Data science strategies leading to the development of data scientists’ skills in organizations

Maria José Sousa¹ • Pere Mercadé Melé² • António Miguel Pesqueira³ • Álvaro Rocha⁴ • Miguel Sousa⁵ • Salma Noor⁶

Received: 27 May 2020 / Accepted: 28 April 2021 / Published online: 16 May 2021
© Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

The purpose of this paper is to compare the strategies of companies with data science practices and methodologies and the data specificities/variables that can influence the definition of a data science strategy in pharma companies. The current paper is an empirical study, and the research approach consists of verifying against a set of statistical tests the differences between companies with a data science strategy and companies without a data science strategy. We have designed a specific questionnaire and applied it to a sample of 280 pharma companies. The main findings are based on the analysis of these variables: overwhelming volume, managing unstructured data, data quality, availability of data, access rights to data, data ownership issues, cost of data, lack of pre-processing facilities, lack of technology, shortage of talent/skills, privacy concerns and regulatory risks, security, and difficulties of data portability regarding companies with a data science strategy and companies without a data science strategy. The paper offers an in-depth comparative analysis between companies with or without a data science strategy, and the key limitation is regarding the literature review as a consequence of the novelty of the theme; there is a lack of scientific studies regarding this specific aspect of data science. In terms of the practical business implications, an organization with a data science strategy will have better direction and management practices as the decision-making process is based on accurate and valuable data, but it needs data scientists skills to fulfil those goals.

Keywords Data science • Pharma • Health sector • Big data • Skills • Data technostructure • Data management structure

¹ Business Research Unit, ISCTE-Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal
² Department of Statistics and Econometrics, University of Malaga, Campus el Ejido s/n, 29071 Malaga, Spain
³ Bavarian Nordic, Grafenauweg 8, 6301 Zürich, Switzerland
⁴ ISEG, Universidade de Lisboa, Rua Miguel Lupi, 1200 781 Lisboa, Portugal
⁵ Department of Economics, Essex University, Wivenhoe Park, Colchester CO4 3SQ, UK
⁶ Department of Computer Science, SBBWU, Peshawar, Pakistan
1 Introduction

Data science in the pharma industry is a new issue, and it can bring significant advantages for early adopters as the benefit of making decisions based on accuracy and data quality is a competitive advantage for companies.

The access to high-quality and large datasets combined with data science techniques [2, 32] will optimize business processes and the health sector. Data science can be a transformative driver of the health sector, namely the health tech industry (creating new data science applications to analyse and visualize in an optimal way the big data available for all the stakeholders of the health system), healthcare providers (as they are the main drivers for deploying better health services for citizens); pharmacists (adjusting themselves to the needs of healthcare providers and citizens in general, making decisions on new product research and development to improve the quality of life); and other stakeholders involved in all health processes, sharing big data [27].

Central concerns with any big data–relevant venture are the privacy of data and ethics, though the use of big data can effectively produce better outcomes and more tailored responses with improved quality of life. However, personal data are sensitive, and legal and ethical issues need to be considered when using data science to analyse this kind of data.

In the European Union, the definition of specific policies and technologies to enable data science as a base for big data analysis [21] will facilitate the creation of global value chains in the health sector (pharma, healthcare providers, citizens, and all other health sector stakeholders), and it will contribute to digital single-market strategy. The value created by data science [6–8] can help to transform the health sector to increase its quality, decrease costs, and improve accessibility for all citizens.

The novelty of this research from the perspective of neural computing and applications (NCAA), besides analysing the pharma industry regarding the data science strategy of pharma companies and taking it as a pilot example for other companies and policymakers, is based on the presupposition that data science methods are unquestionably fundamental to produce information for better decision-making processes. In this context, data scientists use neural network mechanisms to increase attention, learning, and decision-making, supported by applications that help them in those processes. This is reflected in the goals of this research, which identifies the need for data scientists to develop specific skills to support with more sustainability and accuracy the data science strategies and models implemented by organizations.

The paper is organized as follows. Section 2 provides the literature review of the relevant research based on data science methods. Section 3 presents the methodology. Section 4 presents the analysis and discussion, and Sect. 5 concludes the paper and presents the implications for future research.

2 The literature review on data science

Data science, as the name implies, is the science of understanding, analysing, and transforming data through statistical and computational methodologies to solve real-life problems. As suggested by Hayashi [16], the main purpose of data science is to use knowledge of well-established theory and methods in supporting sciences like statistics, data mining, and machine learning to resolve and unmask the hidden structures or features in data and provide solutions for complex natural or social phenomena.

In recent times, the term ‘big data’ have often appeared in relevance to data science. However, it has been observed that data science problems may not always require the use of big data. Alternatively, the use of big data in conjunction with data science methodologies may prove useful in solving some specific problems. Similarly, as observed in the recent literature, deep-learning methodologies are often applied instead of traditional machine learning technologies when dealing with big data in a data science project as their use with big data has shown better results than most of the traditional machine learning approaches.

Data science has attracted intense and growing attention from significant health and life sciences organizations [13], including big pharmaceutical companies that maintain a traditional data-oriented scientific and clinical development approach for various parts of the business and management processes, where data are not shared across different departments like market access or marketing.

Progressing digital transformation stimulates a considerable growth of digital data. Data are an asset for any business organization, and having the capacity to understand all the connected trends and patterns and extract meaningful information and knowledge from data is referred to as data science [12].

The field of data science technology is enriched by knowledge and techniques from statistics as well as computing. On one hand, data science refers to traditional statistics [9] that are produced for argumentation analysis or specific, methodical problems, with additional capacity for exploratory analysis and the integration of data crunching and data mining. On the other hands, data science technologies are also borrowed from the field of software development, which has a strong basis in traditional platforms like data warehouses. This field thus can

© Springer
aggregate and maintain various large datasets for data management and storage on distributed development platforms through distributed computation or integration software [33].

The discovery of adaptive solutions through big data using data science technologies for long-standing pharmaceutical processes is often labelled Pharma 4.0. As discussed by Steinwandter et al. [24], cyberphysical systems and dark factories, which are well-known concepts in Pharma 4.0, often require data science techniques and tools as core components to resolve long-standing pharmaceutical issues and have been found to be useful.

Deep learning and big data analyses have recently gained much attention in leading medical sciences and pharma research as observed by Tariq et al. [29]. This is characterised by the initiative of many medical and pharmaceutical organizations to gather large amounts of relevant data of their domain.

This recent research on the possible application of data science processes to solve long-standing pharmaceutical problems and the growing trend in the industry to utilize big data and deep-learning technologies to better understand various processes and make knowledgeable decisions are the motivations for our study. The adaptation of these technologies requires a better understanding of their advantages and disadvantages, and hence, this research is curtailed to identify and study the factors involved.

It is fundamental for the strategic decision-making process of a pharma organization to identify challenges, capitalize on opportunities, and predict future trends and behaviours of HCPs, KOLs, and other stakeholders (Grom, 2013). For this purpose, many data science techniques [32] can be applied, such as the ones grouped below:

- a. Descriptive statistical analysis, which is used to summarize data from a sample using parameters such as mean, median, standard deviation, variance, and others.
- b. Inferential statistical analysis, which is used to draw conclusions from data through hypotheses (null and alternative hypotheses).
- c. Predictive analytics, which uses predictive algorithms and machine learning techniques to define the probability of future results, behaviours, and patterns, based on existing data.
- d. Prescriptive analytics, which aims to find the optimal recommendations for a decision-making process.
- e. Causal analysis, which searches for data to understand the causes.
- f. Exploratory data analysis, which is an alternative to inferential statistics, and its emphasis is on detecting general trends and patterns in the data and tracking associations.
- g. Mechanistic analysis, which is used in big data analysis in industries.

The understanding and knowledge of these techniques are essential for data scientists as their tasks involve complex data analysis and require a high level of understanding of learning processes to be used accurately in organizations, as discussed by [5].

Data science is gaining middle ground in all pharma companies for the efficient utilization of resources: storage, time, and efficient decision-making to exploit new methods and procedures. The critical challenges are the management of exponentially growing data, its meaningful analysis, deploying low-cost processing tools and practices while minimizing potential risks relating to safety, inconsistency, redundancy, and privacy. In this context, various variables need to be considered by organizations to define a data science strategy [6–8].

The volume of data generated by sources such as patients, hospitals, physicians, suppliers, and others nowadays is considerable, and to obtain valuable information from such enormous and heterogeneous data requires a multimodal learning methodology to make the insights from such combined information available to decision-makers and policymakers.

In the pharma industry, there is a high percentage of unstructured, internal data [1] from liaisons with stakeholders and from products. Furthermore, the use of external data such as lifestyle information for research, development, and marketing is vital to gain useful information and define future strategies. The identification of new trends and research and the development of new products could also be facilitated through this insight into data [17]. An optimal analytical approach should, as much as possible, generate recognizable patterns to allow for cross-checking results and enabling trust in the solutions.

Data quality is a significant issue as decisions are taken based on data that is sensitive, and there is a great responsibility and expectation regarding data accuracy and the quality of analytics tools [28].

Availability of and access to data [1] should also enable expert-driven self-service analytics to allow experts to control the analytics process. However, data are excessive and ever growing. There are several repositories, and new data are being generated daily by billions of connected devices or self-generated by people. It is necessary to find more appropriate and effective ways to leverage this data according to privacy and ethical principles, to access it, to understand the purposes for its use [11], and to improve and optimize the quality of the processes.

Data portability involves the transfer of data in a structured, commonly used, and machine-readable format from one organization (controller) to another organization.
The increase in data captured through the Internet and the Internet of Things needs to be dealt with, keeping in mind the regulations of data ownership standards. There is a need to raise awareness and trigger debate for policymakers and develop data protection and privacy laws and legislation to protect patients and companies.

The literature also discusses the cost of data, not only the cost of online data storage (cloud computing) but also the cost of gathering data and analysing it and the cost of using data to create innovations and advance society and the quality of life of citizens.

Another issue discussed in the literature is the lack of pre-processing facilities in data mining processes. Data pre-processing may require techniques that involve transforming raw data into any required and machine-understandable format. Real-world data are often incomplete, inconsistent, and likely to contain errors. Data pre-processing prepares raw data for further processing and is used in database-driven applications such as customer relationship management and rule-based applications (like neural networks). This process is fundamental for the pharma industry to handle data gathered from diverse sources.

It is important to note here that when we work with big data, we are still missing specific technology, mainly applications that help to analyse the data and make it readable for all stakeholders.

When it comes to medical data, privacy and security are of primary concern as medical data are highly sensitive information; there are strong regulations at the national and European Union level. The preservation of privacy in the practical implementation of a data science strategy also requires specific analytical tools and cybersecurity systems.

Stakeholders are the actual beneficiaries of data science’s potential and its facilitation in making the decision process more efficient. However, there is still a market gap in the skills needed to treat raw data and transform it into knowledge to support research, development, and other functions of a pharma company.

Finally, it is essential to discuss the importance of regulatory risks regarding the protection and security of data to make sure individuals with dubious intentions do not access data.

From the above exploration of the literature, we have gathered the most relevant variables regarding pharmaceutical business data requirements; these will guide our empirical research and help us formulate the answer to the main research question, identified below:

**Research Question:** What are the main differences between pharma companies with and without a data science strategy, and what is data science’s relation to the shortage of skills?

### 3 Methodological approach

To answer the research question identified at the end of Sect. 2, we have applied an approach based on quantitative methods, supported by a questionnaire to identify differences in data science challenges and framework conditions among organizations with or without a data science strategy.

The information was collected via a structured questionnaire that was prepared after a review of the literature. A convenience sample was used (non-probabilistic sampling procedure). When it is difficult to obtain a complete sampling, convenience sampling is suitable. The fieldwork was carried out between April and June of 2019 with a group of 280 individuals. To provide greater representativeness of the data, we selected individuals from companies around the world for a confidence level of 95% (and $p = q = 0.5$) and an increase in data error for the estimate of the proportion of 5.8%. The next table shows a summary of the information regarding the data collection and technical matters of the sample (Table 1).

The sample includes pharma workers from all over the world, as shown in following Table 2:

### 4 Data analysis and discussion

To analyse the differences between the organizations that do not have a data science strategy, a covariance analysis (ANCOVA) was carried out. The Shapiro–Wilks test was carried out, and the null hypothesis of normality was rejected. Rheinheimer and Penfield said that F-Snedecor was robust in the face of violations of normality. Tabachnick and Fidell also say (p. 204) that the robustness of the variance analysis is assured with large samples, with approximately 20 degrees of liberation. Blanca et al. also support this conclusion. The multivariate design with covariates is intended to reduce the damage caused by other covariant variables, such as the psychological variable, where data science influences the creation of innovations. In this way, variance due to individual differences was estimated from the regression between the dependent variable and the covariable. The scores in the dependent variable were statistically adjusted to the covariable. Finally, an ANOVA was performed on these adjusted scores. Thus, the analysis controls the
effect of the covariable so that it eliminates the variation due to the mismatch of the ANOVA error.

To investigate technological differences, the following hypotheses were analysed.

\[ H_0: \bar{X}_j \text{Adjusted Data Science} = \bar{X}_j \text{Adjusted No Data Science} \]

\[ H_1: \text{No } H_0. \]

The adjusted mean is obtained from the following expression:

\[ Y_{j\text{adjusted}} = \frac{Y_j}{C_0} b (X_j - \bar{X}) \]

where: \( \bar{X}_j \text{ adjusted} \): mean of the dependent variable of the \( j \)th group, \( \bar{X}_j \): mean of the dependent variable without the adjustment of the \( j \)th group, \( b \): pending communal regression, \( \bar{X}_j \): covariable mean in the \( j \)th group, \( \bar{X} \): total mean (of all groups) in the covariable.

In following Table 3, the adjusted means for each group are analysed.

In Table 3 is displayed the adjusted means, \( F \)-statistics, and \( p \) values. The analysis shows that there are statistically significant differences in all variables related to challenges (\( p \) value < 0.01 in all cases), always showing a higher score in organizations with data science.

### Table 1  Fact sheet

| Fieldwork | April through June 2019 |
|-----------|-------------------------|
| Sample size | 280 surveyed |
| Sample type | Convenience and geographic quota sampling |
| Survey type | Structured online questionnaire |
| Geographical area | 118 Europe, 102 U.S., 69 Asia |
| Business activities in the EU | Yes: 60.7%; No: 39.3% |
| Sampling error | 5.8%, assuming \( p = q = 0.5 \) and a confidence level of 95% |

### Table 2  Region of the participants

| Region | Count—Which country is your organization from? |
|--------|-----------------------------------------------|
| APAC, EMEA, U.S | 13 |
| Belgium | 1 |
| EMEA | 15 |
| Europe | 1 |
| Europe and Asia | 35 |
| Switzerland | 116 |
| U.S | 37 |
| U.S. and EU | 6 |
| Worldwide | 56 |
| Total | 280 |

### Table 3  Adjusted means, \( F \)-statistics, and \( p \) values

| Variables | Data science | No data science | \( F \) | \( p \) |
|-----------|--------------|-----------------|-------|------|
| Overwhelming volume | 4.128 | 3.153 | 102.385 | 0.000*** |
| Managing unstructured data | 3.987 | 3.526 | 23.797 | 0.000*** |
| Data quality | 4.101 | 3.631 | 14.633 | 0.000*** |
| Availability of data | 4.367 | 3.614 | 14.633 | 0.000*** |
| Access rights to data | 4.358 | 3.500 | 86.401 | 0.000*** |
| Data ownership issues | 4.377 | 3.571 | 83.375 | 0.000*** |
| Cost of data | 4.278 | 3.568 | 51.542 | 0.000*** |
| Lack of pre-processing facilities | 4.315 | 3.613 | 45.606 | 0.000*** |
| Lack of technology | 4.119 | 3.861 | 6.377 | 0.012** |
| Shortage of talent/skills | 4.035 | 3.738 | 7.934 | 0.005*** |
| Privacy concerns and regulatory risks | 4.231 | 3.912 | 15.069 | 0.000*** |
| Security | 4.288 | 3.837 | 18.908 | 0.000*** |
| Difficulties of data portability | 4.445 | 3.994 | 23.106 | 0.000*** |

\* = \( p < 0.1 \); ** = \( p < 0.05 \); *** = \( p < 0.01 \)
Following this, an exploratory factorial analysis of the variables related to challenges was carried out to see which factors could be extracted from this analysis.

The extraction method is the main component analysis method, and varimax was used as the rotation procedure. The first factor obtained explained 63.965% of the total variance of the matrix of challenges, and this dimension has eight items and is classified as the data management dimension. The second factor extracted explained 10.908% of the total variance has five items and is called the data technostructure dimension. These two extracted factors explain 74.874% of the total variance (Table 4). The Kaiser–Meyer–Olkin index is 0.762 higher than 0.7, and the Bartlett test of sphericity is fulfilled by presenting a chi-square of 4633.704 (~p value < 0.000).

It was also important, based on the factors obtained, data management and data technostructure, to validate the scale using confirmatory factor analysis (Table 5).

### Table 4 Exploratory factor analysis

| Item | Skewness | Kurtosis | Components | Data dimensions |
|------|----------|----------|------------|----------------|
|      |          |          | 1          | 2              |
| Overwhelming volume | 0.739 | 0.147 | 0.841 | Data management |
| Managing unstructured data | 0.253 | 1.057 | 0.691 |
| Data quality | -0.048 | 1.625 | 0.763 |
| Availability of data | -0.501 | 0.370 | 0.912 |
| Access rights to data | -0.150 | 1.444 | 0.840 |
| Data ownership issues | -0.203 | 1.262 | 0.829 |
| Cost of data | -0.131 | 1.517 | 0.637 |
| Difficulties of data portability | -0.066 | 0.207 | 0.627 |
| Lack of pre-processing facilities | -0.174 | 1.645 | 0.728 | Data technostructure |
| Lack of technology | 0.552 | 0.173 | 0.892 |
| Shortage of talent/skills | 0.312 | 0.696 | 0.930 |
| Privacy concerns and regulatory risks | 0.582 | 0.778 | 0.688 |
| Security | 0.733 | 0.026 | 0.606 |

% variance explained 63.965% 10.908%

Kaiser–Meyer–Olkin index 0.762

Bartlett’s test of sphericity Chi-square = 4633.704; sig < 0.000

Extraction method: principal component analysis, varimax rotation method with Kaiser

### Table 5 Psychometric properties

| Factor | Items | Loads | Average loads | A | AVE | CRI |
|--------|-------|-------|---------------|---|-----|-----|
| Data management | DM1: Overwhelming volume | 0.794 | 0.837 | 0.945 | 0.706 | 0.950 |
| | DM2: Managing unstructured data | 0.814 |
| | DM3: Data quality | 0.88 |
| | DM4: Availability of data | 0.908 |
| | DM5: Access rights to data | 0.883 |
| | DM6: Data ownership issues | 0.888 |
| | DM7: Cost of data | 0.846 |
| | DM8: Difficulties of data portability | 0.686 |
| Data technostructure | DT1: Lack of pre-processing facilities | 0.834 | 0.810 | 0.902 | 0.666 | 0.908 |
| | DT2: Lack of technology | 0.9200 |
| | DT3: Shortage of talent/skills | 0.91 |
| | DT4: Privacy concerns and regulatory risks | 0.691 |
| | DT5: Security | 0.6960 |

Following this, an exploratory factorial analysis of the variables related to challenges was carried out to see which factors could be extracted from this analysis.

The extraction method is the main component analysis method, and varimax was used as the rotation procedure. The first factor obtained explained 63.965% of the total variance of the matrix of challenges, and this dimension has eight items and is classified as the data management dimension. The second factor extracted explained 10.908% of the total variance has five items and is called the data technostructure dimension. These two extracted factors explain 74.874% of the total variance (Table 4). The Kaiser–Meyer–Olkin index is 0.762 higher than 0.7, and the Bartlett test of sphericity is fulfilled by presenting a chi-square of 4633.704 (~p value < 0.000).

It was also important, based on the factors obtained, data management and data technostructure, to validate the scale using confirmatory factor analysis (Table 5).
The Cronbach’s alpha ($\alpha$) is higher than 0.7 [10], the composite reliability index (CRI) is higher than 0.7 [14], and the average variance extracted (AVE) is higher than 0.5 [14]. The measures of validity are also adequate, the coefficients of standardized loadings are higher than 0.5, and their means are higher than 0.7 (Hair et al. [15]). Moreover, the confidence interval of the correlations is less than 1 [3] (Table 6).

Therefore, we have validated the scale for data management with eight indicators and the scale for data technostructure with five indicators, the lack of technology and the shortage of talent/skills having the highest loads.

Next (Table 7), we analysed globally the differences that may exist between data management and data technostructure between organizations with a data science strategy.

In Table 7, it is possible to verify that the organizations with a data science strategy have the best performance in both management and technostructure. Therefore, in terms of practical business implications, an organization with a data science strategy will have better direction and management practices as the decision-making process is based on accurate and valuable data, based on the data science skills of the workers.

Once the measurement model was validated, it was essential to analyse, in an exploratory way, if there is a relationship between data technostructure and data management regarding if the organization has a data science strategy. The next table shows the standardized coefficients used to analyse this structural relationship [14], 1982, Table 8). For the analysis of structural equation models, they were analysed through the maximum likelihood with the Satorra–Bentler correction, which is robust with non-normal data [25, 26].

| Table 6 Discriminant validity assessment: AVE and square correlations between constructs |
|---------------------------------------------------------------|
| Factor | Data management | Data technostructure |
|--------|-----------------|----------------------|
| Data management | 0.706 | 0.504 |
| Data technostructure | (0.666–0.754) | 0.666 |

The main diagonal represents AVE (in bold). Square correlations are reported above the main diagonal. Confidence intervals ($\alpha = 0.05$) for correlations are reported below the main diagonal.

| Table 7 ANCOVA. data management and data technostructure. Adjusted means, $F$ statistics, and $P$ values |
|---------------------------------------------------------------|
| Dimensions | Data science | No data science | $F$ | $p$ |
|------------|--------------|-----------------|-----|-----|
| Data management | 4.255 | 3.570 | 71.016 | 0.000*** |
| Data technostructure | 4.197 | 3.792 | 24.735 | 0.000*** |

$* = p < 0.1; ** = p < 0.05; *** = p < 0.01$

*Table 6 Discriminant validity assessment: AVE and square correlations between constructs

Table 7 ANCOVA. data management and data technostructure. Adjusted means, $F$ statistics, and $P$ values

Based on this analysis, it is possible to conclude that there is empirical evidence of the existence of a robust positive relationship between data technostructure and data management in organizations that have data science. However, the relationship is not statistically significant if the organization does not have a data science strategy (Fig. 1), which is justified by the fact that they are not focussed on that type of strategy and that they have not reached the maturity to understand the importance of having data science skills for data analysis.

5 Conclusions

This study presents an analysis of the application of data science to the pharma industry. Data science is a new interdisciplinary science that requires strong practical ability and an adaptive organizational culture to effectively implement the described techniques and models to support the pharma industry in daily activities. The review is conducted along two major dimensions, data technostructure and data management, as these are the two main components of a data science strategy. From the survey results, we can conclude that companies essentially need to empower their data technostructure for better results. It was also observed that most of the companies had a higher and increasing interest in data management. Many pharma...
companies are in the process of realizing the importance of the skills needed for a data scientist and implementing data science in their analytics processes. According to the study, there is empirical evidence about the relationship between data technostructure and data management as they need to be defined and managed at nuclear dimensions for the competitiveness of the pharma industry. For future work, we intend to execute a survey with medical affairs practitioners and compare the collected data with our results in this study. It would be interesting to disaggregate the concepts of data management and data technostructure further and perform a deeper analysis.

**Funding** This study was funded by Fundação para a Ciência e Tecnologia, Grant: UIDB/00315/2020.

**Declaration**

**Conflict of interest** The authors declare that there is no conflict of interest.

**References**

1. Adam NR, Wieder R, Ghosh D (2017) Data science, learning, and applications to biomedical and health sciences. Ann N Y Acad Sci 1387(1):5–11
2. Akerkar R, Sajja PS (2016) Intelligent techniques for data science, 1st edn. Springer, Switzerland
3. Anderson JC, Gerbing DW (1988) Structural equation modeling in practice: a review and recommended two-step approach. Psych Bull 103(3):411
4. Blanca MJ, Alarcón R, Arnau J, Bono R, Bendayan R (2017) Non-normal data: is ANOVA still a valid option? Psicothema 29(4):552–557
5. Brownson RC, Colditz GA, Proctor EK (2017) Dissemination and implementation research in health: translating science to practice. Oxford University Press, Oxford
6. Cao L (2017) Data science: a comprehensive overview. ACM Comput Surv (CSUR) 50(3):43
7. Cao L (2017) Data science: challenges and directions. Commun ACM 60(8):59–68
8. Cao L (2016) Data science: nature and pitfalls. IEEE Intell Syst 31(5):66–75
9. Cleveland WS (2001) Data science: an action plan for expanding the technical areas of the field of statistics. Int Stat Rev 69(1):21–26
10. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334
11. Cruz-Correa R, Ferreira D, Bacelar G et al (2018) Personalised medicine challenges: quality of data. Int J Data Sci Anal 6:251. https://doi.org/10.1007/s41060-018-0127-9
12. Dinov ID (2019) Quant data science meets dexterous artistry. Int J Data Sci Anal 7:81
13. Dinov ID (2016) Volume and value of big healthcare data. J Med Stat Inf 4(1):1–7
14. Fornell C, Larcker DF (1981) Structural equation models with unobservable variables and measurement error. J Mark Res 18(1):39–50
15. Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate Data Analysis. Seventh Edition. Prentice Hall, Upper Saddle River, New Jersey
16. Hayashi C (1998) What is data science? Fundamental concepts and a heuristic example. In: Data science, classification, and related methods 1998. Springer, Tokyo, pp 40–51
17. Jain S (2017) Bridging the Gap Between R&D and commercialization in the pharmaceutical industry: role of medical affairs and medical communications. Int J Biomed Sci 3(3):44–49
18. Malley B, Ramazzotti D, Wu JT (2016) Data Pre-processing. In: Secondary Analysis of Electronic Health Records. Springer, Cham. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543629/. https://doi.org/10.1007/978-3-319-43742-2_12
19. Mercadé-Mélè P, Molinillo S, Fernández-Morales A (2017) The influence of the types of media on the formation of perceived CSR. Span J Market-ESIC 21:54–64
20. Mercadé-Mélè P, Molinillo S, Fernández-Morales A, Porcu L (2018) CSR activities and consumer loyalty: the effect of the type of publicizing medium. J Bus Econ Manag 19(3):431–455
21. Radermacher WJ (2018) Official statistics in the era of big data opportunities and threats. Int J Data Sci Anal 6:225. https://doi.org/10.1007/s41060-018-0124-z
22. Rheinheimer DC, Penfield DA (2001) The effects of type I error rate and power of the ANCOVA F test and selected alternatives under nonnormality and variance heterogeneity. J Exp Educ 69(4):373–391
23. Salas J, Domingo-Ferrer J (2018) Some Basics on privacy techniques, anonymization, and their big data challenges. J Math Comput Sci 12:263. https://doi.org/10.1007/s11786-018-0344-6
24. Steinwandter V, Borchert D, Herwig C (2019) Data science tools and applications on the way to Pharma 4.0. Drug Discov Today. 24(9):1795–1805
25. Satorra A, Bentler PM (1988) Scaling corrections for chi-square statistics in covariance structure analysis. In: Procedings of the American Statistical Association
26. Satorra A, Bentler PM (1994) Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye A, Clogg CC (eds) Latent variables analysis
27. Schneeweiss S (2014) Learning from big health care data. N Engl J Med 370(23):2161–2163
28. Skiena SS (2017) The data science design manual. Springer, Cham
29. Tariq MI, Memon NA, Ahmed S, Tayyaba S, Mushtaq MT, Mian NBK543629/, https://doi.org/10.1007/978-3-319-43742-2_12
30. Tabachnick BG, Fidell LS, Ullman JB (2007) Using multivariate statistics. Sixth Edition. Pearson, Boston
31. Torra V, Navarro-Aribas G (2016) Big data privacy, and anonymization. In: Lehmann A, Whitehouse D, Fischer-Hübner S, Fritsch L, Raab C (eds) Privacy and identity management. facing up to next steps. Privacy and Identity, 2016. IFIP Advances in Information and Communication Technology, vol 498. Springer, Cham
32. Wenwu H, Guomai L (2017) Exploration and research on the core course construction of data science and big data technology specialty, education review (2017)
33. Wilkinson MD et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018–1–160018–9
34. Wohlfarth M (2019) Data portability on the internet. Bus Inf Syst Eng 61:551

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.