science methods and its combinations can be used to accelerate QbD. It was shown that by investing more time in thoughtful data analysis and that quantitative risk assessment has the potential to reduce experimental efforts holistically. The herein presented achievements and their impact on each are shown in Figure 3.

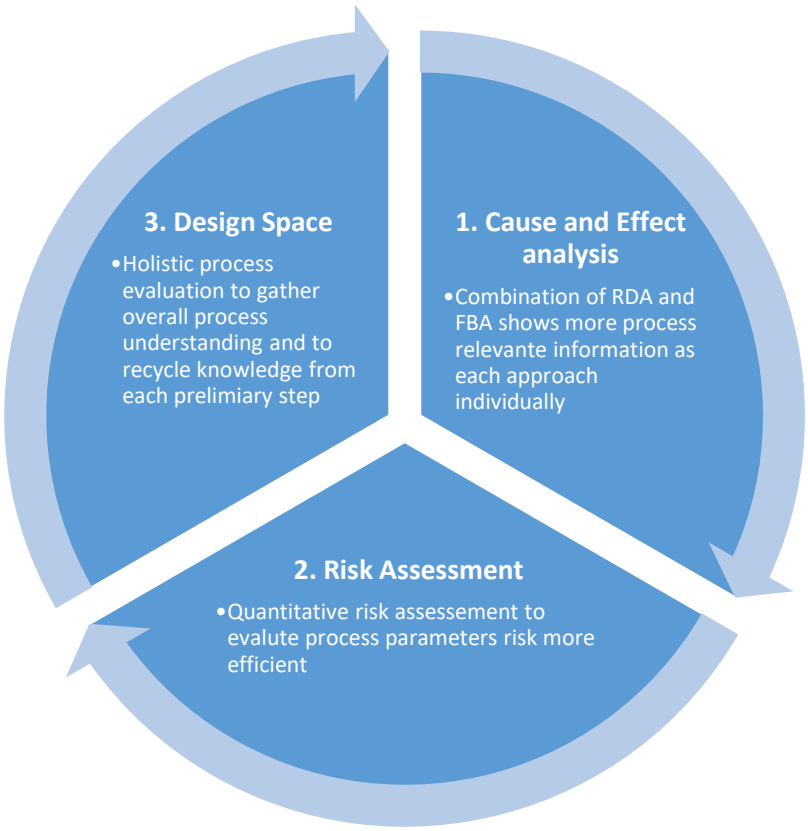

Figure 3: Summary of the achievement gather from the different manuscripts presented here.

In order to achieve the overall aim, focusing more on data science methods and result interpretation, the three essential steps presented in Figure 3 are necessary. Within this section, these steps and its achievements are summarized:

1. First, a comprehensive cause and effect analysis is required to assess all gather process data and to identify the impacts of particular time information on the critical quality attributes. It is shown in [19] and [20] that the combination of the two most popular methods in cause assessment of the RDA and the FBA leads to a comprehensive assessment tool with which all available process data can be analyzed. This new procedure, therefore, makes it possible to identify suspect time series first and from which the relevant information can be obtained, with targeted unfolding procedures. Subsequent multivariate hypotheses testing finally result in cause identification.
2. Risk Assessment interpretation by using a novel risk assessment approach for quantitative risk assessment of all process parameters is described in [21]. This quantitative evaluation is essential to estimate the exact influence of single process parameters and to evaluate their behavior in the whole process. The current approach to assess RA, the RPN approach, can be avoided with the herein method described. In particular, the interpretations from the process experts are linearized and can thus make a statement about the effect of individual process

parameters on the CQA. This evaluation leads to a more precise adaptation of the subsequent experiments and therefore has a great potential to reduce the number of necessary experiments or increase the power of the model.

3. Design space evaluation presented in [22] shows the benefits of combining the individual data science methods applied in the QbD approach. First, it starts with an investigation of the significant influences on the CQAs within each unit operation independently to increase specific process knowledge. This new awareness can be used to create an RA and to estimate the effect sizes of the individual process parameters and the distribution within the operating ranges of the parameter. The subsequent simulation for the quantitative parameter assessment opens up the possibility of a holistic process assessment. The final result of the simulation allows us to evaluate the design space of the process better and to conduct targeted experiments to obtain the optimal settings for a holistic process control strategy.

3.2. Impact and Deliverables

The increasing importance of being able to evaluate the process holistically is a significant issue, especially in process characterization. In order to understand the process, many experiments are often carried out. This experimental effort costs time and money and often has no crucial benefit for the expert. To further increase the benefit, often specific experiments will further conduct to understand the process even better.

This cycle leads into an endless loop, and always further experiments to understand the process better and better are needed. In this thesis, it is shown that not always the experimental execution leads to an improvement of the knowledge about the process. Already the right combination of existing data analysis methods and a useful combination of these leads to a significant added value for the expert since additional process knowledge can be easily generated. Therefore, there is a need to invest more time for data analysis and extract all possible knowledge from all different data sources. Here, the number of experiments can already be reduced significantly by performing useful data analysis correctly, and effects on critical process parameters can be evaluated faster. This additional generated knowledge helps the process expert to evaluate the subsequent RA better and to incorporate appropriate effect sizes of the individual parameters at the beginning of the risk assessment.

Risk assessment has the purpose of evaluating critical process parameters in the process. This criticality assessment is currently based on the RPN. Three factors are multiplied by each other, and the resulting value is used to assess whether a process parameter is critical or not. There is no quantitative statement of the criticality with this value. Therefore, it is not possible to estimate the influence of the process parameter on the process at this stage of the analysis. Here a method is presented where it is feasible to translate the process knowledge into quantitative risk evaluation based on simple mathematical modeling. The linearization method developed for this purpose enables it to estimate the effect sizes of the individual process parameters holistically and to estimate the factor's impact on each quality attribute. This approach allows us to plan the following experiments more sustainable to get the required information effectively.

Finally, it is crucial for a design space evaluation to choose the optimal process conditions to be able to evaluate it. With an evaluation of all available process data and a comprehensive process simulation, this is possible and described here. The way from a unit operation process evaluation to a holistic procedure is therefore possible. With this procedure, the number of experiments can be effectively reduced and thus also the costs and time for process characterization. Thus, a product can be brought to market faster because the effort for experiments can be optimized explicitly with the help of Data Science methods.

4. Outlook

This thesis covers just a subset in the field of including data science in the biopharmaceutical industry by considering existing data science methods and its application in QbD. This work shows how simple data analysis methods can add significant value to the process expert. While machine learning algorithms, artificial intelligence, or Bayesian statistical methods may result in other data mining suggestions and can also improve the process knowledge. Such techniques and the applied algorithms are often hard to understand by current process experts within the biotechnological industry, and therefore also considered as black box model. Through the herein used data evaluation methods can create self-learning applications based on historical data to suggest appropriate process settings for the future. All of these sophisticated mathematical techniques are not covered here, since it was an aim, to develop an easy to understand improvement workflow considering existing techniques within the QbD approach to decrease the experimental effort by investing more time in data analysis.

Furthermore, another commonly used technique within the biopharmaceutical and chemical industry was significantly improved by existing data science methods, the RA, as shown in this thesis. The risk evaluation approach is also well developed and often used in the industry, and therefore, a significant application area of the improved approac. Linearization of Severity and Occurrence in combination with holistic process simulation is used for this purpose. Since the herein assumed linear relation based on the Severity estimation allows that mainly main effects can be simulated, it will be effortless to adjust the current assumption with interaction and quadratic effects of the model parameter. This updated approach could be a potential research topic in the future when the integrated process modeling technique is investigated further. Additionally, a potential in the development of a hybrid integrated process model is expected to be developed shortly. This hybrid solution could potentially combine risk assessment information with real measurements in order to adjust the risk assessment or to model unit operations were no data are measured.

Although the herein presented techniques are mainly developed to decrease the drug time to marked, it needs to be stated that the herein presented methods can use in any other industrial sector, where the root causes analysis approaches, risk assessment techniques, and experimental design planning is applied.

5. References

1. Macaulay T GlaxoSmithKline Chief Data Officer Mark Ramsey using data to transform drug discovery. In: CIO UK. https://www.cio.co.uk/cio-interviews/gsk-cdo-mark-ramsey-explains-how-data-is-transforming-drug-discovery-3673555/. Accessed 22 Oct 2019

2. Aktuelles - Pressemeldungen - Detail - Rentschler Biopharma. https://www.rentschler-biopharma.com/de/aktuelles/pressemeldungen/detail/view/artikel-in-plattform-life-sciences-zum-schwerpunktthema-smarte-medizin-veroeffentlicht/. Accessed 22 Oct 2019

3. Doran PM (2013). In: Bioprocess engineering principles, 2nd ed. Elsevier/Academic Press, Amsterdam ; Boston, pp 218–253

4. Steinwandter V, Borchert D, Herwig C (2019) Data science tools and applications on the way to Pharma 4.0. Drug Discovery Today. https://doi.org/10.1016/j.drudis.2019.06.005

5. ICH Q8 (R2) (2009) Pharmaceutical Development Q8 (R2)

6. Rathore AS (2009) Roadmap for implementation of quality by design (QbD) for biotechnology products. Trends Biotechnol 27:546–553. https://doi.org/10.1016/j.tibtech.2009.06.006

7. Charaniya S, Hu W-S, Karypis G (2008) Mining bioprocess data: opportunities and challenges. Trends in Biotechnology 26:690–699. https://doi.org/10.1016/j.tibtech.2008.09.003

8. Herwig C, Garcia-Aponte OF, Golabgir A, Rathore AS (2015) Knowledge management in the QbD paradigm: manufacturing of biotech therapeutics. Trends in Biotechnology 33:381–387. https://doi.org/10.1016/j.tibtech.2015.04.004

9. Rathore AS, Kumar D, Kateja N (2018) Role of raw materials in biopharmaceutical manufacturing: risk analysis and fingerprinting. Current Opinion in Biotechnology 53:99–105. https://doi.org/10.1016/j.copbio.2017.12.022

10. CMC Biotech Working Group (2009) A-Mab: a Case Study in Bioprocess Development. 278

11. Rathore AS, Mittal S, Pathak M, Arora A (2014) Guidance for performing multivariate data analysis of bioprocessing data: Pitfalls and recommendations. Biotechnology Progress 30:967–973. https://doi.org/10.1002/btpr.1922

12. Guideline IHT (2005) Quality risk management. Q9, Current step 4:408

13. Guideline IHT (2011) Development and manufacture of drug substances (chemical entities and biotechnological/biological entities) Q11. FDA

14. von Stosch M, Hamelink J-M, Oliveira R (2016) Hybrid modeling as a QbD/PAT tool in process development: an industrial E. coli case study. Bioprocess and Biosystems Engineering 39:773–784. https://doi.org/10.1007/s00449-016-1557-1

References

15. Herwig C, Glassey J, Kockmann N, et al (2017) Better by Design, Qualtiy by design must be viewed as an opportunity not as a regulatory burden. The Chemical Engineer Issue 915:

16. Zurdo J, Arnell A, Obrezanova O, et al (2015) Early Implementation of QbD in Biopharmaceutical Development: A Practical Example. BioMed Research International 2015:1–19. https://doi.org/10.1155/2015/605427

17. Guerra AC (2018) Machine learning in biopharmaceutical manufacturing. European Pharmaceutical Review 23:62–65

18. Yang Y, Ye Z, Su Y, et al (2019) Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B 9:177–185. https://doi.org/10.1016/j.apsb.2018.09.010

19. Borchert D, Suarez-Zuluaga DA, Sagmeister P, et al (2019) Comparison of data science workflows for root cause analysis of bioprocesses. Bioprocess and Biosystems Engineering 42:245–256. https://doi.org/10.1007/s00449-018-2029-6

20. Suarez-Zuluaga DA, Borchert D, Driessen NN, et al (2019) Accelerating bioprocess development by analysis of all available data: A USP case study. Vaccine 37:7081–7089. https://doi.org/10.1016/j.vaccine.2019.07.026

21. Borchert D, Zahel T, Thomassen YE, et al (2019) Quantitative CPP Evaluation from Risk Assessment Using Integrated Process Modeling. Bioengineering 6:16. https://doi.org/10.3390/bioengineering6040114

22. Borchert D, Suarez-Zuluaga DA, Thomassen Y, Herwig C Risk Assessment and Integrated Process modelling – An Improved QbD approach for the Development of the Bioprocess Control Strategy. Submitted to AIMS Bioengineering