Data Science In Pharmaceutical Industry

António Pesqueira[1], Maria José Sousa[2], Álvaro Rocha[3], Miguel Sousa[4]

[1] Alexion Pharmaceuticals, Zurich, Switzerland
antonio.pesqueira@live.com

[2] Instituto Universitário de Lisboa, Portugal
mjdcsousa@gmail.com

[3] Universidade de Coimbra, Portugal
amrochagmail.com

[4] Essex University, UK
miguel.ac.sousa@gmail.com

Abstract. Data Science demand from Medical Affairs (MA) functions in the pharmaceutical industry are exponentially increasing, where business cases around more modern execution of activities and strategic planning are becoming a reality. MA is still lagging in terms of implementing data science and big data technology in the current times, which means a reflecting immaturity of capabilities and processes to implement these technologies better. This paper aims to identify possible gaps in the literature and define a starting point to better understand the application of Data Science for pharmaceutical MA departments through the identification and synthesis of data science criteria used in MA case studies as presented in the scientific literature. We applied 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 Takeda, a Pharma organization. The findings indicate that there is good evidence in the empirical relation between Data Technostructure and Data Management dimensions of the Data Science strategy of the organization..

Keywords: Data Science, Pharmaceutical Industry, Literature Review, Big data technologies

1 Introduction

Over the last ten years, the Pharmaceutical Industry has been under greater scrutiny from regulators, healthcare professionals (HCPs), and patients. Resulted from those factors, pharmaceutical companies are now relying tremendously on strategic functions like Medical Affairs (MA).

For quite some time, MA' primary role was defined mainly as a scientific exchange, information support, managing daily regulatory reporting requirements or driving medical evidence generation (e.g., Phase IV studies, Real World Evidences or collaborative research) with a strong focus on priority diseases and developed products (Dyer 2011).

In nowadays, MA is becoming a central function and core element of all pharma operations.

As MA was set up to operate independently from sales pressures, their role has grown from the last decade to creating strategic relationships with healthcare professionals (HCPs), Key Opinion Leaders (KOLs), and other stakeholders (e.g., payers, regulators, investigators, and others).

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.

Even though Data Science is a crucial concept within the pharmaceuticals industry, to the best of our knowledge, no study explored which criteria are being used in the context of Data Science applicable to MA. There is a need for research to assess which Data Science practices proposed by practitioners are beneficial. In this context, the main goal and contribution of this paper are to report the state of the are in the field of Data Science in MA employing a systematic literature review. We followed the guidelines proposed by Kitchenham and Charters (2007). In this paper, we detail our study and also point the gaps and future directions for research in Data Science for MA functions..

2. Theoretical Framework

2.1 Introduction of Data Science and Big Data Technologies

Data Science has attracted intense and growing attention from significant healthcare and life sciences organizations, 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. Consequently, the data volume is forecasted 44 trillion gigabytes until 2020 (EMC Digital Universe 2014).

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.

The topics of data science technologies encompass two different aspects. Data science refers to 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. On 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 later integrate into distributed computation or integrated software.

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).

The critical challenges for medical affairs 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.

Data Sscience is gaining middle ground in all MA establishments for the efficient utilization of resources: storage and time and efficient decision making to exploit new methods and procedures.

2.2 Medical Affairs in the pharmaceutical industry

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.

Medical Affairs (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 (payers, regulators, HCPs) 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).

In the past, pharmaceutical companies considered MA just a support function and one that could even slow down marketing and commercial activities.

Furthermore, in this fasting moving market dynamics, the way MA engage with all the involved stakeholders is becoming more fundamental than ever, where MA is considered as the best function to provide scientific and clinical expertise to support approved medicines, work closely with R&D in the developing new drugs throughout post-approval activities and to be better prepared to respond to customer demands and to develop and maintain stronger long-term relationships with key opinion leaders, scientific societies, payers and patient groups (Plantevin et al. 2017).

To increase the transparency and efficiency of all the activities developed and its relationship with physicians, it has undergone significant changes in the way MA can understand all the surrounding data and the way that can quickly manage all the hidden connections and patterns to understand the engagement outcomes and stakeholders needs.

Many of these changes have led to an increase in the responsibility of MA but still not the full capacity to make a practical usage of the data.

3. Review Method

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.

A summary of the protocol used in this review is given in the following sub-sections.

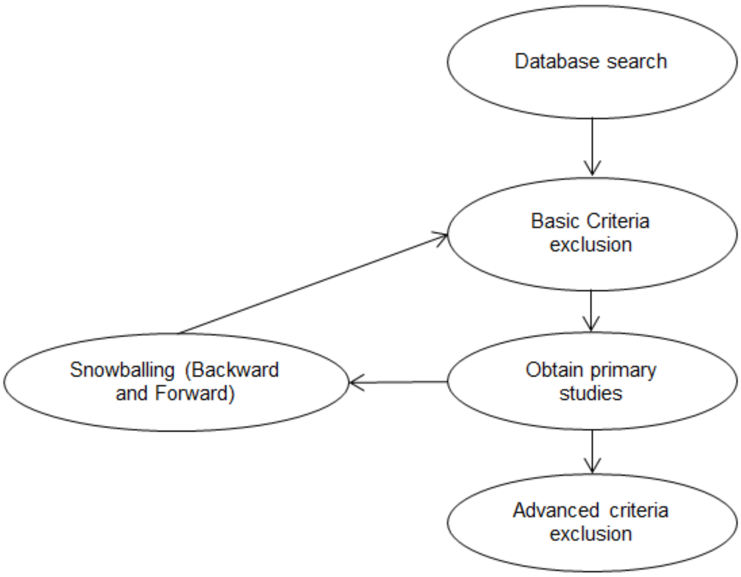

Database search
Basic Criteria exclusion
Snowballing (Backward and Forward)
Obtain primary studies
Advanced criteria exclusion

Figure 1: Overview of the search and selection process

.

3.1 Research Questions

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. Given this, we formulated the following research question (RQ): 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?

3.2 Search Strategy

To minimize the probability of missing relevant articles, publications, we used a combined search strategy, which is based on database search and snowballing (backward and forward). First, we defined a search string, which is presented in Section 2.2.1 and used it to search databases containing scientific papers in the context of data science, as shown in Section 2.2.2. After applying the essential criteria exclusion, shown in Section 2.3, the resulting papers were defined as the starting set for the snowballing. After executing the snowballing iterations, we applied the advanced criteria exclusion, which is related to the actual data extraction and quality assessment. We show an overview of the search and selection process in Figure 1.

We decided to use this strategy 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).

3.3 Search terms

Starting with the research questions, suitable keywords were identified using synonyms and related terms. The following keywords were used to formulate the search string:

- Population: Data Science and Medical Affairs. Alternative keywords: Medical Data Science, Data Science in Medical Affairs, Medical Affairs Data, and Data Science in Pharmaceutical.
- Intervention: Data Science. Alternative keywords: data science in medical affairs and medical data.
- 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”

3.4 Data Sources

Since our goal is to cover the literature published in Data Science, we chose the following digital databases for data retrieval: ACM Digital Library; Science Direct; Springer; Web of Science; Wiley Online Library; Google Scholar. We did not include IEEExplore 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.

3.5 Selection criteria

Before applying the selection criteria given the topic of the review, we defined generic exclusion criteria: Published in non-peer reviewed publication channels such as books, thesis or dissertations, tutorials, keynotes, and others. OR Not available in English OR A duplicate.

We implemented the first two criteria in the search strings that were executed in the digital libraries, wherever possible.

Afterward, the remaining papers were evaluated through two sets of selection criteria: basic and advanced.

3.5.1 Basic criteria

We applied the necessary criteria to evaluate if papers are relevant to the aims of our paper by reading the titles and abstracts. These criteria were applied to papers that passed the generic exclusion criteria and were identified through database search or snowballing. In this context, we included papers that: Are related to data science AND Are related to medical affairs.

Following the procedure presented in Ali (Ali et al. 2014), we decided that papers are classified as: Relevant, Irrelevant 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 the next section of this paper.

3.5.2 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.

3.5.3 Snowballing

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, we used Springer and Google Scholar to identify the title and abstract of the papers, citing our set of selected papers. We applied the essential criteria, shown in Section 2.3.1, 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.

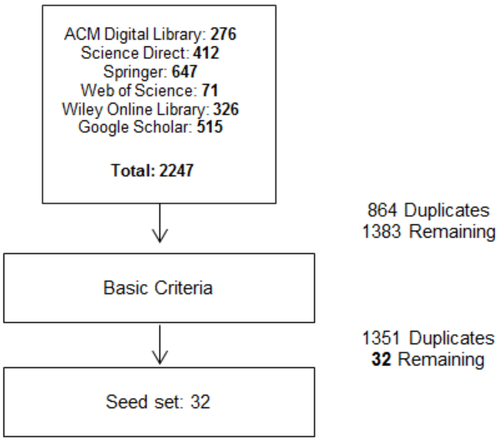
3.6 Data Extraction

We used a spreadsheet editor to record relevant information. With the spreadsheet, we were able to map each datum extracted with its source. From each paper, we extracted general information such as the year and name of the publication channel and data related to the RQs. The following data were extracted from the papers: i. type of article (journal, conference, magazine and scientific journal, webpage, and others), 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), we used the classification presented by Tonella et al. (2007): experiment, observational study, experience report, case study, or systematic review. For (x), we used the classification scheme presented by Shaw (2003): analysis, evaluation, experience, example, persuasion, or blatant assertion. Also, as we can see from figure 2, we initially identified 2247 papers through the different data sources, and then we start applying the different criteria as defined in the following points: 2.3.1 and 2.3.2.

The final seed set result was 32 remaining papers, where 864 duplicates were removed from the essential criteria definitions and then 1351 duplicates removed for the seed set creation.

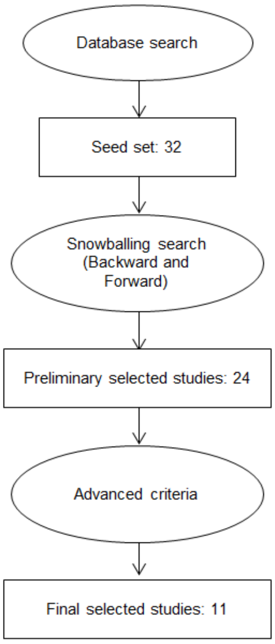
Figure 2: Overview of the database search



4. Results

In this section, we present the results for the systematic review process and the research questions as well. In Figure 3, we present an overview of the number of studies passing through the different stages of the study. We show details of the results of the database search in Figure 2 and of the results of the snowballing process, in which we iterated twice, 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

In Figure 4, we show the number of papers per year. In Figure 5, we show the distribution of papers per type of publication channel. Also, in table 2, we present the used techniques in the 11 papers studied and as a conclusion for the request question 1.

Figure 4: Number of papers per year.

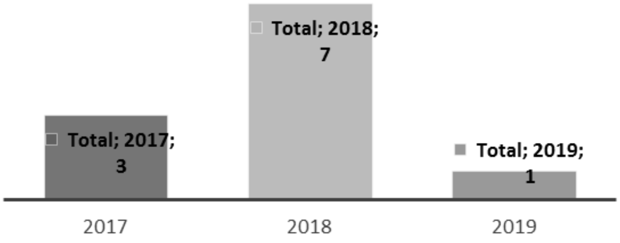
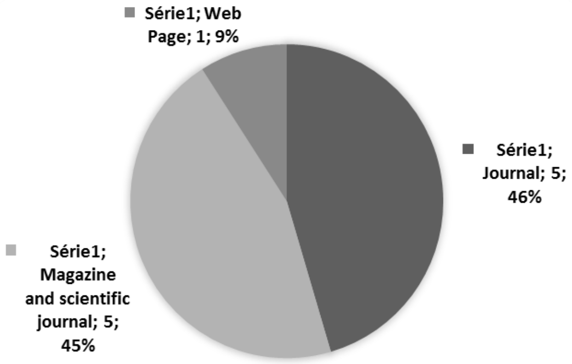


Figure 5: Distribution of papers per type of publication channel.



5. Conclusions

This study presents a systematic literature review on the application of data science to medical affairs. We used a hybrid approach composed of database search and snowballing. The primary search fetched 1383 unique results, from which 32 papers were selected as a seed set for snowballing. After the data extraction, only 11 papers were included in the study. Data from these papers were analyzed to answer each research question. There is a variety of data science techniques used in medical affairs. Furthermore, some papers used a multilevel approach to perform more advanced statistical analysis and using high-performance computing capacity from open software tools.

Data scientists or medical affairs practitioners can use the results of this study as a guide for them to apply data science techniques on their projects or compare them with their methodologies. Moreover, based on the results of this study, we recommend that there is a strong need to publish more papers presenting how data science is applied in medical affairs studies and research projects and to conduct empirical studies to assess the results of applying this practice.

For future work, we intend to execute a survey with medical affairs practitioners and compare the collected data with our results in this study.

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 medical affairs in daily activities.

In conclusion, 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.

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 indeed 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 to excel within the pre-, and post-launch period.

Acknowledgements

We would like to thank to AISTI (Associação Ibérica de Sistemas e Tecnologias de Informação / Iberian Association for Information Systems and Technologies) for the financial support granted in this study.

References

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

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

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

Dyer S.(2011) "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.

Fornell, C., & Bookstein, F. L. (1982) 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. (1981) Structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*. 1981; 18(1), 39–50.

Grom T. (2013) Medical affairs: beyond the science. Showcase feature.

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. (2010) *Multivariate Data Analysis*. New Jersey, Pearson Prentice Hall.

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.

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

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

Mercade-Mele, P., Molinillo, S., Fernández-Morales, A., & Porcu, L. (2018) CSR activities and consumer loyalty: The effect of the type of publicizing medium. *Journal of Business Economics and Management*. 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.

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, 551–571.

PharmaForum. Medical Affairs (2017) The heart of a data-driven, patient-centric pharma; 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. (2017) Reinventing the Role of Medical Affairs. 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

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, 102–107.

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

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

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 (2013) 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. (2013) 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, (2017) Exploration and Research on the Core Course Construction of Data Science and Big Data Technology Specialty, *Education Review*.

Wolin MJ, Ayers PM, Chan EK.(2001) The emerging role of Medical Affairs within the modern Pharmaceutical Company. *Drug Information Journal*, 35; 547–555.

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