Data Science Strategies, Structures and Management Models, in relation to Shortage of Data Scientists Skills in Organizations

Abstract

The purpose of this paper is to analyze the differences between companies who have a data science strategy and the data specificities/variable that can influence the definition of a data science strategy in pharma companies.

The current paper is empirical and the research approach consists of verifying with a set of statistical tests the differences between companies with a Data Science Strategy and companies without Data Science Strategy. It was designed a specific questionnaire and applied 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 Data Science strategy and companies Without a Data Science strategy.

The paper offers a depth analysis between companies with or without 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 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. Introduction

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

The access to high-quality and large datasets combined with data science techniques (Wenwu and Guomai, 2017; Akerkar and Sajja, 2016) will optimize the processes of the companies and the health sector. Data science can be a driver for the transformation of the health sector, namely the Healthtech industry (creating new data science applications in order to analyse and visualize in an optimal way the big data available for all the stakeholders of the health system); the healthcare providers (as they are the main driver for deploying better health services for the citizens); the pharma (adjusting their self's to the needs of the healthcare providers and the citizens in general, making decisions on new products research and development to improve the quality of life); and others stakeholders involved in all the health processes, sharing big data (Schneeweiss, 2014).

A central concern is the privacy of data and ethics, as big data can effectively conduct to better outcomes and more tailored responses (Yao, Zhu, and Cui, 2018) with improved quality of life. However, personal data is sensitive, and legal and ethical issues need to be considered when using data science to analyze this kind of data.

At European Union the definition of specific policies and technologies to enable data science as a base for big data analysis (Radermacher, 2018) will facilitate the creation of global value chains in the health sector (pharma, healthcare providers, citizens, and all the other health sector stakeholders) and it will contribute towards the de digital single market strategy. The value created by Data Science (Cao, 2016, 2017a, 2017b) can help to transform the health sector to increase its quality, decrease costs, and improve accessibility for all citizens.

In this context, this article aims to analyze the pharma industry regarding the data science strategy of pharma companies and to take it as an example or a pilot-example for other companies and also for policymakers.

2. Literature Review on Data Science

Data Science has attracted intense and growing attention from significant health and life sciences organizations (Dinov, 2016), including the big pharmaceutical companies that maintain a traditional data-oriented scientific and clinical development fields, as very far parts of the business and management structures., where data is not shared across different departments like market access or marketing.

The progressing digital transformation stimulates a considerable growth of digital data. Data is an asset for any business organization, and having the capacity to understand all the connected trends, patterns, and extract

meaningful information and knowledge from the data is referred to as data science (Dinov, 2019).

The topics of data science technologies encompass two different aspects. Data science refers to traditional statistics (Cleveland, 2001) that are produced on argumentation analysis or specific, methodical problems, with additional capacity for exploratory analysis and integration of data crunching and data mining. In another hand, data science technologies also are resulted from traditional software development that has a strong basis on traditional platforms like data warehouses, having the main capacity to aggregate several quantities of data managed and stored on distributed development platforms that integrate into distributed computation or integrated software (Wilkinson, 2016).

It is fundamental for the strategic decision-making process (Ziying and Letian, 2018) of a pharma organization to identify challenges, capitalize on opportunities, and to predict future trends and behaviors of HCPs, KOLs, and others stakeholders (Grom, 2013). For this purpose many data science techniques (Wenwu and Guomai, 2017), can be applied and grouped in: a) descriptive statistical analysis which is used to summarize data from a sample using test as Mean; Median, Standard Deviation, Variance and others; b) Inferential Statistical Analysis which are used to make conclusions from data through hypothesis (null and alternative hypothesis); 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 search for data to understand the causes; f) Exploratory Data Analysis (EDA) which is an alternative to inferential statistics, and the emphasis is on detect general trends and patterns in the data and to track associations; g) Mechanistic Analysis which is used in the big data analysis in industries.

This techniques are framed by data scientist skills, as they involve complex data analysis, and need high level of learning processes to be used in an accurate way in organizations, as Brownson (2017) as showed in his studies.

Data Science is gaining middle ground in all pharma companies for the efficient utilization of resources: storage and time and efficient decision making to exploit new methods and procedures. Moreover, the critical challenges are the management of the exponentially growing data, its meaningful analysis, deploying low-cost processing tools and practices while minimizing the potential risks relating to safety, inconsistency, redundancy, and privacy. In this context, some variables need to be considered by organizations in order to define a data science strategy (Cao, 2016, 2017a, 2017b):

The volume of data (Martin-Sanchez and Verspoor, 2014) is big and to obtain valuable information from such enormous and heterogeneous sources (patients, hospitals, physicians, suppliers, and others), data need to be treated using 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 (Adam, Wieder, and Ghosh, 2017), from the liaisons with the stakeholders and also from the products. Furthermore, the use of external data such as lifestyle information, for research and development and also for marketing is vital to gain knowledge from that information and define the strategies and the new trends on research and development of new products (Jain, 2017). An optimal analytical approach should, as much as possible, generate recognizable patterns in order to allow for cross-checking results and enabling trust in the solutions.

The data quality is a significant issue as the decisions are taken based on data that is sensitive, and there are a high responsibility and expectations regarding data accuracy and the quality of analytics tools (Skiena, 2017).

Availability and access of data (Adam, Wieder, and Ghosh, 2017) it should also enable expert-driven self-service analytics to allow the experts to control the analytics process. There are also several repositories and new data generated daily by billions of connected devices or self-generated by people. It is necessary to find more appropriate and effective ways to leverage these data in line with privacy and ethical principles, to access it, to understand the purposes for its use and quality (Cruz-Correia, Ferreira, Bacelar, et al., 2018) in order to improve and optimize the processes.

Data portability encloses the right to transfer data in a structured, commonly used, and machine-readable format from one organization (controller) to another organization (Wohlfarth, 2019).

The increasing data captured through the Internet and the Internet of Things needs analysis about the data ownership (Torra and Navarro-Arribas, 2016), it is needed to raise awareness and trigger debate for policymakers and develop data protection and privacy laws and legislation to protect patients and companies.

On another perspective the literature discusses the cost of data, and it is not only the costs with online data storage (cloud computing), but also the costs to gather the data, to analyze it, and to use the data to create innovations (Dinov, 2019), and to evolve the society and the quality of life of the citizens.

Moreover, there is also a lack of pre-processing (Malley, Ramazzotti, and Wu, 2016) facilities as data mining processes, which is a technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and is likely to contain many errors. Data pre-processing prepare 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 as a way to treat data from all a diversity of sources.

In this line, we verify that when we discuss big data is still missing specific technology, mainly apps that help to analyze the data (Radermacher, 2018) and make it readable for all the stakeholders.

Privacy and security are primary concerns (Salas and Domingo-Ferrer, 2018) as medical data is highly sensitive information, besides there are strong regulations at national and at European Union level. The privacy-preserving for the practical implementation of a data science strategy also requires specific analytical tools and cybersecurity systems.

The stakeholders are the crucial element to use data science potential and to make the decision process more efficient, nevertheless there is still a shortage of talent and skills to treat raw data and make it in knowledge for research and development (Brownson, 2017), but also for all the other functions of a pharma company.

Finally, it is essential to discuss the importance of regulatory risks regarding the protection and security of data (Salas and Domingo-Ferrer, 2018) to make sure individuals with dubious intentions do not access data.

From the literature emerged all of those variables, that will guide the empirical research and the answer to the main research question:

RQ: What are the main differences between a pharma company with and without a data science strategy and its relation to the shortage of skills?

3. Methodological Approach

In order to answer the research question, it was applied a quantitative methodologic approach 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 (Mercadé et al., 2017,2018). The fieldwork was carried out between April and June of 2019 with a participation of 280 individuals. In order to provide greater representativeness of the data, we have 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 the technical matters of the sample (Table 1).

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 US; 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%

4. Data Analyses and Discussion

To analyze the differences between the organizations that do not have a strategy on Data Science, a covariance analysis (ANCOVA) has been carried out. Data were checked for normality using the Kolmogorov-Smirnov test, and from the multivariate design with covariates, it is intended to reduce the damage caused by other covariant variables, such as the psychological variable of which data science influences the creation of innovations. In this way, the variance due to the individual differences is estimated from the regression between the dependent variable and the covariable. The scores in the dependent variable are statistically adjusted to the covariable. Finally, an ANOVA is performed on these adjusted scores (Tabachnick and Fidell, 2007). 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 have been analyzed.

H_0: $\bar{Y}_{j\ \text{Ajusted Data Science}} = \bar{Y}_{j\ \text{Ajusted No Data Science}}$

H_1: No H_0

The adjusted mean is obtained from the following expression:

$\bar{Y}_j$ adjusted = $\bar{Y}_j$ – b*($\bar{X}_{j}$ - $\bar{X}_{..}$)
Where :

$\bar{Y}_j$ adjusted: mean of the dependent variable of the jth group
$\bar{Y}_j$: mean of the dependent variable without the adjustment of the jth group
b: pending communal regression

$\bar{X}_{j:}$ covariable mean in the jth group

$\bar{X}_{..:}$ total mean (of all groups) in the covariable

In the following table is analyzed the adjusted means for each group.

Table 2. Adjusted means, F statistics and p value

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	62,34	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

In Table 2 are 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.

Following, an exploratory factorial analysis of the variables related to challenges has been carried out to see the factors that are extracted from this analysis. 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 and has five items and is called as the Data Technostructure dimension. The two extracted factors explain 74.874% of the total variance (Table 3).

Table 3. Exploratory Factor Analysis

	Components		
ITEM	1	2	Data dimensions
Overwhelming volume	0.841		Data Management
Managing unstructured data	0.691		
Data quality	0.763		
Availability of data	0.912		
Access rights to data	0.840		
Data ownership issues	0.829		
Cost of data	0.637		
Difficulties of data portability	0.627		
Lack of pre-processing facilities		0.728	Data Technostructure
Lack of technology		0.892	
Shortage of talent/skills		0.930	
Privacy concerns and regulatory risks		0.688	
Security		0.606	
% variance explained	63.965%	10.908 %	
Kaiser-Meyer-Oklin index	0.762		
Bartlett's test of sphericity	Chi-square = 4633,704; sig <0,000		
Notes: Extraction method: principal component analysis, varimax rotation method with Kaiser			

It was also important, based on the factors obtained, “Data Management” and “Data Technostructure," to validate the scale using confirmatory factorial análisis (Table 4).

Table 4: Psychometric properties.

Factor	Items	Loads	Average Loads	α	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	DT1: Lack of pre-processing facilities	0,834	0,810	0,90	0,66	0,90

Tecnoestructure	DT2: Lack of technology	0,9200		2	6	8
	DT3: Shortage of talent/skills	0,91				
	DT4: Privacy concerns and regulatory risks	0,691				
	DT5: Security	0,6960				

The Cronbach's Alpha (α) is higher than 0.7 (Cronbach, 1951), the Composite Reliability Index (CRI) is higher than 0.7 (Fornell and Larcker, 1981), and the Average Variance Extracted (AVE) is higher than 0.5 (Fornell and Larcker, 1981). 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., 2010). Moreover, the confidence interval of the correlations is less than 1 (Anderson and Gerbing, 1988) (Table 5).

Table 5: 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

Note: The main diagonal represents Average Variance Extracted –AVE (in bold). Square correlations are reported above the main diagonal. Confidence intervals (α = 0.05) for correlations are reported below the main diagonal

Therefore, we have validated the scale for Data Management with 8 indicators and the scale for Data Technostructure with 5 indicators, being the lack of technology, and the shortage of talent/skills, the with the most high loads.

Next, we intend to analyze globally the differences that may exist between Data Management and Data Technostructure between the organizations with a Data Science strategy.

Table 6. ANCOVA. Data Management and Data Technostructure. Adjusted means, F statistics, and p-value

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

On the table 6, 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 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, based on the data scientist skills of the workers.

Once the measurement model has been validated, it is essential to analyze, in an exploratory way, if there is a relationship between Data Technostructure and Data Management regarding if the organization has a strategy on Data Science. Next table shows the standardized coefficients to analyze this structural relationship (Fornell and Larcker, 1981, 1982) (table 7):

.

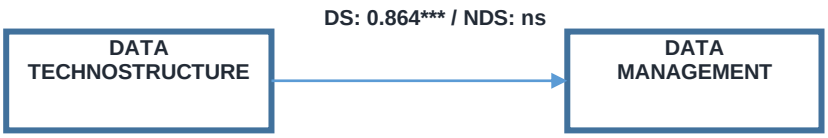
Table 7: Evaluation of the structural relationship

	Data Science		No Data Science	
Structural relationship	Coef.	Valor t*	Coef.	Valor t*
DT → DM	0.86435	38.31***	0.1103	0,97

*=p<0,1; **=p<0,05;
***=p<0,01

Base don 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 (Figure 1), which is justified by the fact that they are not focus on that type of strategy, and that they have not reached the maturity to understand the importance of having data scientists skills for data analysis.

Figure 1: Structural relationship



Note: DS: DATA SCIENCE / NDS: NO DATA SCIENCE; ns: not significant; ***: p-value < 0.01

5. Conclusions

This study presents an analysis of the application of data science to the pharma industry. Data science is a new interdisciplinary specialty, which requires strong practical ability and adaptive organizational culture to effectively implement the described techniques and models to support the pharma industry in daily activities. Review emerges then two dimensions a) Data Technostructure; and b)Data Management, which are the main pillars of a data science strategy. From the survey, results can be concluded that companies will consider essential to empower Data Technostructure and that they have a higher and increasing interest in Data Management, and they also assumed the importance of the skills needed for a data scientist and to implement data science in their analytics processes. According to the study, there is empirical evidence between the relationship of Data Technostructure with Data Management, as they need to be defined and managed as the 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 convenient to disaggregate the concepts of Data Management and Data Technostructure further and perform a causal analysis.

References

Adam, N.R., Wieder, R., Ghosh, D.: Data science, learning, and applications to biomedical and health sciences. Ann. N. Y. Acad. Sci. **1387**(1), 5–11 (2017)

Akerkar, R., Sajja, P.S.: Intelligent Techniques for Data Science, 1st ed. Springer, Switzerland (2016).

Anderson, J. C., & Gerbing, D. W. Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin. 1988;103(3), 411.

Brownson, R.C., Colditz, G.A., Proctor, E.K.: Dissemination and Implementation Research in Health: Translating Science to Practice. Oxford University Press, Oxford (2017)

Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (CSUR) **50**(3), 43 (2017a)

Cao, L.: Data science: challenges and directions. Commun. ACM **60**(8), 59–68 (2017b)

Cao, L.: Data science: nature and pitfalls. IEEE Intell. Syst. **31**(5), 66–75 (2016)

Cleveland, W.S. Data science: an action plan for expanding the technical areas of the field of statistics. Int. Stat. Rev. 69(1), 21–26 (2001)

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16(3), 297–334.

Cruz-Correia, R., Ferreira, D., Bacelar, G. et al. Personalised medicine challenges: quality of data. Int J Data Sci Anal (2018) 6: 251. https://doi.org/10.1007/s41060-018-0127-9

Dinov, I.D. Quant data science meets dexterous artistry. Int J Data Sci Anal (2019) 7: 81

Dinov, I.D.: Volume and value of big healthcare data. J. Med. Stat. Inf. **4**(1), 1–7 (2016)

Fornell, C., & Bookstein, F. L. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. Journal of Marketing Research. 1982; 19, 440–452.

Fornell, C., & Larcker, D. F. Structural equation models with unobservable variables and measurement error. Journal of Marketing Research. 1981; 18(1), 39–50.

Jain S. Bridging the Gap Between R&D and commercialization in the pharmaceutical industry: role of medical affairs and medical communications. Int J Biomed Sci. 2017;3(3):44–49.

Malley B., Ramazzotti D., Wu J.T. Data Pre-processing. In: Secondary Analysis of Electronic Health Records. (2016) Springer, Cham

Martin-Sanchez, F., Verspoor, K.: Big data in medicine is driving big changes. Yearb. Med. Inform. (2014).

Mercadé-Melé, P., Molinillo, S., & Fernández-Morales, A. The influence of the types of media on the formation of perceived CSR. Spanish Journal of Marketing-ESIC. (2017), 21, 54-64.

Mercadé-Melé, P., Molinillo, S., Fernández-Morales, A., & Porcu, L. CSR activities and consumer loyalty: The effect of the type of publicizing medium. Journal of Business Economics and Management. (2018),19(3), 431-455.

Radermacher, W.J. Official statistics in the era of big data opportunities and threats. Int J Data Sci Anal (2018) 6: 225. https://doi.org/10.1007/s41060-018-0124-z

Salas, J. & Domingo-Ferrer, Some Basics on Privacy Techniques, Anonymization, and their Big Data Challenges J. Math.Comput.Sci. (2018) 12: 263. https://doi.org/10.1007/s11786-018-0344-6

Schneeweiss, S.: Learning from big health care data. N. Engl. J. Med. 370(23), 2161–2163 (2014)

Skiena, S.S.: The Data Science Design Manual. Springer, Cham (2017)

Torra V., Navarro-Arribas, G. 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.. (2016) Springer, Cham

Wenwu He, Guomai Liu, Exploration and Research on the Core Course Construction of Data Science and Big Data Technology Specialty, Education Review (2017).

Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**,160018-1–160018-9 (2016)

Wohlfarth, M. Data Portability on the Internet. Bus Inf Syst Eng (2019) 61: 551.

Yao, L.; Zhu, L.; Cui, C. Exploration of Data Science Course Construction and Personnel Training in Big Data Era, Computer Generation (2018).

Ziying Wang, Letian Gao, The Scientific Characteristics of Big Data in Computer Age and Its Decision-making Significance. Decision and Information (2018).