Data Science Projects in Pharmaceutical Industry

António Miguel Pesqueira, Alexion Pharmaceuticals, Switzerland
antonio.pesqueira@live.com

Maria José Sousa, ISCTE - Instituto Universitário de Lisboa, Portugal
maria.jose.sousa@iscte-iul.pt

Pere Mercadé Melé, University of Málaga, Spain
pmercade@uma.es

Álvaro Rocha, University of Coimbra, Portugal
amrrocha@gmail.com

Miguel Sousa, University of Essex, UK
miguel.ac.sousa@gmail.com

Renato Lopes da Costa, ISCTE - Instituto Universitário de Lisboa, Portugal
renatojlc@gmail.com

Abstract. The purpose of this paper is to discuss the relevance of data science in Medical Affairs (MA) functions in the pharmaceutical industry, where data is becoming more important for the execution of activities and strategic planning in the health industry.

Tis study analyses pharmaceutical companies who have a data science strategy and the variables that can influence the definition of a data science strategy in pharma companies in opposite to other pharmaceutical companies without a data science strategy. The current paper is empirical and the research approach consists of verifying the characteristics and differences between those two types of pharmaceutical companies. A questionnaire specifically for this research was developed and applied to a sample of 280 pharma companies. The development and analysis of the questionnaire was based on a Systematic Literature Review of studies published up to (and including) 2017 through a database search and backward and forward snowballing. In total, we evaluated 2247 papers, of which 11 included specific data science methodologies criteria used in medical affairs departments. It was also made a quantitative analysis based on data from a questionnaire applied to a Pharma organization. The findings indicate that there is good evidence in the empirical relation between Data Science and the strategies of the organization.

Keywords: Data Science, Pharmaceutical, Medical Affairs, Literature Review, Projects

1. Introduction and Conceptualization

Pharmaceutical Industry has an enormous influence of the regulators, healthcare professionals (HCPs), and patients, which has lead to the emergence of the importance of strategic functions like Medical Affairs (MA).

This function primary roles included, in the past years, a scientific exchange, information support, managing daily regulatory reporting requirements or driving medical evidence generation (for example, phase IV studies, real world evidences or collaborative research), with a strong focus on priority diseases and developed products (Dyer 2011). In the current times the MA assumed a central role of all pharma operations, operating independently from sales pressures, and assuming the responsibility to creating strategic relationships with healthcare professionals (HCPs), Key Opinion Leaders (KOLs), and other stakeholders (for example, regulators, investigators, and others) (Plantevin et al. 2017). Also, the MA function is growing the importance and capacity in the technology processes improvements, technology adoption, and primarily being the focus with improving HCPs, KOLs, and stakeholder engagement activities.

The conceptualization of this research is based on one main concept - data science – which has assumed a significant role in healthcare and life sciences organizations, including the large 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. Data regarding to this research is considered an asset for pharmaceutical organizations, and Data Science allows the pharma and health professionals to identify trends, patterns, and extract meaningful information and knowledge from the data, using a considerable set of methods and techniques, which will be discussed later on this paper.

Moreover, the digital transformation global process is stimulating the growth of digital data, encompassing two different aspects: a) the traditional statistics that are produced on argumentation analysis or specific, methodical problems, with additional capacity for exploratory analysis and integration of data crunching and data mining; b) data science technologies, which sometimes results from traditional software development with strong basis on traditional platforms like data warehouses, but having the capacity to aggregate large quantities of data to be managed, and stored on distributed development platforms, being later integrated into distributed computation or integrated software.

The critical challenges for MA are the management of big data, its meaningful analysis, deploying low-cost processing tools and practices while minimizing the potential risks relating to safety, inconsistency, redundancy, and privacy, and Data Science can be the answer to this challenges with an efficient utilization of resources: storage and time and efficient decision making to exploit new methods and procedures.

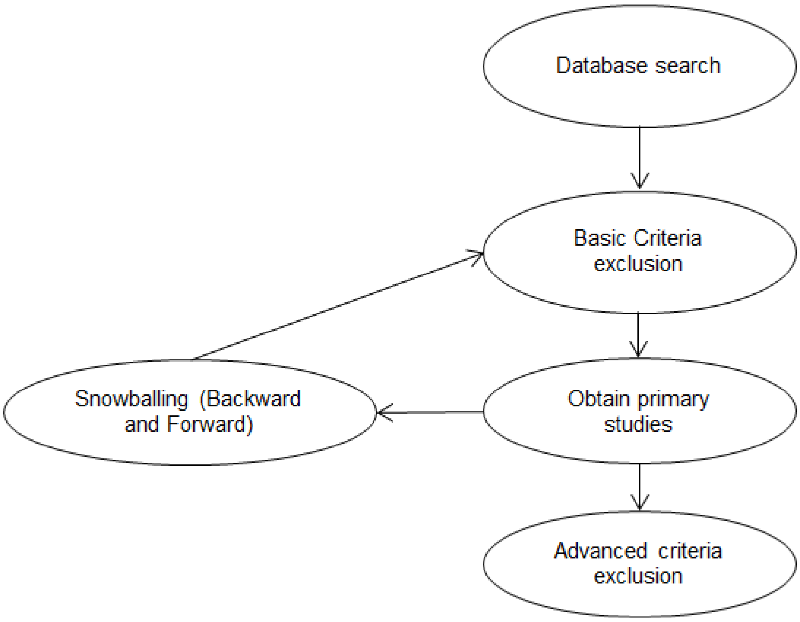
A pharmaceutical industry model includes two main pillars: an R&D function being responsible for developing new medicines/molecules and a commercial team in charge of marketing and selling those products during a post-clinical phase and after all clinical development and trials are completed. MA serves as a connecting bridge between R&D and Sales/Marketing, facilitating the transition of products and knowledge from R&D to the market access and commercialization stages. Despite many changes over the last years, all the stakeholders continued to demand high levels of scientific knowledge, and to have better interactions in terms of transparency and information sharing with the industry in its interactions. Also, here, the role and importance of MA in a more complex healthcare marketplace environment are increasing exponentially (Jain, 2017). Moreover, it is fundamental for the strategic decision-making process of a pharmaceutical organization to identify challenges, capitalize on opportunities, and to predict future trends and behaviours of HCPs, KOLs, and other stakeholders (Grom, 2013).

In this context, the main innovation of this study is related to the importance that big data is assuming in the pharmaceuticals industry, and the lack of studies exploring which criteria is being used in the context of Data Science applied to MA. There is a need for research to assess which Data Science practices proposed by practitioners and academics are the most relevant. In this context, the contribution of this research to theory and practice is to identify Data Science methods and techniques throughout a systematic literature review.

2. Systematic Literature Review Methodological Approach

According to Kitchenham and Charters (2007), a systematic review is an evidence-based technique that uses a well-structured and repeatable methodology to identify, analyze and interpret all the relevant academic papers related to a specific research question or phenomenon of interest. A fundamental assumption of this technique is the involved protocol, which is the plan that will describe the conduct of the systematic review. It includes the research questions, search process, selection process, and data analysis procedures (Figure 1).

Figure 1: Overview of the search and selection process


Database search
Basic Criteria exclusion
Obtain primary studies
Snowballing (Backward and Forward)
Advanced criteria exclusion

This research aims to identify possible gaps in the literature and define a starting point to define Data Science for Medical Affairs practitioners, employees or representatives, through the identification and synthesis of the Data Science criteria used in Medical Affairs projects as presented in the scientific literature. To minimize the probability of missing relevant articles, publications, it was used a combined search strategy, which is based on database search and snowballing (backward and forward). First, was defined a search string used to search databases containing scientific papers in the context of data science. After applying the essential criteria exclusion, the resulting papers were defined as the starting set for the snowballing process. After executing the snowballing iterations, it was applied the advanced criteria exclusion, which is related to the actual data extraction and quality assessment. Figure 1 shows an overview of the search and selection process. This strategy was used to avoid missing publications, papers, or articles due to limitations and inconsistencies of digital libraries. They have different formats to handle the Boolean expressions, as discussed in Brereton et al. (2007), and we were not sure how reliable is their ability to handle searches with long strings. Finally, there is evidence in the literature on the risks of missing papers using only one approach (Badampudi et al. 2015).

The search terms were based on the research questions using synonyms and related terms. The following keywords were used to formulate the search string: a) Population: Data Science and Medical Affairs. Alternative keywords: Medical Data Science, Data Science in Medical Affairs, Medical Affairs Data, and Data Science in Pharmaceutical; b) Intervention: Data Science. Alternative keywords: data science in medical affairs and medical data; c) Context: Industry or academia. Our target population was papers performed in the industry or academy, and we intended to capture papers in that context regardless of the type of research performed.

To define a first version of the search string, the keywords within a category were joined by using the Boolean operator 'OR,' and the two categories were joined using the Boolean operator 'AND.' This was done to target only papers in the context of data science related to medical affairs. To simplify the strings and include additional synonyms, we defined the following search string:
("data science" OR "medical affairs" OR "medical" OR "data" OR "medical data science" AND (medical AND (data OR science) AND (data science OR science OR (medical AND (affairs OR data clinical OR medical science OR data affairs)) OR "data science in medical affairs"

Regarding the data sources, the goal was to cover the literature published about Data Science, so the following digital databases were selected for data retrieval: a) ACM Digital Library; b) Science Direct; c) Springer; d) Web of Science; e) Wiley Online Library; and f) Google Scholar. IEEExplore was not included because it could not handle our search string due to its size. On the other hand, Web of Science and Google Scholar also indexes IEEE papers.

Before applying the selection criteria given the topic of the review, it was defined generic exclusion criteria: a) Published in non-peer reviewed publication channels such as books, thesis or dissertations, tutorials, keynotes, and others. OR; b) Not available in English OR; c) A duplicate. The first two criteria was implemented in the search strings executed in the digital libraries, and the remaining papers were evaluated through two sets of selection criteria: a) basic and b) advanced.

1. Basic criteria

The fist criteria was based on the titles and abstracts of the papers. These criteria were applied to papers that passed the generic exclusion criteria and were identified through database search or snowballing. In this context, the papers included were related to data science AND related to medical affairs.
Following the procedure presented by Ali et al.(2014), the papers were classified as: a) Relevant; b) Irrelevant; c) or Uncertain (in the case, the available information on the title and abstract is inconclusive). Only the papers evaluated as relevant were select for inclusion in this research.

2. Advanced criteria

The advanced criteria are related to the actual data extraction, in which the full-text of the papers were thoroughly read. The studies published in multiple papers and only including the extended version of the study. Additionally, all the papers that were not relevant to assess the request questions were excluded as they did not contain any relevant information. In other words, a paper was only included if it contained examples of data science applied and used in a medical affairs context.

The snowballing approach was, first, performed on the set of papers identified through the database search and included using the necessary criteria. For each paper in the set, we applied the backward and forward snowballing. To execute the forward snowballing, it was used Springer and Google Scholar to identify the title and abstract of the papers, citing our set of selected papers - the essential criteria was used to include these papers.
To execute the backward snowballing, first, we distributed the papers to be evaluated, and the reviewer was responsible for applying the generic exclusion criteria shown, as presented in section 2.3.
This was done by evaluating the title in the reference list and, if necessary, the place of reference in the text. Afterward, the included studies were evaluated using the essential criteria, in which the reviewer assessed each paper.

For the data extraction it was used a spreadsheet editor to record relevant information, and to map each article metadata. The general information extracted was (Table 1):

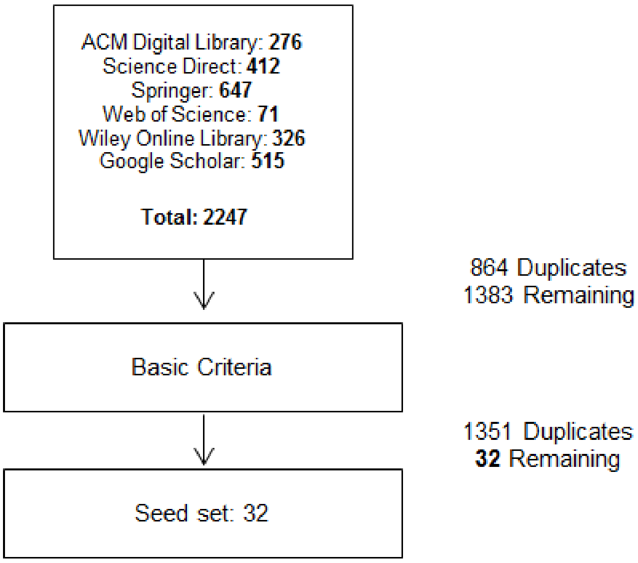
Table 1: Metadata extracted from the papers

i.	type of article
ii.	name of the publication channel
iii.	year of publication
iv.	data science used the method
v.	statistical analyze applied
vi.	number of cases
vii.	research type
viii.	research question type
ix.	empirical research type
x.	research validation

For question (vii), we used the classification presented by Wieringa et al. (2006): validation research, evaluation research, solution proposal, philosophical papers, opinion papers, or experience papers. For (viii), we used the classification presented by Shaw (2003): method or means of development; a method for analysis or evaluation; design, evaluation, or analysis of a particular instance; generalization or characterization; or feasibility study or exploration. For question (ix), it was used the classification presented by Tonella et al. (2007): experiment, observational study, experience report, case study, or systematic review. For (x), it was used the classification scheme presented by Shaw (2003): analysis, evaluation, experience, example, persuasion, or blatant assertion.

Also, as seen in figure 2, initially was identified 2247 papers through the different data sources, and then it was applied the different criteria identified previously. The final result was 32 papers, where 864 duplicates were removed from the essential criteria definitions and then 1351 were duplicates which were removed from the final selection of articles.

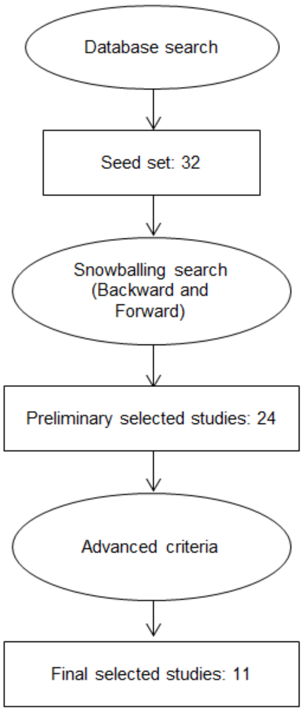
Figure 2: Overview of the database search



3. Results

In this section, the results of the systematic review process are presented. Figure 3 shows an overview of the different stages of the articles search. The details of the results of the database search can be seen in Figure 2 and the results of the snowballing process (iterated twice) can be seen in Figure 3.

Figure 3: Number of papers in study selection.


Database search
Seed set: 32
Snowballing search (Backward and Forward)
Preliminary selected studies: 24
Advanced criteria
Final selected studies: 11

Following it is possible to see the list of the included studies in this research (Table 2).

Table 2: Overview of the selected studies.

Paper Number	Authors	Year	Title	Details	Publication channel
P1	P Chou, K Hong, C Lei, H Zhang	2017	Correlation between Cancer Research Trends and the Importance of Cancers based on Mortality and Diagnosis Rates: An Analysis of Altmetric Data	STEM Fellowship Journal, 2017 - NRC Research Press	Journal
P2	M Vassilaki, JA Aakre, WK Kremers	2018	Association Between Functional Performance and Alzheimer's Disease Biomarkers in Individuals Without Dementia	Journal of the …, 2018 - Wiley Online Library	Journal
P3	S Broes, D Lacombe, M Verlinden, I Huys	2018	Toward a tiered model to share clinical trial data and samples in precision oncology	Frontiers in medicine, 2018 - frontiersin.org	Magazine and scientific journal
P4	J Magalhães, Z Hartz, A Antunes	2017	An Overview of the Open Science in Times of Big Data and Innovation to Global Health	International Journal of …, 2017 - redalyc.org	Journal
P5	A Matranga	2017	Digital Technology & Patients Tools in Clinical Trials	S Consulting - 2017 - kayentis.com	Web Page
P6	K Ottoboni, F Lewis, L Salmaso	2018	An Empirical Comparison of Parametric and Permutation Tests for Regression Analysis of Randomized Experiments	Statistics in …, 2018 -	Magazine and

				amstat.tandfonline.com	scientific journal
P7	RO Roberts, JA Aakre, WK Kremers, M Vassilaki	2018	Prevalence and outcomes of amyloid positivity among persons without dementia in a longitudinal, population-based setting	JAMA …, 2018 - jamanetwork.com	Magazine and scientific journal
P8	A Wagg, I Milsom, S Herschorn, J Heesakkers	2018	Mirabegron in older vs. younger patients: Safety and tolerability data from a large integrated database	European Urology …, 2018 - Elsevier	Magazine and scientific journal
P9	P Paul, PS Aithal, A Bhuimali	2018	Health Information Science and its growing popularities in Indian self-financed universities: Emphasizing Private Universities—A Study	2018 - papers.ssrn.com	Journal
P10	D Kato, H Tabuchi, S Uno	2018	Three-year safety, efficacy and persistence data following the daily use of mirabegron for overactive bladder in the clinical setting: A Japanese post-marketing study	LUTS: Lower Urinary Tract …, 2018 - Wiley Online Library	Magazine and scientific journal
P11	By: Carnovale, Carla; Mahzar, Faizan; Scibelli, Sara; et al.	2019	Central nervous system-active drug abused and overdose in children: a worldwide exploratory study using the WHO pharmacovigilance database	EUROPEAN JOURNAL OF PEDIATRICS Volume: 178 Issue: 2 Pages: 161-172 Published: FEB 2019	Journal

In Figure 4, it is showed the number of papers per year, and figure 5, shows the distribution of papers per type of publication channel. Also, in table 2, it is showed the techniques used in the 11 papers studied and the conclusion of the RQ1.

Figure 4: Number of papers per year.

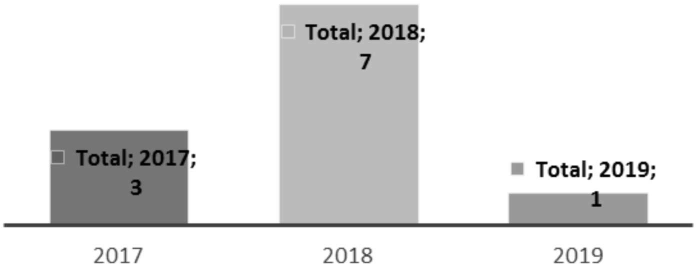
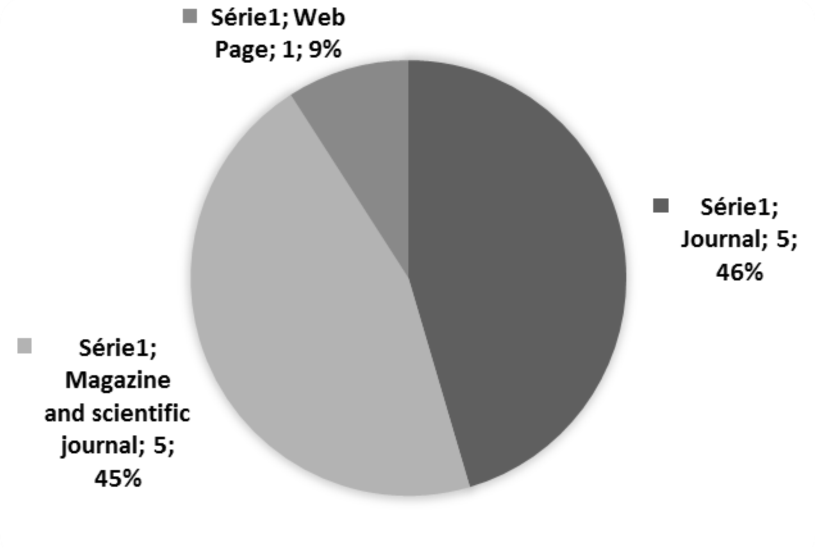


Figure 5: Distribution of papers per type of publication channel.



RQ1: What are the most used Data Science techniques in Medical Affairs case studies, research papers, or academic investigation articles.

This section shows which are the data science techniques most used in Medical Affairs case studies, research papers, or academic investigation articles. A total of 7 techniques were identified, being applied logistic analysis the most used, as shown in Table 3.

Table 3: Overview of the selected techniques used in identified papers

Paper Number	Publication channel	Technique	Tool used
P1	Journal	Linear and logistic regression	R and Python
P2	Journal	Test of Hypotheses, Logistic Regression, and Association Rules	SPSS and R
P3	Magazine and scientific journal	Linear and logistic regression	SPSS and R
P4	Journal	Linear and logistic regression	Python and SPSS
P5	Web Page	Linear regression	R and Python
P6	Magazine and scientific journal	Regression and classification analysis	R and Python
P7	Magazine and scientific journal	Linear and logistic regression	R and Python
P8	Magazine and scientific journal	Linear and logistic regression	Python and SPSS
P9	Journal	Linear and logistic regression	Python and SPSS
P10	Magazine and scientific journal	Regression Analysis	R and Python
P11	Journal	Regression analysis	R and Python

Furthermore, it is also possible to observe that seven papers used applied linear statistical analysis and three applied multiple regression analysis techniques. Different types of statistical techniques are applied in medical publications using data science methodology, and computing capacities from open software tools, like R, Python, and SPSS.

Moreover, the literature shows that MA usage of data science can have several different applications with tremendous benefits. Some of the analyzed applications from the literature review done are described in the below points:

- Logistic Regression, which is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. The advanced knowledge of concepts and theories in the therapeutic area(s), including competitive landscape and current medical and scientific knowledge (e.g., disease states, product label, statistics interpretation), requires advanced knowledge of research, including but not limited to observational and clinical study design, hypothesis testing, basic statistical methods, clinical study analyses and basic understanding of other types of research.
- Classification, the process of learning a model that elucidates different predetermined classes of data. It is a two-step process, comprised of a learning step and a classification step. In the learning step, a classification model is constructed, and the classification steps the constructed model is used to prefigure the class labels for given data.
- Linear Regression, a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables, in particular, are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
- Multiple Regression Analysis, an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target, or criterion variable). Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors and widely used in medical or clinical publications.

RQ2 What are the main responsibilities and activities from Medical Affairs that will require Data Science in supporting the execution of those.

At this point it is important to make an analysis to the related concepts of medical affairs and the primary responsibilities and activities that are connected with data science, for a better understanding on how the data science techniques are supporting medical affairs. Medical Affairs is in the core of responsibilities, a global function. Medical Affairs involves a wide range of overall strategy activities and take the lead on activities that can span to several different operating regions, such as publication planning; draw general guidelines for selecting which studies to sponsorship (company-owned or from investigator-initiative research); defining real-world data and outcome research needs; Advisory board and preceptorships. Typically Medical Affairs is a division within the R&D department and interacts with many different stakeholders, both internal and external. Regarding internal HCPs, it interacts with marketing, regulatory affairs, clinical research, medical information, pharmacovigilance, quality & compliance, market access, sales, and legal.

Considering external HCPs, the Medical Affairs team establishes straight, peer-to-peer interactions between collaborators from the pharmaceutical company and healthcare professionals, mainly physicians, scientific societies, and payers. Also, from the literature review, it was able to understand that a medical affairs major areas of responsibility are shown in the following table (4):

Table 4: MA major areas of Responsibility

Provide a medical perspective and support to the development of new medicines,
Provide the medical input necessary to support marketed medicines throughout their life cycle;
Ensure that the overall drug safety reporting process fully adheres to applicable requirements;
Perform internal audits and quality reviews;
Identify risks from a medical point of view and implement risk management activities;
Review promotional, educational and corporative materials;
Act as the ethical conscience of the company;
Perform in-country clinical studies feasibilities;
Conducting non-registration (typically Phase IV) clinical studies;
Supports investigator-initiated studies;
Executing health economics and outcomes studies in partnership with the Market Access department.
An essential role in disseminating written information to the scientific community;
Lead the development of the sponsorship strategy for external education programs for healthcare professionals, particularly in areas of focus within the company product portfolio;
Identify new business opportunities, unmet scientific/medical/patient needs, and to contribute to challenge the status quo and suggest improvement actions, both for internal or external HCPs.

4. Empirical Study

This section of the article deals with the analytical part of research, where the design of research is developed, the approach of statistical analysis and the development of hypotheses, which are decurrent from the research questions: RQ1: What are the most used statistical techniques in Medical Affairs case studies, research papers or academic investigation articles, where data scientists were used, and conventional data science tools were selected; RQ2: What are the main responsibilities and activities from Medical Affairs that will require Data Science in supporting the execution of those.

a) Data collection and sample

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). 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 (5) shows a summary of the information regarding the data collection and the technical matters of the sample.

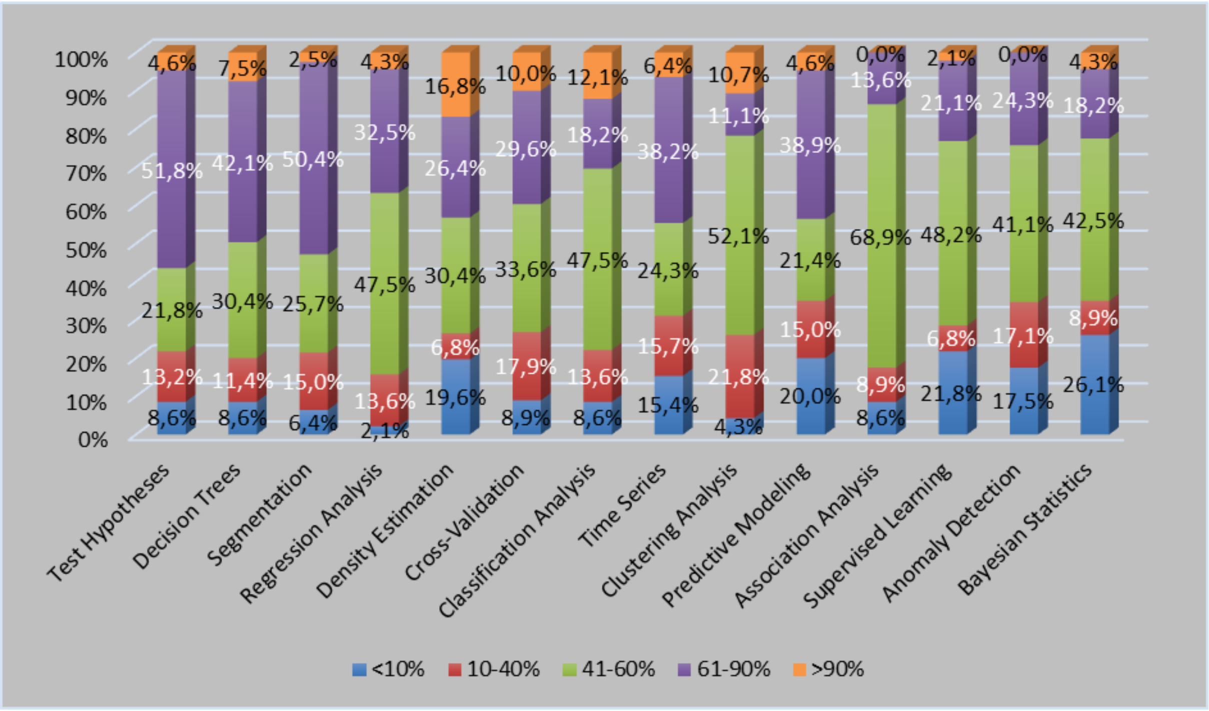
Table 5. Fact Sheet

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%

b) The hypothesis of the research study

In order to examine what are the most used statistical techniques in Medical Affairs, it was executing a descriptive analysis of what is approximately the percentage that is further processed for value generation.

Figure 5: Descriptive Analysis of Data Science Techniques



The Data Science Techniques most used are Test Hypotheses (57,9%), Decision Trees (57,2%), and Segmentation (57,1%). The less used techniques are Supervised Learning (44,7%), Anomaly Detection (43,9), and finally, Bayesian Statistics (42,4%). Moreover, it was analyzed the percentage of use of that technique according to if the organization has or does not have a strategy on Data Science (table 6).

Table 6: Descriptive Analysis of Data Science Techniques

Data Science Techniques	**Total Average** **N =280**	**Have a strategy on DS** **n =190**	**Do not have a strategy on DS** **n = 90**
Test Hypotheses	57,9%	66,5%	39,7%
Decision Trees	57,2%	60,4%	50,4%
Segmentation	57,1%	64,0%	42,4%
Regression Analysis	55,7%	58,4%	50,1%
Density Estimation	53,6%	60,2%	39,7%
Cross-Validation	53,4%	61,7%	36,1%
Classification Analysis	52,8%	60,7%	36,1%
Time Series	51,6%	50,7%	53,4%
Clustering baseAnalysis	50,2%	51,7%	47,1%
Predictive Modeling	49,1%	55,1%	36,3%
Association Analysis	47,3%	49,5%	42,7%
Supervised Learning	44,7%	47,1%	39,7%
Anomaly Detection	43,9%	46,9%	37,7%
Bayesian Statistics	42,5%	47,8%	31,4%

It was also executed an analysis of the differences in the use of Data Science Techniques based on the criteria if the organization has a Data Science strategy. In these cases, it is possible to contrast if the variables are independent or not through a chi-square test. Then a contingency table (7) was made to verify the intensity of the relationships between independent variables. Cramer test is used to analyze whether there is an association between the variables and their intensity (Cramer, 1946).

Table 7: Chi-square test and V of Cramer test – according to the organization having a strategy on DS

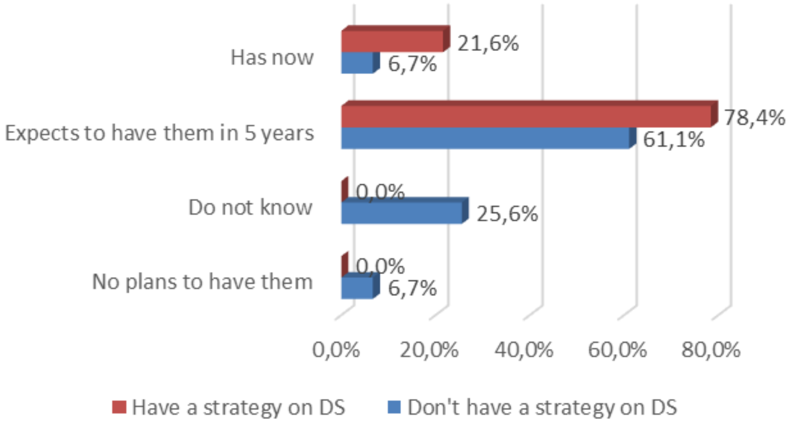
Data Science Techniques	**Chi-square test**	**Cramer test**
Test Hypotheses	159.58***	0.755***
Decision Trees	14.8***	0.23***
Segmentation	92.75***	0.576***
Regression Analysis	26***	0.305***
Density Estimation	57.26***	0.452***
Cross-Validation	147.28***	0.725***
Classification Analysis	77.17***	0.525***
Time Series	16.83***	0.245***
Clustering Analysis	109.03***	0.68***
Predictive Modeling	107.81***	0.621***
Association Analysis	129.55***	0.305***
Supervised Learning	20.26***	0.269***
Anomaly Detection	54.963***	0.443***
Bayesian Statistics	39.99***	0.378***

Note: ***: p-value < 0.01

If the organization has a strategy on DS, the tendency is to have a higher percentage of data science techniques.

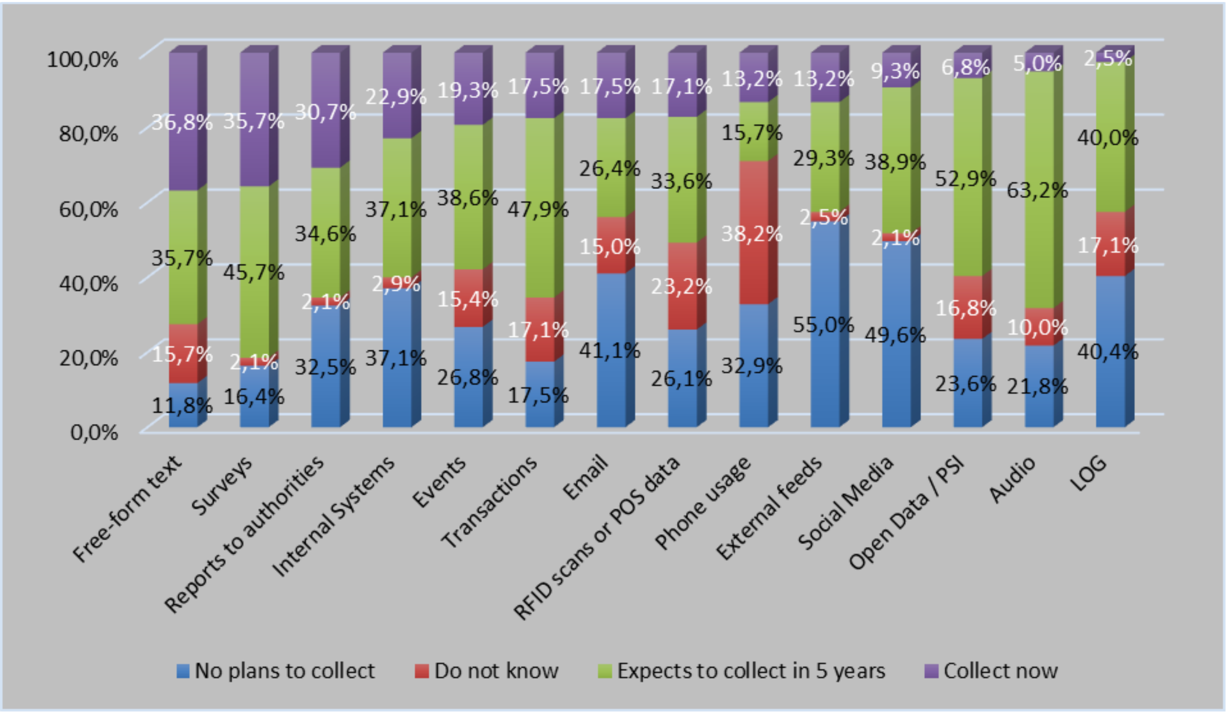
The following chart (figure 6) shows if the organization with DS strategy has the right analytical tools to handle data science. The results show that 21.6% of the organizations which have a strategy on DS have the right tools currently, and 61.3% expect to have them in 5 years. On the other hand, 6.7% of the organizations which do not have a strategy on DS don’t have plans to have analytical tools, and 25.6% do not know if they will have those tools, but 61.1% expect to have them in 5 years, and only 6.7% has now planned to have data analytical tools.

Figure 6: Does your organization have the right analytical tools to handle DS?



Moreover, it is essential to analyze the sources of data collection, or the sources they expect to collect data in the future.

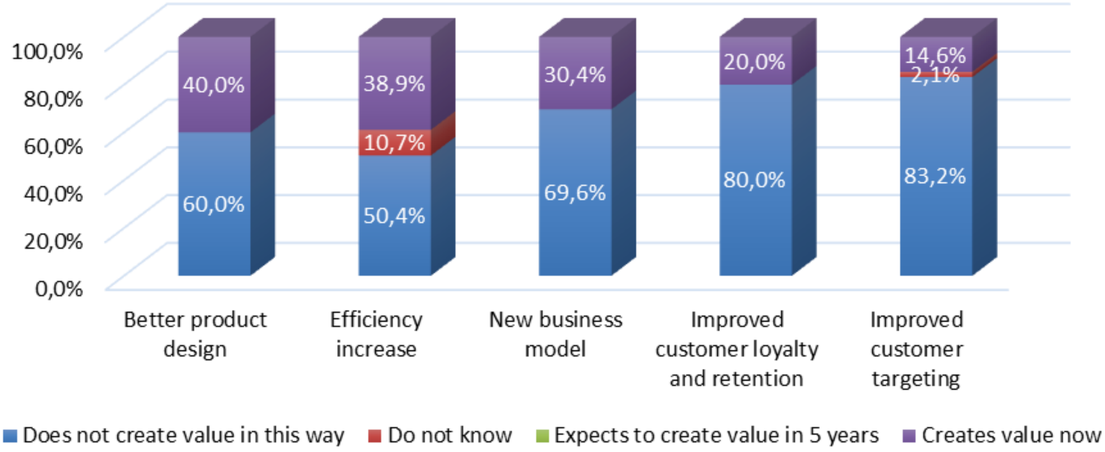
Figure 7: Data Sources



According to the chart, 36.8% of the organizations collect now free-form text, and 35.7% are from surveys and 30.7% from reports to authorities. The data sources with low potentialities to collect data are open data/public sector information (6.8%), followed by audio (5%) and finally, LOG (2.5%).

In order to examine what are the primary responsibilities and activities from Medical Affairs required by DS, it was analyzed the data-driven innovation impacts. The question was in which way does DS create, or is expected to create, value in the organization.

Figure 8: Data-driven innovation impacts



The analysis shows that 40% of organizations consider that they currently create value in "Better product design" and 38.9% in “Efficiency increase." On the other hand, “Improved customer targeting” is where it is considered to create less value.

5. Conclusions

The purpose of this article was to demonstrate, and communicate the value of data science in a medical affairs function in enhancing the knowledge of medicines and the associated therapeutic areas in which a company focus its research efforts, in providing thorough understanding of its medicines: interpret emerging scientific trends, clinical data and the competitive landscape and align internal stakeholders on a balanced benefit/risk proposition. All of the described research questions were validated in this article where data science can play a decisive role in giving the necessary tools and processes to a medical affairs department in communicating to the medical and scientific communities in an accurate, fair and balanced manner about the benefits and the risks of the medicines, enabling prescribers and other healthcare decision-makers to make informed decisions with patients and use medicines safely and effectively. Data Science also gives concrete support to medical affairs in working cross-functionally with colleagues from Marketing, Sales, Regulatory and Access to guide the acquisition and integration of clinical data so that existing clinical evidence is communicated accurately, reflecting the value of the medicines, to help to inform the right capital allocation decisions in the advancement of the lifecycle of the brands and the company's pipeline and to ensure launch readiness, organizing and training medical affairs colleagues and providing them with the tools necessary for becoming more strategic. For this research a systematic literature review on the application of data science to medical affairs was performed following to the application of an online survey. As the main techniques for the systematic literature survey it was used an hybrid approach composed of database search and snowballing. After all the selection process, only 11 articles were included in this research. The metadata from the paper helped to answer the research questions, and the main conclusions are that there is a variety of data science techniques used in medical affairs, and that the most part of the research used a multilevel approach to perform advanced statistical analysis, using high-performance computing capacity from open software tools. The results base of that conclusion will be a base for data scientists or medical affairs practitioners as a guide to apply data science techniques on their projects. Other main conclusion of this research is that there is a need in the academic and practitioner community to have more knowledge on the Data Science methods and tools, and also on the benefits of data science applied in medical affairs activities and research projects. Other main conclusion is that the role of Medical Affairs within a pharmaceutical company serves to spearhead the dissemination (and in some cases, the generation) of unbiased clinical and scientific information about medicine to the healthcare community and to offer medical and scientific expertise, and that Data Science can help to make this function even more strategic, mainly regarding decision-making, and forecasting the company future.

References

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

Cramer, H. Mathematical methods of statistics (Vol. 9). Princeton university press; 1999.

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

Dyer S. "Medical Science Liaison – Aligning the activities and goals of Medical Science Liaison teams for strengthened corporate sustainability." The Medical Science Liaison Corporation, 2011.

Deepika Badampudi, Claes Wohlin, and Kai Petersen. 2015. Experiences from Using Snowballing and Database Searches in Systematic Literature Studies. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE '15). ACM, New York, NY, USA, Article 17, 10 pages.
DOI:http://dx.doi.org/10.1145/2745802.2745818

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.

Grom T. Medical affairs: beyond the science. Showcase feature; 2013.
Available from: http://www.pharmavoice.com/article/medical-affairsbeyond-the-science/.
Accessed January 30th, 2019.

Hair, J. F., Black, W. C., & Babin, B. J. Anderson. R. Multivariate Data Analysis. New Jersey, Pearson Prentice Hall; 2010.

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.

Kitchenham Barbara and Charters Stuart. 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE-2007-01. School of Computer Science and Mathematics, Keele University.

Levene, H. Contributions to probability and statistics. Redwood City, California: Standford University Press; 1960.

Li Yao, Longfei Zhu, Cheng Cui, Exploration of Data Science Course Construction and Personnel Training in Big Data Era, Computer Generation (2018).

Mercade-Mele, 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.

Nauman Bin Ali, Kai Petersen, and Claes Wohlin. 2014. A Systematic Literature Review on the Industrial Use of Software Process Simulation. J. Syst. Softw. 97, C (Oct. 2014), 65–85. DOI:http://dx.doi.org/10.1016/j.jss.2014.06.059

Paolo Tonella, Marco Torchiano, Bart Du Bois, and Tarja Systä. 2007. Empirical studies in reverse engineering: state of the art and future trends. Empirical Software Engineering 12, 5 (2007), 551–571. DOI:http://dx.doi.org/10.1007/ s10664-007-9037-5

PharmaForum. Medical Affairs – the heart of a data-driven, patient-centric pharma; 2017. Available from: https://pharmaphorum.com/views-and-analysis/medical-affairs-heart-data-driven-patient-centricpharma/. Accessed January 30th, 2019.

Plantevin L, Schlegel C, Gordian M. Reinventing the Role of Medical Affairs; 2017. Available from: http://www.bain.com/publications/articles/reinventing-the-role-of-medical-affairs.aspx

Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohammed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software 80, 4 (2007), 571 – 583. DOI:http://dx.doi.org/10.1016/j.jss.2006.07.009

Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. 2006. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. Requirements Engineering 11, 1 (2006), 102–107. DOI: http://dx.doi.org/10.1007/s00766-005-0021-6

Rollins BL, Perri M. Pharmaceutical Marketing. Burlington: Jones & Bartlett Learning; 2014.

Shaw. M. 2003. Writing good software engineering research papers. In 25th - International Conference on Software Engineering, 2003. Proceedings. 726–736.

DOI:http://dx.doi.org/10.1109/ICSE.2003.1201262

Strategic Benchmarking Research, 2014. Accessed January 30th, 2019, at: http://pt.slideshare.net/bestpracticesllc/pop-253-a-report-summary-strategic-kol-management.

Suresh B, Buxton C, Ferrer J, Piervincenzi R, Nathoo A. "Managing talent in the medical affairs function: Creating value through a strengths-based approach." McKinsey & Company, and Korn/Ferry International, July 2013. Accessed January 30th, 2019, at: http://www.mckinsey.com/~/media/McKinsey/dotcom/client_service/Pharma%20and%20Medical%20Products/PMP%20NEW/PDFs/Managing_Medical_Affairs_Talent.ashx.

Tyson G, Doyle K. "Optimizing the Impact of the Medical Affairs Function." Campbell Alliance, 2013. Accessed January 31th, 2019, at: http://www.campbellalliance.com/articles/Campbell_Alliance_Optimizing_the_Impact_of_the_Medical_Affairs_Function.pdf.

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

Wolin MJ, Ayers PM, Chan EK. “The emerging role of Medical Affairs within the modern

Pharmaceutical Company”. Drug Information Journal, 2001.Volume 35; pages 547–555.

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