ELSEVIER

Teaser This contribution reviews the current state of data science in biopharmaceutical environments and outlines a roadmap towards a more powerful and flexible biopharma life cycle in the future.

Data science tools and applications on the way to Pharma 4.0

Valentin Steinwandter[1], **Daniel Borchert**[1] and **Christoph Herwig**[2,3]

[1] Exputec GmbH, Mariahilferstr. 147/2/2D, Vienna, Austria
[2] Institute of Chemical Environmental and Bioscience Engineering, Research Area Biochemical Engineering, Vienna University of Technology, Gumpendorferstr. 1a, Vienna, Austria
[3] CD Laboratory on Mechanistic and Physiological Methods for Improved Bioprocesses, Vienna University of Technology, Gumpendorferstr. 1a, Vienna, Austria

Multiple obstacles are driving the digital transformation of the biopharmaceutical industry. Novel digital techniques, often marketed as 'Pharma 4.0', are thought to solve some long-existing obstacles in the biopharma life cycle. Pharma 4.0 concepts, such as cyberphysical systems and dark factories, require data science tools as technological core components. Here, we review current data science applications at various stages of the bioprocess life cycle, including their scopes and data sources. We are convinced that the scope and usefulness of these tools are currently limited by technical and nontechnical problems experienced during their development and deployment. We suggest that the establishment of DevOps mind- and toolsets could improve this situation and would be essential cornerstones in the further development of Pharma 4.0 systems.

1 Introduction

New techniques developed in the fields of information technology and data science are currently disrupting the manufacturing industry. Industry 4.0 is considered to be the fourth industrial revolution, after the first; [mechanical production facilities (starting at the end of the 18th century)], the second [electricity, assembly lines and mass production (1870s)], and the third revolutions [automation, computers and electronics (during the 1970s)]. The term 'Industry 4.0' originates from a German initiative named 'Industrie 4.0' from 2011 that had the goal to improve the competitiveness of the manufacturing industry in Germany [1]. Two years later, in 2013, the German Government started to support the idea and published recommendations for its implementation [2].

Although the phrase 'Industry 4.0' lacks an exact and generally valid definition, most researchers agree that Industry 4.0 is about the implementation of cyberphysical systems (CPS), the Internet of

Valentin Steinwandter leads the software development group at Exputec, which works on a data science framework for the biotech industry. After his MSc in biotechnology, he started to work at the interface between data science, software, and biotechnology. He recently finished his PhD at the Technical University of Vienna, where he focused on various data science applications and workflows throughout the bioprocess life cycle.

Daniel Borchert is a biotechnologist with training in nanobiotechnology and nanoscience; he joined Exputec in 2016 as a product manager and consultant. He is responsible for data science services and software implementations at industrial biopharma and contract manufacturing companies. He has been working with leading companies within the biotech as well as the food industry all over the world. Additionally, he is PhD candidate at the Technical University of Vienna, with a research focus on the development and application of data science tools to accelerate bioprocess development.

Christoph Herwig, a bioprocess engineer from RWTH Aachen, worked in industry in the design and commissioning of large chemical facilities before enter his interdisciplinary PhD studies at EPFL, Switzerland in bioprocess identification. Subsequently, he positioned himself at the interface between bioprocess development and facility design in biopharmaceutical industry. Since 2008, he has been a full professor of biochemical engineering at the Vienna University of Technology. His research focuses on the development of data science methods for integrated and efficient bioprocess development along QbD principles. Product fields include circular economy and biopharmaceuticals within industry-driven projects. In 2013, he founded Exputec, which addresses data science solutions for the biopharma life cycle.

Corresponding author: Herwig, C. (christoph.herwig@tuwien.ac.at)

https://doi.org/10.1016/j.drudis.2019.06.005

Things (IoT), and smart factories. According to CyPhERS, a CPS comprises computation, communication, and control components tightly combined with physical processes of different nature (e.g., mechanical, electrical, and chemical) [3]. Smart Factories are a specific example of a CPS and of high interest in biopharmaceutical manufacturing. IoT is often considered as a subset of CPS, interconnecting 'real' and 'digital' components through the use of embedded chips, sensors, and networks [4]. However, the exact meaning of these terms differs between market segments.

Over the past few years, the pharmaceutical industry has begun to show interest in the adoption of Industry 4.0 concepts. This is surprising, given the pharmaceutical industry is still very focused on its precursor, the digital transformation. Many processes that are now state of the art in pharma originated from old, paper-based concepts. However, Pharma 4.0 is a hot topic because Industry 4.0 concepts have the potential to solve some of the biggest obstacles currently facing the industry (Table 1).

The defined goal of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) is 'to ensure that safe, effective, and high quality medicines are developed and registered in the most resource-efficient manner'. Although most would agree that current pharmaceuticals are safe, effective, and high quality, there exists serious doubt whether the life cycle of pharmaceutical development, manufacturing, and registration is resource efficient. Multiple studies have attempted to estimate the total costs of development for a new pharmaceutical product. The numbers vary widely, given that different estimates start with different assumptions. However, the estimates of total development costs until approval range from US$161 million (2009) to US$2.87 billion (2013) [5–7]. Even more crucial, the time span between preclinical studies and the approval of a drug is, on average, 12 years [8].

Additionally, big pharmaceutical companies encounter an increasing need for competitive, optimized processes as their current main sources of income, patent-protected products, run out. This will lead to higher competition with generic products (i.e., biosimilars). Whereas the past decades allowed high profit margins without optimal product conditions, this could change as soon as more competitors push into the market. Another driver for Pharma 4.0 is the Quality by Design (QbD) initiative of the US Food and Drug Administration (FDA), demanding data- and risk-based approaches for the development and manufacturing of pharmaceuticals. Many concepts of Pharma 4.0 are considered to have the potential to solve some of these issues.

Here, we focus on a small but central part of Pharma 4.0, reviewing currently applied data science tools and methods and showing where they are applied today, where and how they can be implemented in the future, and which benefits could be expected from them. To highlight the importance of data science in Industry 4.0, we explain what Industry 4.0 implies. It is not only about the automation and digitalization of processes (which has already happened), but also about the automation of decisions. Cyber-physical systems (i.e., interconnected systems of physical machines that are controlled by software and algorithms) are likely to become autonomous units, being able to function without manual interventions, and delivering quality by design. Algorithms will be likely to take over many of the decisions currently made by slow and error-prone humans, and make them in real-time, using all the available data. Therefore, it is crucial to understand where we are currently, at which points in the biopharmaceutical life sciences algorithms already help, and how this could evolve. Thus, we review state-of-the-art data science methods, considering their data sources and scopes along the product life cycle, from process development, through process validation, routine manufacturing, and product life cycle management. Furthermore, we investigate which directions data science methods and implementation techniques need to be developed in to allow its incorporation as a central part of Pharma 4.0 in the biopharmaceutical product life cycle. We also highlight challenges that need to be solved and how the knowledge and experience from other industries can help to solve these.

2 Data science challenges because of biopharma particularities

Data scientists in the biopharmaceutical industry have to conquer different obstacles. To understand those difficulties, we discuss the nature of the bioprocesses and their life cycle, because they directly impact the requirements for data management and analysis. Here, we show the context and present the difficulties and specialties of data science for biopharma, whereas later we present commonly used data science tools in detail and categorize them based on their use during the bioprocess life cycle.

TABLE 1

Current obstacles in the biopharmaceutical manufacturing industry and how data science could solve them

Current obstacle	Possible solution
Development and approval of new pharmaceuticals is too expensive, and the time to clinic is too long	Simplified application of data science tools for smart experimental design, data-driven decision-making, and optimization of processes
Processes are not transparent	Difficult or impossible-to-measure variables can be estimated indirectly using predictive tools
Routine manufacturing processes are not efficient, run suboptimally, or result in failed batches	Optimize performance of already established processes by applying optimization methods; implement improvements by considering ICH Q12 guidelines for product life-cycle management
Effect of multidimensional changes in process parameters, as they occur along process value chain, on CQAs is unknown	Apply process modeling to obtain data-driven estimates of parameters changes, considering whole process value chain
People with the correct set of skills are rare and, therefore, expensive and limited in their time	Algorithms can take over decisions that are currently based on manual analysis of data
People have limited capacity for analyzing and interpreting high-dimensional data sets	Algorithms can be set up to take over data interpretation

2.1 Bioprocesses are mostly batch based

First, we discuss a typical, exemplary production process. Although continuous processes are well established in many industries, biopharmaceutical manufacturing relies significantly on batch processes. Such a process can be split into serially performed unit operations (UOs) [9] (Fig. 1). The number of unit operations, as well as the splitting and pooling steps within such a process, are product dependent. Usually, the two main categories for unit operations in biopharma are upstream processing (USP) and downstream processing (DSP). The aim of USP is to generate the desired product in a certain concentration, whereas DSP results in its separation, purification, and polishing [10]. Variations in one of the Critical Process Parameters (CPPs) of one unit operation can have change the process conditions and product quality attributes in one of the subsequent steps [11].

2.2 Processes can split and merge

Fig. 1 shows that, at a certain point, the process splits from a serial process to a parallel one. Two pooling steps follow this split before the product is finally purified. One reason for such a complex process, splitting and pooling, is often the limitation in working volume of certain devices used within the process. Chromatographic steps and the columns used in particular are limited in the working volume [12]. To make data science possible, the data sources and data management systems need to consider these difficulties and allow both the structured storage of data and the simple retrieval of all related data.

2.3 The main goal is risk reduction

Although the main goal of data science in many industries is to improve productivity or reduce production costs, this usually is not the main aim of data science in biopharma. Given the regulatory environment, few experiments involved, high costs for failed batches, and extremely high risks of products being developed but not successfully commercialized, the main focus for data science in biopharma is process transparency and process robustness. The first aim is risk reduction, after which companies start to think about productivity [13]. This strongly impacts the applied tools and methods used.

2.4 There are too few experiments

Fast process development, rapid clinical trials, high production costs of complex processes, as well as the risk of failed batches result in one of the major issues when it comes to statistics and data science: the

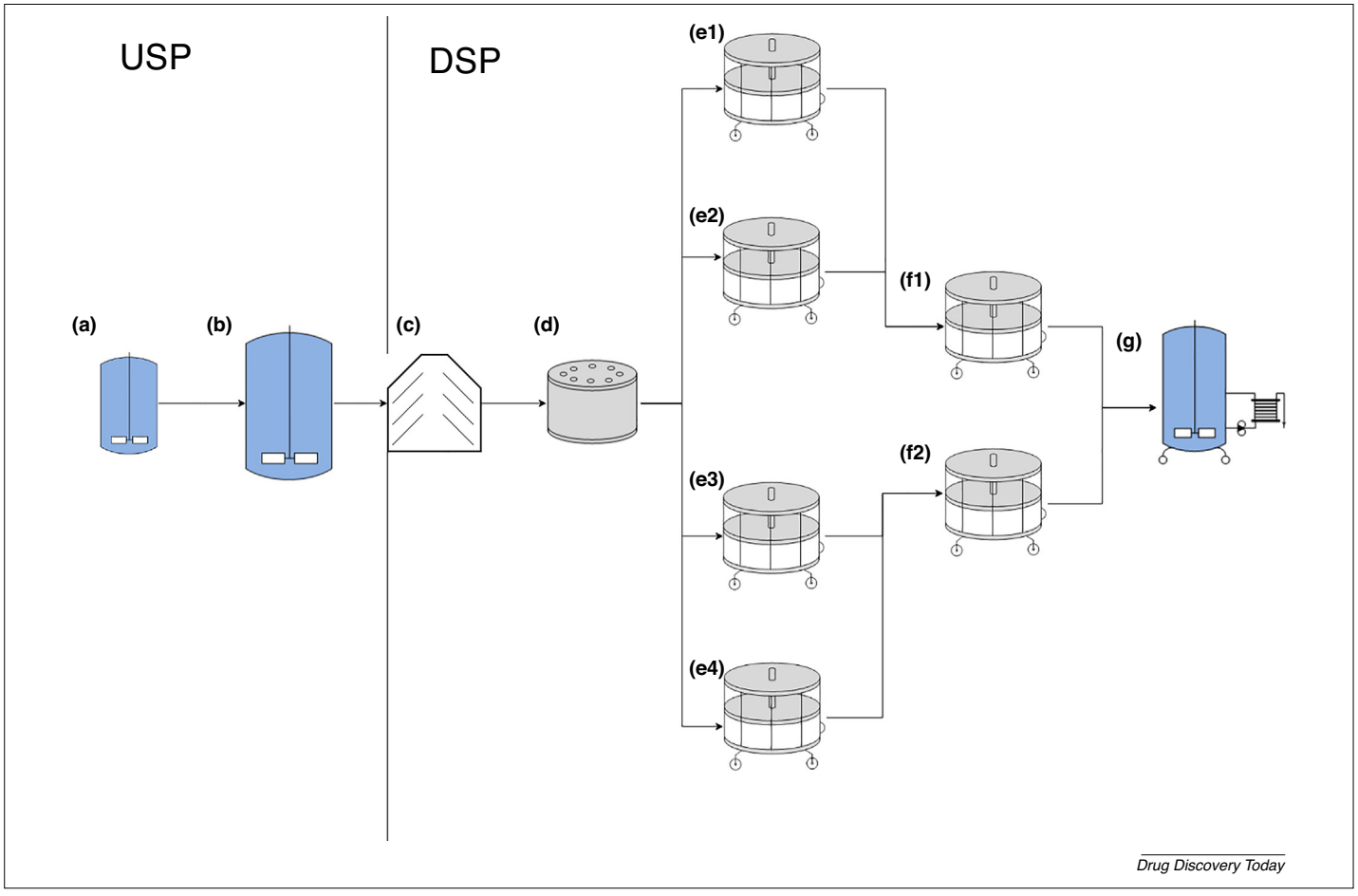

FIGURE 1

The ideal process flow of a batch-based production process in biopharma: The entire process is generally divided into upstream processing (USP) and downstream processing (USP). USP aims to produce the active pharmaceutical ingredient (API), whereas DSP handles all the separation and purification steps. In this example, USP has two distinct steps, a smaller seed reactor (a) followed by a larger main fermentation (b). DSP usually starts with separation (c) followed by a filtration (d). For the first step of the purification, the filtered product is split into multiple chromatographic columns (e.1–e.4). Pooling steps between the different chromatographic purifications are possible (f.1–f.2). Finally, a diafiltration step (g) removes the remaining impurities.

$n<<p$ problem [14]. 'n' is the number of observations and 'p' the number of variables per observation; a low n is the standard case in biopharmaceutical statistics. The n<<p problem is a statistician's nightmare because frequent statistical assumptions (normality and homogeneity of variances, independence of errors, etc.) are hard, if not impossible, to assess [15]. This 'problem' needs to be considered during the statistical analysis, because statistical good practice demands to have more observations than variables ($n>p$) and most available statistical methods have this prerequisite. Nevertheless, special techniques have been developed to deal with such data sets. Currently, the most prominent techniques are shrinkage methods, such as Least Absolute Shrinkage and Selection Operator (LASSO), ridge regression, and elastic net [16,17].

3 Tasks and tools during the bioprocess life cycle

Here, we present various data science methods and how they are applied through the biopharmaceutical life cycle. The methods and tools are categorized depending on their application in a biopharmaceutical process life cycle into four categories: (i) process development and scale-up; (ii) process validation and characterization; (iii) routine manufacturing; and (iv) life-cycling, including Q12 and Continued Process Verification (CPV). For each category, we discuss: (i) the scope and expected results of the data science tools (i.e., which problems should data science solve) and (ii) the used data sources (i.e., in which formats and from which interfaces the data can be consumed); and (iii) provide a description about the actual tools.

We do not discuss the huge amount of data science applications used for chemometrics, such as chromatography or spectroscopy. Even though these systems are an integral part of Process Analytical Technology (PAT) and Pharma 4.0, they are outside the scope of this publication. Chemometrics are already well described elsewhere [18].

3.1 Process development and scale-up

Although the market price of pharmaceuticals appears to be independent of their development costs [19], the development of new pharmaceutical products is expensive and, because of the low success rates (most products never get approved), a risky venture. Pharmaceutical companies try to reduce these risks by analyzing all of the available data to assess the potential of their future possible products. Recently, the pharmaceutical industry began to think about collaborative research and development (R&D) to reduce the cost of failures and to increase success rates [20]. Increased collaboration in this field would further increase the need for data science tools and additionally require more straightforward implementation and use, compared with current possibilities.

3.1.1 Scope and expected results

The aim of process development is to establish a robust and scalable process, to optimize already developed processes, and/or to accelerate new processes. The optimized process [21] should fulfill the regulator requirements of the authority in any performance mode and should save production time and costs. Additionally, in the case of complex bioprocesses, the identification or improvement of operation ranges [22], as well as the improvement and refinement of process knowledge and understanding [23,24], are vital.

3.1.2 Data sources

We have been working on data science applications for many years, sometimes together with small biotech start-ups and sometimes with big international companies. Mostly, the data sources for the data science were, simply put, a big mess. The data collection and cleansing phase were always a major part of the whole project and the main results were sometimes not an insightful model but a well-structured and cleaned-up data set.

At this stage of the bioprocess life cycle, many different data sources exist. Generally, the use of Excel and similar spreadsheets is widespread, and some data are even available just as handwritten notes in lab notebooks. Other data sources are printouts from stand-alone devices (e.g., analytical measurement devices). These manual notes and printed data sheets have to manually transferred into an electronic format to make computational data analysis possible. We also encountered data simply stored inside of Matlab programming scripts, mixed up with analysis code, and hardly identifiable as raw data.

In more sophisticated cases, data are stored in commercially available databases, such as in Laboratory Information Management Systems (LIMS) or in database systems set up and managed by the companies themselves. In case of high-frequency time-series data, also known as online measurements, multiple systems exist. Multibioreactor systems store the recorded data in internal databases, with the ability to export the data as Excel or CSV files [25,26]. Although there exists no common data format between the different companies, most vendors usually stick to the same file formats for many years.

3.1.3 Methods

The diversity of data sources and the different process development aims require flexibility in data analysis methods. One Factor At a Time (OFAT) experiments for the optimization of processes was the standard method for a long time [27]. The pharmaceutical industry required more time compared with other industries for the adoption of experimental planning supported by multivariate statistics [i.e. the Design of Experiments (DoE)]. DoE uses statistical models for the efficient planning of experiments and is currently one of the most commonly used tools [21]. Less-known alternatives to multivariate statistical methods are optimizer-based approaches (e.g., by applying a simplex algorithm) [22,28]. Both DoEs and optimizer-based approaches often result in response surface diagrams that are used for the identification of process operation ranges. A more advanced tool is the use of model-based experimental design, which combines mechanistic process modeling [29] with statistical analysis [30,31]. Additionally, a DoE combined with prior knowledge [23] as well as hybrid models [e.g., the combination of artificial neuronal networks (ANNs) and Multivariate Data Analysis (MVDA)] [32] are tools used in the process development.

Besides experimental design methods, root-cause or cause-and-effect analyses are widely used in the industry to generate new process knowledge [24,33]. For scale-up, standard MVDA data analysis tools have been applied for some years [34]. Recently, a combination of different MVDA methods with genetic algorithms for the prediction of scale-up variables was investigated [25]. Combined factor screening and characterization definitive screening design [26] are sophisticated analytical methods currently applied in process development.

3.2 Process validation and characterization

According to the FDA, process validation should provide scientific evidence for the safety, efficacy, and quality of the product. These criteria cannot be achieved by in-process controls or final product testing, but must be implemented into the process [35].

3.2.1 Scope and expected results

As part of the process validation, pharmaceutical companies use process characterization studies to provide the requested scientific evidence for process understanding and robust process control [36]. For the characterization of biopharmaceutical processes, mainly data from process development are used. This includes data from small to pilot-scale studies.

Given various economic and practical reasons, performing the process characterization studies on manufacturing scale is unfeasible. Therefore, scaled-down models of large-scale manufacturing runs are developed and qualified. During these steps, it is the goal to prove that there is no significant difference between the small-scale experiments and the large-scale runs and, therefore, that more process characterization experiments need not be run on a manufacturing scale [37].

3.2.2 Data sources

The data sources from these studies are similar to those discussed elsewhere in this article. However, the data sources mentioned in relation to routine manufacturing also have an important role here.

3.2.3 Methods

One of the main goals during process validation is to increase the understanding of a process. Therefore, process modeling methods are applied [38]. They focus on single process steps (e.g., one unit operation) or a complete process chain (e.g., through an integrated process model) [11]. It is the goal of the process characterization study to understand what happens if multiple CPPs are near their control limits. Furthermore, it is possible to use the outcomes of process characterization studies to decrease the likelihood of out of specification events and make the process more robust in general [11].

Proving statistical significance might sound like an easy task, but most statistical tools are good to show differences, not equality. For example, a two-sample *t*-test has a null hypothesis of equality; the analyst must show that the two samples are different to show inequality. This is a problem, because discarding the null hypothesis might not be possible when the precision or the number is low. The two one-sided test (TOST) is a univariate equivalence test that starts with a null hypothesis and appropriately handles cases with poor precisions or low numbers of samples [39,40]. For multivariate equivalence testing, the Generalized Mahalanobis Distance (GMD) approach is applied [41,42].

3.3 Routine manufacturing

During the phase of routine manufacturing, fully developed and validated processes are applied on a large scale. Given that the costs for one of these runs are immense, the main goal is to ensure that the processes run within the ranges specified during the process development and validation phase, preventing out of specification events and failed batches. Given high regulatory restrictions, possible improvements to the processes cannot be applied without re-running the previous validation phases, which mostly results in 'frozen' processes once they have been approved.

3.3.1 Scope and expected results

The main scope of data science in routine manufacturing is the process monitoring (i.e., to confirm that all CPPs lie within the expected ranges or early fault detection) [43]. Other possible applications are the mining of process data to discover high-productivity characteristics [44], or the online state and parameter estimation and real-time control with or without models, enabling Advanced Process Control (APC) strategies [45].

3.3.2 Data sources

Some of the most important data sources in routine manufacturing are manufacturing execution systems (MES). These systems are used to control and document the manufacturing processes. In particuar, process variables measured in real time with online or at-line sensors or devices directly end up in these systems.

LIMS systems are important data sources when it comes to analytical measurement results. Excel files are also used; the GMP validation, in this case, is challenging but possible [46]. Other data sources are paper-based or electronic batch records. For various reasons (e.g., misunderstood regulatory guidelines [47]), the transition from paper-based to electronic batch records remains ongoing, which is a significant issue. The manual extraction of data from documents that are hundreds of pages long is a time-consuming and error-prone task and, therefore, often not executed. However, for some data, batch records are the only source. Electronic batch records, often integrated into the MES system, allow the straightforward use of manufacturing data for data science purposes.

3.3.3 Methods

One of the simplest approaches to monitor a process and detect outliers is to monitor each single measured variable by its own and assign lower and upper control limits. If one of the variables exceeds those limits, an out of specification (OOS) event is fired. This method has several drawbacks, especially when the number of variables increases.

Therefore, multivariate statistical approaches are used. These are especially convenient when many variables need to be monitored, because they stay easy to interpret. They also have an advantage if the criteria for the boundaries are more complex. For instance, a batch might be OOS although none of the variables per se is out of its single range. However, a certain combination of variables within a multidimensional space can be a clear indicator of a fault, even though the single variables do not show this information [48]. Possible multivariate tools applied are principal component analysis (PCA) [49], partial least squares (PLS) [50], and derived values, such as squared prediction errors (SPE), Q statistics, or Hotelling's T^2 [51].

Soft-sensors estimate variables that cannot be measured directly. Soft-sensors, or software sensors, are typically based on physical or chemical laws (e.g., by using first principle assumptions) [52,53] or are data driven [43]. Previously, they have been applied for the control of bioprocesses based on physiological rates [52,54]. In addition, ANNs were applied for the determination of unmeasurable media components (biomass, recombinant protein, etc.) [55].

3.4 Life-cycling, including Q12 and continued process verification

Life-cycling in this context describes a holistic view of the process development and manufacturing phases of a pharmaceutical

process. In contrast to the phases described earlier, this process phase is never over or finished. It starts with the process development phase and continues after the manufacturing phase has begun [56].

3.4.1 Scope and expected results

The main goal of life-cycling is to set up a quality and process control environment that guarantees a high quality of the products even though the process itself needs to be changed. Changes in manufacturing runs should not enforce a complete revalidation of the process. The processes should already be defined in a way that they can be changed and optimized during their life cycle, without compromising quality and an increased risk for the patients. The possible advantages of postapproval changes are higher productivities, shorter times to market, and more secure products. It is expected that Pharma 4.0 and its components will be indispensable for achieving these goals.

The ICH Q12 guideline aims to make postapproval chemistry, manufacturing and controls (CMC) changes more predictable, leading to continuous improvements and higher efficiency. It should allow adjustments and prevent frozen, suboptimal processes because of regulatory burdens that are too high [56]. Compared with traditional process monitoring, its main goal is to be proactive through predictive monitoring (e.g., by early detection of small drifts in the manufacturing processes).

Real-time release testing reduces the number of end-product tests by making use of fully automated and integrated manufacturing and testing. It uses real-time in-process tools to continuously monitor and control the pharmaceutical manufacturing process [57]. However, in biopharma, real-time release testing is in its early stages.

Personalized medicines require a completely new concept of process control and monitoring. If the manufacturing of a drug is modified based on the needs of individual patients, the quality and process control tools need to adapt to this fact.

3.4.2 Data sources

For the implementation of life-cycling approaches, more so than in the previous sections, complete, well-structured, and immediately available data are required. It is not enough to look at single unit operations or substeps of the manufacturing, but is instead necessary to look at the whole process in both directions (i.e., between unit operations and along the process development life cycle).

Large commercial providers of data historian solutions started to market their solution by using the Industry 4.0 buzz phrase [58,59]. Alternative solutions also come from smaller software providers and are sometimes set up by the companies themselves. Data historians might be the main data sources for the purpose of life-cycling, because they collect, store, and provide data for centuries. At the same time, commercial historians have their most significant advantage in the efficient storing of data with the best compression algorithms. By contrast, others are of the opinion that data historians are no longer appropriate, seeing major problems in their pricing models (pay per tag), which makes their use expensive for highly variable data types. Additionally, hard drive space is currently very cheap [60].

3.4.3 Methods

The data science methods described in the previous sections are tools, applied in small, well-defined areas, for clearly defined use cases during specific phases of the bioprocess life cycle. Methods for life-cycling are more general, and, because of their higher complexity, some are ongoing developments. All use cases discussed here require the long-term storage and simple reuse of once-generated knowledge.

Knowledge management tools organize the acquisition, storage, analysis, and dissemination of knowledge during the bioprocess life cycle [61]. Theoretical constructs for knowledge management systems have been around for many years and numerous publications have already discussed the proposed concepts. However, the practical implementation of these concepts in the real world is lagging behind [62].

Although it is currently relatively simple to store large amounts of data, it is more complex to store large amounts of knowledge in a form that allows simple reuse and recombination. Specifically, the knowledge that was generated during the process development and process validation phase should be reused for manufacturing. Feedback loops between these interconnected steps should; for example, if manufacturing runs show new variations that were not in the validation phase, the reasons (e.g., other raw material or equipment faults) should be investigated.

Models are a highly condensed form of knowledge. Comprehensive models, often referred to as digital twins [63,64], should allow one to estimate whether and how changes to an existing process affect other steps in the process chain. Digital twins should learn from newly generated data (e.g., also using the input of PAT devices). Updated knowledge could lead to changes in the addition, change, or removal of CQAs or CPPs.

4 Discussion

Here, we summarize the challenges for data science in a Pharma 4.0 environment and discuss which technologies and processes might help to solve these challenges in the future.

4.1 Development and deployment of data science tools

Currently, technical and nontechnical burdens cause multiple obstacles during the development and deployment of data science tools (Table 2). The actual tasks of algorithm development and deployment strongly rely on data access, the presence of test environments, and fast deployment with short feedback cycles. However, the current situation in the industry is entirely different. Data access is highly restricted, for example because of intellectual property (IP) protection (the company shares as many data as required) or technical reasons (the interfaces are not in place or difficult to use) [65,66]. Test environments with simulated manufacturing sites and simulated real-time modes are often not good enough, which makes the development of an algorithm even more difficult. Often, algorithms need to be developed based on file exports. In other cases, virtual machines are available to do the development and first tests in pseudo-real-time modes. The deployment to production is also a complex process full of issues. First, the operational environments and manufacturing software are usually not developed with data science and algorithms in mind [67]. Furthermore, applying any changes in an already validated environment is difficult, if not impossible [56]. Finally, it takes much time, often months or even years, until a developed algorithm can finally run on production data.

TABLE 2

Current obstacles in the development, deployment, and testing of data science tools in the biopharmaceutical industry[a]

Obstacle	Technical burden	Nontechnical burden	Refs
Data access	Data access is either not possible or difficult because of nonstandardized or nonavailable interfaces; too many different data sources exist (see also Table 3 in the main text)	Data access is highly restricted because of IP protection	[65,66]
Development	Data science tools that integrate into the whole process chain are impossible to develop and test because proper virtual environments for development and testing are not available	Conservative mindsets and long update cycles do not allow fast development with short feedback cycles	[67]
Deployment	Algorithms cannot be deployed because the software environments currently in production facilities were not designed with data science applications in mind	Tools cannot be set up and used because significant changes require revalidation of the whole process	[56,67]
Knowledge	It is not trivial to store developed models, tools, test results, and so on in a way that others can reuse the algorithms	Many algorithms are developed in academic environments, but academia has not enough knowledge of the day-to-day problems in the industry; by contrast, the industry knows the problems but does not have time, money, or required skilled people to develop and deploy necessary algorithms	[61,62]

[a] These issues become even more important if the required tools become more complex, as they are supposed to do in the Pharma 4.0 context.

4.2 Many different data sources

A significant issue on its own is the vast amounts of different data sources. Table 3 provides an overview of the most important data sources present during the bioprocess life cycle. Some of the data sources (file-based exports) are simple to set up and use, but have some drawbacks during the analysis (e.g., because of inconsistent formats and an increased effort required for data cleansing and restructuring). Other sources (e.g., data from data warehouses or data historians) have already been structured and standardized during the set-up phase of the systems. Here, the initial set-up effort of the systems itself is higher, although is easier to access and use the data for data science purposes.

4.3 With the current systems and mindsets, Pharma 4.0 is not going to happen

The complexity of biopharmaceutical processes as well as restrictive IT/OT environments in pharmaceutical companies are a significant burden for the implementation of data science tools. We think that the current feedback cycles for the development and deployment of such algorithms are too long. The data science applications used currently, as discussed earlier, are mostly passive in nature and used for monitoring or analytical purposes. Rarely, algorithms are part of real-time control loops and are rarely solely responsible for crucial, proactive decisions. Industry today is still far away from Pharma 4.0 and outcomes such as self-learning factories and production of personalized medicines.

4.4 How to make Pharma 4.0 possible

Many of the obstacles discussed earlier are not limited to data science in pharma. It is generally difficult to find the correct compromise between operational stability and fast development. Other industries, such as software development and operations industries, had and have the same challenges. They have developed tools and cultures to overcome these obstacles. As we discuss here, some of these tools and cultures might help to move the pharmaceutical industry to a technological level at which more powerful data science applications become realistic.

4.5 DevOps culture and technology

We are convinced that development and feedback cycles can be shortened by adapting strategies from software development environments. Agile development, continuous integration, automated tests, and deployment strategies have been used in this field for many years and are state of the art in most projects. These principles can be adapted to the biopharmaceutical manufacturing industry and data science applications.

For instance, data scientists need sandbox environments where they have access to the data they need. Sandbox environments are 'playgrounds' strictly separated from production environments that allow the straightforward development and test of new tools without the risk of breaking the production systems.

For the implementation of such tools, virtual environments are desirable. Given the complexity of real production systems, it will not be possible to implement virtual clones of them in the near future. However, the developments of these virtual environments could start with the virtualization of single compartments and, at some point, link these together to become fully virtual representations of the production facilities. Virtual clones of production facilities would be valuable tools for the fast development, implementation, and testing of new data science applications. The deployment of these applications should be standardized and, at any point (even in the production environment), the whole code stack should be able to roll back to a previous version.

Table 4 details what such a deployment strategy could look like. It is similar to state-of-the-art strategies for the deployment of complex software products and environments.

4.5.1 Open standards and interfaces for data and knowledge

It would be wrong to say that there are no standards for data formats and interfaces: the problem is that there are too many. For various technical and historical reasons, the number of possible data sources is extremely high and causes significant amounts of work for data scientists. The industry needs to move towards powerful, standardized interfaces, because this is a clear bottleneck of current development. These obstacles will become even more

TABLE 3

Overview of data sources

Category	Examples	Advantages	Disadvantages	Fast set-up and low IT/OT costs	Data science value and 'usability'	Refs
Tabular files	.csv,. txt files, Excel, LibreOffice, DASGIP, or ambr exports automatically or manually generated	Easy to set up and use; portable file format, which does not require a complex IT environment	Too many different variants and inconsistencies because of manual file changes lead to significant effort required for data collection and cleansing when a large number of files need to be analyzed	+++++	+++	[24,46]
Hand-written laboratory results	Paper-based laboratory notebook	No IT required	No automated data analysis possible; data need to be transferred into another system first	+++++	+	[69,70]
Direct interface to analytical device or sensor	OPC (UA, HDA, DA), serial ports, field bus	Simple set-up if open interface is used; direct access to data in rea time; most direct way to implement smart sensors if all data are available	Most of metadata is missing; only simple algorithms can be used without any further additional data sources (e.g., through user input or additional data source)	+++	++	[71]
Distributed control systems	Emerson Delta-V, Siemens Simatic PCS7	Real-time connection to connected systems	Main purpose is automation, not data delivery; therefore, often difficult to access (historical) data	++	++	[72,73]
SCADA/MES systems	Sartorius MFCS, Infors Eve, Applikon BioXpert	Structured data formats and access through standardized interfaces; many process data are available for analysis	Needs set up and availability of interface; often, metadata to the process itself are missing (but required for the analysis)	+++	+++	[25,26]
LIMS systems	LabWare, LabVantage, Starlims, McKesson Lab	Standardized data source for analysis results from laboratory	Requires additional meta data to align these data with process data from other systems	+	+++	[74]
Data warehouses	Amazon Redshift, Microsoft SQL Server, SAS Data Management, Oracle Database	Developed for simplified and fast access for analytical purposes; holds data from different departments and origins	Significant effort to set up systems	+	+++++	[75]
Data historians	OSI Pi, Wonderware	Developed and used to store large amounts of mostly time-series process data; usually offers different APIs to allow data access (e.g., SQL, REST)	Significant effort to set up and maintain systems	+	+++++	[58,59]
Data lakes	Distributed file systems, (e.g., Apache Hadoop, Azure Data Lake, Amazon S3)	Developed and used to make saving and storage of structured and unstructured data easy; no data cleansing and preparation necessary to store data	As data are unstructured, data science algorithms need to prepare data before analysis	+++	++	[76]
Paper-based batch records	–	Extremely valuable data that cannot be found anywhere else, (e.g., holding times, exceptions, or out-of-specification events)	Extremely high effort for data collection and cleansing	+++	+	[77]

TABLE 3 (*Continued*)

Category	Examples	Advantages	Disadvantages	Fast set-up and low IT/OT costs	Data science value and 'usability'	Refs
Electronic batch records	Emerson Syncade, ampleLogic eMBR	Same as paper-based batch records, but less effort required for data collection and cleansing	Huge amount of data, still need much manual work and interpretation	+	+++	[78]
Others	.pdf,. doc(x),. odt documents, images	Huge amounts of data available in these formats (e.g., already-interpreted data in form of reports, or raw data in form of images)	Data are often unstructured and require much manual work to make them usable for data science	++++	++	[79]

TABLE 4

Deployment scenarios for the implementation of data science components in pharmaceutical manufacturing environments[a].

Environment	Description	Virtual
Development and sandbox environments	Data scientist works on new algorithm; developers and IT integrators work on implementation	√
Testing and integration environment	Algorithm and implementation are tested in test environments; unit and integration tests are set up and automatically applied	√
Staging environment in virtual facility	Staging environment is virtual clone of production facility; same historical data as in the production environment are available; IT section is equal to that at the production facility, whereas the physical components are simulated, as well as the processes themselves	√
Production environment in real facility	New functional increment is deployed to production facility; it is possible to revert deployment and rollback to a previous version	X

[a] The code, at every stage, is versioned and deterministically describes the behavior of the system. This allows version rollbacks and fast feedback loops. Infrastructure as Code (IasC) elements might be vital to reach these goals [80].

important within the coming years, given that the amount of generated process and analytics data is growing continuously (e.g., such as our ability to apply 'near online' measurements of LC-MS systems during biopharmaceutical manufacturing [68]).

However, is not solely about standards for data or interfaces (Table 3); it is also about standards to enable knowledge management, given that this needs to move from theory to real-live implementations and support the Pharma 4.0 visions.

5 Concluding remarks

Data science is a core component of many Pharma 4.0 ideas. Without algorithms that take over decisions and control the processes on their own, ideas such as personalized medicines or dark factories will not become a reality any time soon.

As shown in this review, most of the data science tools applied in the industry are still relatively basic and, most of the time, focus on subparts of the whole process flow and product life cycle. Currently, biopharma is a challenging environment for data scientists. The lack of proper data formats and standardized interfaces, the low number of experiments, conservative IT/OT environments, and regulatory restrictions currently prevent the development, deployment, and use of more powerful tools.

Another problem is the general mindset. Biopharmaceutical companies spend large amounts of money on their IT/OT infrastructure, firewalls, and other security-related systems to protect their data. Sadly, this is a double-edged sword. The more restrictive, conservative, and closed these systems become, the more difficult it is for data scientists and engineers to extract the hidden value from the data. Data and knowledge are thought to be the main capital of this industry, but they are not properly used currently.

Nevertheless, we are convinced that the essential technologies for data science in a Pharma 4.0 environment are already available. Multivariate tools and statistical procedures that combine process knowledge with statistics (e.g., used in process development, validation, and manufacturing) are well established and widely known in industry and academia. The bottleneck that prevents their more powerful use is their limited scope because of the absence of interconnections, forming autonomous systems through the combination of physical machines and algorithms.

DevOps techniques and a DevOps mindset are vital, because data science algorithms require a life-cycle approach; they must be written, deployed, sometimes reverted, reworked, deployed again, tested, and in general, continuously improved, without gigantic deployment, test, or validation overheads. Although DevOps is successfully applied currently, especially for software as a service deployment, it is not yet fully established for IT/OT environments in pharmaceutical companies.

We are convinced that the following tasks are integral parts of evolving Pharma 4.0 tools: (i) set-up systems that allow proper testing of sophisticated tools (e.g., by developing virtual clones of production facilities, virtualizing not only the data flow, but also the physiological part of the processes); (ii) implement frameworks to define the data science infrastructure in the form of versionable code, allowing fast feedback cycles through easily patchable and revertible systems; (iii) use open, interoperable formats. Nothing hinders data scientists more than data that cannot be used, because of either their proprietary format or inconsistency; and

(iv) establish knowledge management tools that work in practice, allowing researchers to store and access knowledge as models, algorithms, or documents.

To make Pharma 4.0 possible, technical and nontechnical changes need to happen. Data scientists and process engineers need better access to data, standardized interfaces, and fast release cycles of new data science tools. Production environments need to become ready for a connected world and, as the complexity increases, proper development and deployment strategies need to become a core part of new facilities and production lines.

References

1 Hermann, M. *et al.* (2016) Design principles for Industrie 4.0 scenarios. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (XXXX, eds), pp. 3928–3937, IEEE. https://dl.acm.org/citation.cfm?id=2918269&picked=prox&preflayout=flat.

2 BMBF, Industrie 4.0; www.bmbf.de/de/zukunftsprojekt-industrie-4-0-848.html (accessed June 10, 2019).

3 CyPhERS (2013) *Characteristics, Capabilities, Potential Applications of Cyber-Physical Systems: a Preliminary Analysis*. CyPhERS

4 Jeschke, S. *et al.* (2017) Industrial internet of things and cyber manufacturing systems. In *Industrial Internet of Things* (Jeschke, S., ed.), pp. 3–19, Springer

5 Sertkaya, A. *et al.* (2016) Key cost drivers of pharmaceutical clinical trials in the United States. *Clin. Trials* 13, 117–126

6 Morgan, S. *et al.* (2011) The cost of drug development: a systematic review. *Health Pol.* 100, 4–17

7 DiMasi, J.A. *et al.* (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* 47, 20–33

8 Van Norman, G.A. (2016) Drugs, devices, and the FDA: Part 1. *JACC Basic Transl. Sci.* 1, 170–179

9 Sharratt, P.N. (1997) Chemicals manufacture by batch processes. In *Handbook of Batch Process Design* (Sharratt, P.N., ed.), pp. 1–23, Springer

10 Kumara Behera, B. and Varma, A. (2017) Downstream processing. In *Microbial Biomass Process Technologies and Management* (B. K. Behera, A. Varma, eds), pp. 109–214, Springer.

11 Zahel, T. *et al.* (2017) Integrated process modeling—a process validation life cycle companion. *Bioengineering* 4, 86

12 Coskun, O. (2016) Separation techniques: chromatography. *North. Clin. Istanb.* 3, 156–160

13 Kannt, A. and Wieland, T. (2016) Managing risks in drug discovery: reproducibility of published findings. *Naunyn Schmiedebergs Arch. Pharmacol.* 389, 353–360

14 BioProcess International (2016) *Addressing the Challenges of Developing Biopharmaceutical Drugs*. BioProcess International

15 Zahel, T. *et al.* (2017) Workflow for criticality assessment applied in biopharmaceutical process validation stage 1. *Bioengineering* 4, 85

16 Rasmussen, M.A. and Bro, R. (2012) A tutorial on the Lasso approach to sparse modeling. *Chemomet. Intell. Lab. Syst.* 119, 21–31

17 Ferreira, A.P. *et al.* (2018) XXXXXX. In *Multivariate Analysis in the Pharmaceutical Industry* (XXX, eds), pp. 297–301, Elsevier Science.

18 Lourenço, N.D. *et al.* (2012) Bioreactor monitoring with spectroscopy and chemometrics: a review. *Anal. Bioanal. Chem.* 404, 1211–1237

19 Io, C. (2018) Assessing pharmaceutical research and development costs.

20 Joseph DiMasi, A. (2018) *JAMA Internal Medicine* 178 (4) 587

21 Wechselberger, P. *et al.* (2012) Efficient feeding profile optimization for recombinant protein production using physiological information. *Bioprocess. Biosyst. Eng.* 35, 1637–1649

22 Konstantinidis, S. *et al.* (2012) The hybrid experimental simplex algorithm – an alternative method for 'sweet spot' identification in early bioprocess development: case studies in ion exchange chromatography. *Anal. Chim. Acta* 743, 19–32

23 Wang, Z. and Georgakis, C. (2017) An *in silico* evaluation of data-driven optimization of biopharmaceutical processes. *AIChE J.* 63, 2796–2805

24 Borchert, D. *et al.* (2019) Comparison of data science workflows for root cause analysis of bioprocesses. *Bioprocess Biosyst. Eng.* 42, 245–256

25 Sokolov, M. *et al.* (2018) Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target. *Biotechnol. J.* 13, 1700461

26 Tai, M. *et al.* (2015) Efficient high-throughput biological process characterization: definitive screening design with the Ambr250 bioreactor system. *Biotechnol. Prog.* 31, 1388–1395

27 Politis, S.N. *et al.* (2017) Design of experiments (DoE) in pharmaceutical development. *Drug Dev. Indust. Pharmacy* 43, 889–901

28 Chhatre, S. *et al.* (2011) The simplex algorithm for the rapid identification of operating conditions during early bioprocess development: case studies in FAb' precipitation and multimodal chromatography. *Biotechnol. Bioeng.* 108, 2162–2170

29 Kroll, P. *et al.* (2017) Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. *Process Biochem.* 62, 24–36

30 Franceschini, G. and Macchietto, S. (2008) Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.* 63, 4846–4872

31 Mears, L. *et al.* (2017) Mechanistic fermentation models for process design, monitoring, and control. *Trends Biotechnol.* 35, 914–924

32 von Stosch, M. *et al.* (2016) Hybrid modeling as a QbD/PAT tool in process development: an industrial E. coli case study. *Bioprocess Biosyst. Eng.* 39, 773–784

33 Gomes, J. *et al.* (2015) Integrating systems analysis and control for implementing process analytical technology in bioprocess development: integrating systems analysis and control. *J. Chem. Technol. Biotechnol.* 90, 583–589

34 Ahuja, S. *et al.* (2015) Application of multivariate analysis and mass transfer principles for refinement of a 3-L bioreactor scale-down model-when shake flasks mimic 15,000-L bioreactors better. *Biotechnol. Progress* 31, 1370–1380

35 FDA (2011) *Process Validation: General Principles and Practices*. FDA

36 Little, T.A. (2017) *Process Characterization Essentials, Model Optimization, and Controlling the Process*. BioPharm International

37 Li, F. *et al.* (2006) A systematic approach for scale-down model development and characterization of commercial cell culture processes. *Biotechnol. Progress* 22, 696–703

38 Kroll, P. *et al.* (2017) Model-based methods in the biopharmaceutical process lifecycle. *Pharm. Res.* 34, 2596–2613

39 Limentani, G.B. *et al.* (2005) Beyond the t-test: statistical equivalence testing. *Anal. Chem.* 77, 221A–226A

40 Ahlstrom, B. *et al.* (2009) *Analysis of Stability Data with Equivalence Testing for Comparing New and Historical Processes under Various Treatment Conditions*. AMGEN

41 Cao, Y. *et al.* (2018) Evaluating manufacturing process profile comparability with multivariate equivalence testing: case study of cell-culture small scale model transfer. *Biotechnol. Prog.* 34, 187–195

42 Hoffelder, T. *et al.* (2015) Multivariate equivalence tests for use in pharmaceutical development. *J. Biopharm. Stat.* 25, 417–437

43 Kadlec, P. *et al.* (2009) Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* 33, 795–814

44 Charaniya, S. *et al.* (2010) Mining manufacturing data for discovery of high productivity process characteristics. *J. Biotechnol.* 147, 186–197

45 Ramero-Torres, S. *et al.* (2018) Biopharmaceutical process model evolution- enabling process knowledge continuum from an advanced process control perspective. *Am. Pharm. Rev.* (July), In: https://www.americanpharmaceuticalreview.com/Featured-Articles/352447-Biopharmaceutical-Process-Model-Evolution-Enabling-Process-Knowledge-Continuum-from-an-Advanced-Process-Control-Perspective/

46 ECA Academy (2014) *What are the EU GMP Requirements for the Validation of Excel Spreadsheets?* ECA Academy

47 FDA (2003) *Guidance for Industry - Part 11, Electronic Records; Electronic Signatures — Scope and Application*. FDA

48 Albert, S. and Kinley, R.D. (2001) Multivariate statistical monitoring of batch processes: an industrial case study of fermentation supervision. *Trends Biotechnol.* 19, 53–62

49 Ayech, N. *et al.* (2012) New adaptive moving window PCA for process monitoring. *IFAC Proc.* 45, 606–611

50 Kourti, T. *et al.* (1995) Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control* 5, 277–284

51 Steinwandter, V. *et al.* (2018) Multivariate analytics of chromatographic data: visual computing based on moving window factor models. *J. Chromatogr. B* 1092, 179–190

52 Wechselberger, P. *et al.* (2012) Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes. *Bioprocess Biosyst. Eng.* 36, 1205–1218

53 Sagmeister, P. *et al.* (2013) Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development. *Chem. Eng. Sci.* 96, 190–198

54 Lu, F. *et al.* (2013) Automated dynamic fed-batch process and media optimization for high productivity cell culture process development. *Biotechnol. Bioeng.* 110, 191–205

55 Caramihai, M. and Severi, I. (2013) Bioprocess modeling and control. In *Biomass Now - Sustainable Growth and Use* (Matovic, M.D., ed.), pp. 159–165, InTech.

56 ICH (2017) *Technical and Regulatory Considerations for Pharmaceutical Product Lifecycle Management - Q12 (Draft Version)*. ICH

57 Jiang, M. *et al.* (2017) Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing. *Biotechnol. Bioeng.* 114, 2445–2456
58 OSIsoft (2016) *The PI System and Industrie 4.0.* OSIsoft
59 Koumoutsakis, J. (2018) Industry 4.0: Looking Beyond the Buzz. *Aspen Technology*
60 Matthew, L. (2015) Will the Data Historian Die in a Wave of IIoT Disruption? *Industrial Transformation Blog*
61 Herwig, C. *et al.* (2015) Knowledge management in the QbD paradigm: manufacturing of biotech therapeutics. *Trends Biotechnol.* 33, 381–387
62 Ragab, M.A.F. and Arisha, A. (2013) Knowledge management and measurement: a critical review. *J. Knowledge Manage.* 17, 873–901
63 Hong, M.S. *et al.* (2018) Challenges and opportunities in biopharmaceutical manufacturing control. *Comput. Chem. Eng.* 110, 106–114
64 Geris, L. *et al.* (2018) The future is digital: In silico tissue engineering. *Curr. Opin. Biomed. Eng.* 6, 92–98
65 OPC Foundation (2008) *Unified Architecture*. OPC Foundation
66 Davies, N. (2017) *Cyber Attacks: How Prepared is Pharma?* The Pharma Letter
67 Harp, D.R. and Gregory-Brown, B. (2017) *IT/OT Convergence - Bridging the Divide*. NexDefense
68 Dong, J. *et al.* (2016) High-throughput, automated protein a purification platform with multiattribute LC–MS analysis for advanced cell culture process monitoring. *Anal. Chem.* 88, 8673–8679
69 Faggella, D. (2018) AI in the life sciences: six applications. *Genet. Eng. Biotechnol. News* 38, 10–11
70 Petrone, J. (2018) Roche pays $1.9 billion for Flatiron's army of electronic health record curators. *Nat. Biotechnol* 36, 289–290
71 Löfgren, A. *et al.* (2018) Designing an autonomous integrated downstream sequence from a batch separation process – an industrial case study. *Biotechnol. J.* 13, 1700691
72 Besenhard, M.O. *et al.* (2018) A multivariate process monitoring strategy and control concept for a small-scale fermenter in a PAT environment. *J. Intell. Manufact.* 29, 1501–1514
73 Dumarey, M. *et al.* (2018) Advances in continuous active pharmaceutical ingredient (API) manufacturing: real-time monitoring using multivariate tools. *J. Pharm. Innov.* 2018 . http://dx.doi.org/10.1007/s12247-018-9348-7 Published online October 25
74 Katz, L. *et al.* (2018) Synthetic biology advances and applications in the biotechnology industry: a perspective. *J. Indust. Microbiol. Biotechnol.* 45, 449–461
75 Tobyn, M. *et al.* (2018) The preeminence of multivariate data analysis as a statistical data analysis technique in pharmaceutical R&D and manufacturing. In *Multivariate Analysis in the Pharmaceutical Industry* (A. P. Ferreira, J. C. Menezes, M. Tobyn, eds), pp. 3–12, Elsevier.
76 Mondal, S. and Khatua, S. *et al.* (2019) Accelerating pairwise sequence alignment algorithm by mapreduce technique for next-generation sequencing (NGS) data analysis. In *Emerging Technologies in Data Mining and Information Security* (Abraham, A., ed.), pp. 213–220, Springer
77 López-Lucas, M.D. *et al.* (2018) Production via good manufacturing practice of exofucosylated human mesenchymal stromal cells for clinical applications. *Cytotherapy* 20, 1110–1123
78 Cornwall, P. *et al.* (2018) Route design, the foundation of successful chemical development. *Bioorg. Med. Chem.* 26, 4336–4347
79 Yee, W. *et al.* (2018) Inclusion of 2-mercaptoethanol in lysis buffer could interfere with isolation of high molecular weight DNA from freshwater microalgae. *Indian J. Microbiol.* 58, 109–113
80 Artac, M. *et al.* (2017) DevOps: introducing infrastructure-as-code. *In 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)* (R. Bilof, eds), pp. 497–498, IEEE.