Proposed use of data mining techniques in healthcare for knowledge generation in patient care: an integrative review

Reinaldo Ribeiro de Oliveira
Sarah Loretta Barbosa dos Santos
Renato Jose Sassi

1 INTRODUCTION

Data mining, also known as database knowledge discovery (KDD), is a process of extracting patterns, trends, and insights from large and complex data sets. It involves various techniques such as association rule mining, clustering, classification, and prediction, which can discover hidden patterns and relationships in data. Data mining has been widely used in various domains, including healthcare, to extract valuable insights that can inform decision making and improve outcomes (**Pooja et al, 2019**)

In healthcare, data mining has gained significant attention due to the availability of large amounts of electronic health records (EHRs) and other health data, as well as the growing need for personalized, evidence-based care. The use of data mining techniques in healthcare has the potential to revolutionize patient care by providing insights into patient characteristics, disease patterns, treatment outcomes, and other relevant information. This can help healthcare professionals make informed decisions about diagnosis, treatment, and patient management, leading to better patient outcomes and improved healthcare delivery.(**N.P.J et al, 2022**)

On the current state of data mining in healthcare, **Drayton-Brooks et al, (2020)** suggests that data mining techniques have been increasingly used in healthcare to extract knowledge and insights from large and complex datasets such as electronic health records, clinical databases, administrative databases, and health insurance claims data.

These datasets contain a wealth of information about patients, including demographics, medical history, diagnoses, treatments, laboratory results, and other relevant data that can be used for knowledge generation in patient care.(**Dreisbach C et al, 2019**)

The use of data mining in healthcare has shown promising results in several areas, such as disease prediction and risk assessment, treatment effectiveness evaluation, adverse event detection, patient stratification, and personalized medicine. For example, **Ara Shaikh et al (2022)** suggests that data mining techniques have been used to identify patterns and trends in patient data that can predict

the risk of developing chronic diseases such as diabetes, cardiovascular disease, and cancer, enabling early intervention and prevention strategies. Data mining has also been used to evaluate the effectiveness of different treatment options for specific patient populations, identify adverse drug reactions, and stratify patients based on their response to treatments, leading to more targeted and personalized care.(**Dubromel A et al, 2020**)

Electronic records can improve the quality and efficiency of healthcare. National and international bodies propagate this resource worldwide (**Shakibaei et al, 2022**) However, the evidence base on the effects and benefits of electronic records is questionable. Health system outcomes are influenced by many components, making it difficult to make claims about specific types of interventions. Moreover, electronic records themselves constitute a complex intervention that offers several strands with positive effects on health systems outcome (**Reis Z et al, 2017**) The motivation for developing this research, was the coexistence with hospital care management, production of care data in the form of archived and unused Electronic Health Records (EHR), The mining and manipulation of these idle data can generate epidemiological knowledge if retrospective, generation of interventions on current risks, prospective if in real time, as well as care interventions that can reduce risks to patients and increase the accuracy of protocol application (**Kariotis T et al**, **2022**). The stimulation of research with the use of data mining techniques in health, can produce institutional, academic and collective benefits and especially to the patient (**Sarwar, T et al, 2023**).

Aim: To demonstrate studies on data mining techniques in electronic health records (EHR) for hospital attendances published between 2015 to 2023.

2 METHODS

This study contemplates the first phase of an integrative review for the theoretical foundation of a doctoral thesis of the master's and doctoral programs of the post-graduate program of informatics and knowledge management of the university Nove de Julho of São Paulo. Methodological evaluations were adopted from a series of literature reviews and systematic review studies. The literature search was performed in PUBMED/MEDLINE, SCIENCEDIRECT, SCOPUS with the keywords "Medical Record System" and "Eletronics Health Records" and "Health Informatics" and "Data mining" and "Artificial Intelligence". The selection process comprised two phases. The first was the search and the second the consolidation and analysis of the findings. The initial search found 1320 references, of which 21 were included in the review. The evaluation combined a score of the studies' quality, a description of the data sources in the case of secondary data analyses, and a qualitative assessment of the publications' conclusions about data mining, risk identification, quality, and efficiency of care.

3 THEORETICAL FOUNDATION

This integrative review aims to provide an overview of the current state of the art of data mining in healthcare, including its potential benefits and challenges. It further explores various data mining techniques that can be used in healthcare and their applications in patient care (**Islam M et al, 2018**). The review also highlights the ethical considerations and potential limitations of using data mining techniques in healthcare and provides recommendations for future research directions in this field.

In the study we review publications on various data mining techniques, their potential applications in healthcare, and the benefits and challenges associated with their use in patient care. We also discuss ethical considerations related to the use of data mining techniques in healthcare settings and propose future directions and research strategies for integrating data mining techniques into clinical practice (**Kariotis T et al, 2022**)

In the scope of data mining techniques in healthcare, data mining techniques encompass a wide range of methods that can be applied to healthcare datasets to extract meaningful patterns and insights. Some of the commonly used data mining techniques in healthcare include machine learning algorithms, association rule mining, and clustering.(**Dubromel et al, 2020**)

Machine learning algorithms such as decision trees, support vector machines, and neural networks are widely used in healthcare for predictive and classification tasks as explained (**WU et al, 2021**). These algorithms can analyze large data sets and identify patterns that may not be apparent through traditional statistical methods. For example, machine learning algorithms can be used to predict the likelihood of a patient developing a specific condition, such as diabetes or cancer, based on their demographic, clinical, and genetic data. Machine learning algorithms can also help identify risk factors associated with adverse events, such as hospital readmissions or medication errors, and assist in the development of targeted interventions to reduce these risks (**Jothi N et al, 2022**)

Association rule mining is another data mining technique that can be used in healthcare to discover patterns and relationships between variables in large data sets. Association rule mining can help identify associations and interoperability between different clinical variables, such as symptoms, diagnoses, and treatments, and can provide insights into the co-occurrence of certain conditions or the effectiveness of specific treatments (**Reis Z et al, 2017**) For example, association rule mining can be used to identify patterns in medication prescription data, such as frequent co-prescribing of certain medications, which can inform medication management strategies.

Clustering is a data mining technique that involves grouping similar data points based on their similarities in terms of attributes or characteristics. Clustering can be used in healthcare to identify subpopulations of patients with similar clinical profiles or disease characteristics. For example, clustering techniques can be applied to electronic health record data to identify groups of patients (**Pooja H et al, 2019**)

Flowchart 01 - Data mining steps
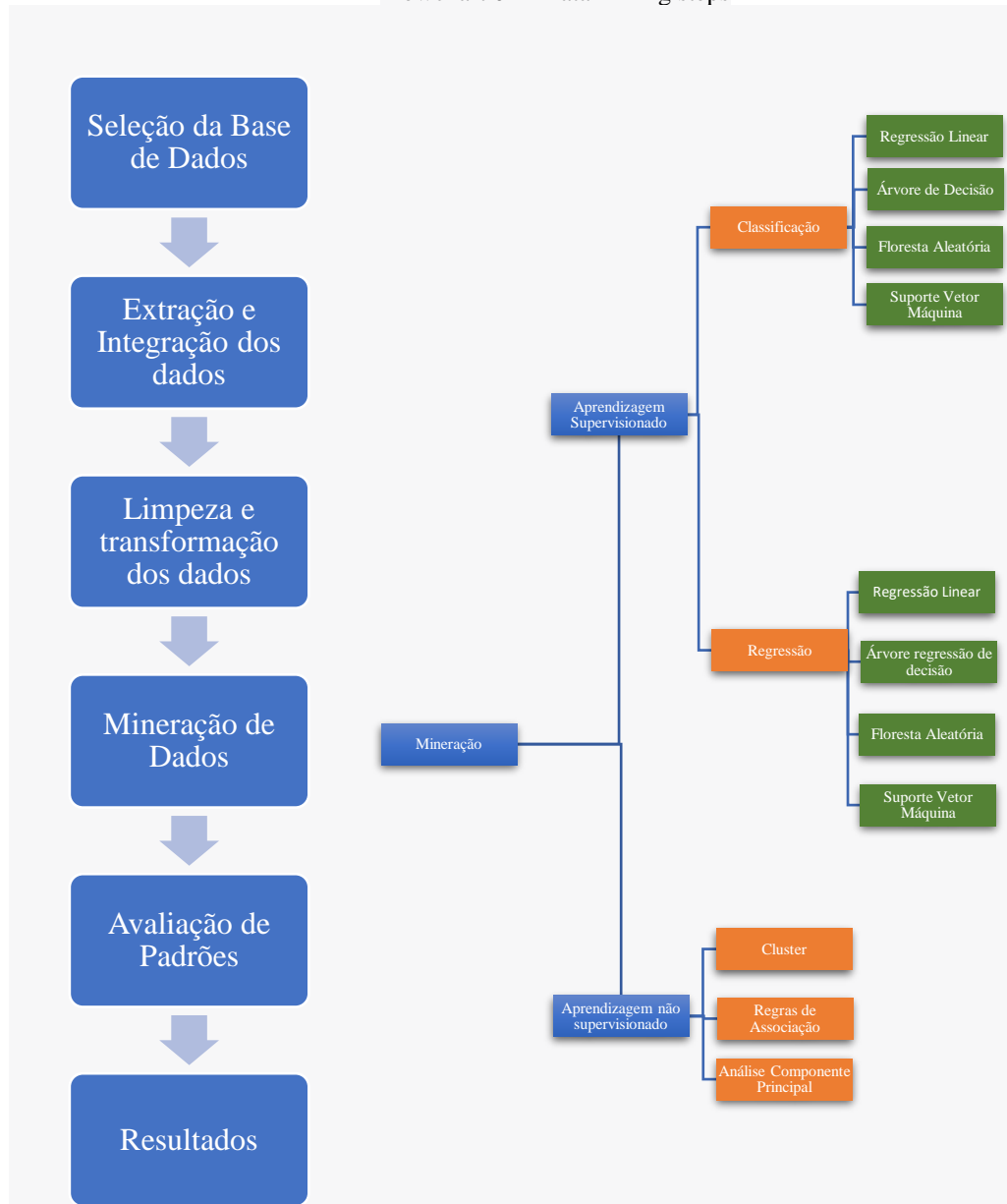
This study contemplates the first phase of an integrative review for the theoretical foundation of a doctoral thesis in progress at the post-graduate program in computer science and knowledge management at the Nove de Julho university in São Paulo. Methodological evaluations were adopted from a series of literature reviews and systematic studies and reviews. The literature search was performed in PUBMED/MEDLINE, SCIENCEDIRECT, SCOPUS, and IEEExplore with the keywords "Medical Record System" and "Eletronics Health Records" and "Health Informatics" and "Data Mining" and "Natural Language Processes".

4 RESULTS

The selection process comprised two phases. The first was the search and the second the consolidation and analysis of the findings. The database searches initially found 1320 publications, 15 of which were included in this review. The 15 studies included in the detailed reviews were available in full. Of the 1320 studies found, 860 (65.16%) were discarded after reading the title, 445 (33.72%) were excluded after reading the abstract, and 32 (2.43%) articles were discarded after reading the full text. Thus, this left 15 studies used for the development of the research. The evaluation combined a score of the studies' quality, a description of the data sources in the case of secondary data analysis, and a qualitative evaluation of the publications' findings on data mining in electronic health records in identifying risks, quality and efficiency of care. Of the 14 articles selected for final review, 8 are from health data science research groups and another 6 are from technology and innovation research groups.

Flowchart 02 - Consolidated database search results
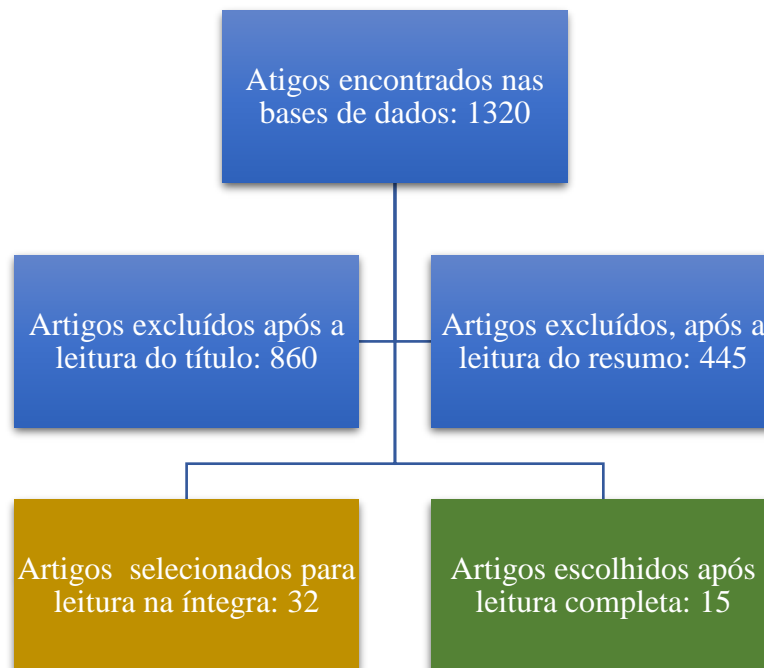
Most of the studies analyzed, were produced in India with (33.3%) in China with (13.3%) in the United States with (13.3%) in Australia with (13.3%) in Brazil with (6.66%) in Malaysia with (6.66%) and in France with (6.66%). All studies used fully publicly available data ("secondary data studies"; 9/15, 60%). A total of (12/15, 80%) studies analyzed the effect of the treatment of the bases produced in electronic medical records, the quality of health data (8/15, 53.33%), (6/15, 40%) the quality of the bases. The primary data studies achieved a mean score of 4.3 (SD 1.37; theoretical maximum 10); Of the secondary data, a mean score of 7.1 (SD 1.26). Of the primary studies, 2 demonstrated the need for base preparation and treatment. All studies demonstrated positive impacts on information mapping with the possibility of healthcare analysis. Finally, 9/15 studies showed

possibilities of punctual interventions in the care process and (9/15, 56%) studies showed increased quality of health care when scanning enables interventions in (80%).

Table 01 - Scope of the studies evaluated

Studies selected for analysis in the integrative review			
Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Wu et al, 2021/China	The objective of this study was to assist clinical researchers in understanding the application of mining technology	Data Mining	Description of methods associated with data mining and commonly used to process clinical big data
Wu, WT., Li, YJ., Feng, AZ. *et al.* Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Med Res* **8**, 44 (2021). https://doi.org/10.1186/s40779-021-00338-z			
Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Wencheng Sun et al, 2018/China	This article focuses on the processing of EHRs (electronic health records) and looks emphatically at the main techniques.	Data and Text Mining	In the future, the larger scale and more complex structure of the RES will make it more difficult to process data in the medical record, but the social and economic benefits it brings will be more noticeable and research at the grassroots will play a greater role in the health field.
Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, Guoyan Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review," Journal of Healthcare Engineering, vol. 2018, Article ID 4302425, 9 pages, 2018. https://doi.org/10.1155/2018/4302425			
Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Jayasri N.P. et al, 2022/India	he goal of this paper was to evaluate the medical database of diabetes patients through a combination of innovative hierarchical decision attention network to more accurately predict the occurrence of diabetes datasets to discover the optimal solution for the patient.	Data Mining	To perform this analysis, the MapReduce platforms are used in addition to the proposed hierarchical algorithm, such as hierarchical decision attention network, AA, and outlier-based multiclass classification.
N.P., J., & Aruna, R. (2022). Big data analytics in health care by data mining and classification techniques. *ICT Express*, *8*(2), 250-257. https://doi.org/10.1016/j.icte.2021.07.001			
Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Richa Jain et al, 2021/India	One of the challenges was how to discover useful and important information effectively amidst the huge amount of data available through data mining techniques.	Data and Text Mining	This article reviewed different predictive and descriptive data mining techniques that have been applied to voluminous health records to aid in the decision-making process.
Richa Jain, Devendran V, "Data Mining Algorithms in Healthcare: An Extensive Review," *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp.728-733, 2021. https://doi.org/10.1109/ICATIECE45860.2019.9063623			
Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Vimal et al, 2023/India	Submit tools to accumulate, manage, analyze, and assimilate large volumes of disparate structured and	Data Mining	AI-based analytical patterns can handle large-scale data more efficiently compared

			to traditional machine learning methods, and based on inference, the diagnostic process can be carried out in medical fields.

Vimal, S., Rho, S., & Pelusi, D. (2023). Special Issue on Artificial Intelligence Empowered Big Data Analytical Patterns for Medical Applications. *Neural Processing Letters*, *55*(1), 1-1. https://doi.org/10.1007/s11063-023-11158-8

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Drayton-Brooks et al, 2020/USA	The purpose of this study was to examine the use of big data and data mining of clinical education for nursing professionals and to establish a foundation for competency-based education	Data Mining	Using a data mining knowledge discovery process, faculty are able to gain greater understanding of clinical practice experiences to inform competency-based nursing education and the use of trusted professional activities for the future

Drayton-Brooks, S. M., Gray, P. A., Turner, N. P., & Newland, J. A. (2020). The use of big data and data mining in nurse practitioner clinical education. *Journal of Professional Nursing*, *36*(6), 484-489. https://doi.org/10.1016/j.profnurs.2020.03.012

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Ara Shaikh et al, 2022/India	Identify the main applications of data mining in medicine	Data Mining	After collecting data, unique strategies are implied to implement important resources for patient care and treatment. There are many useful applications of data mining in the detection, diagnosis and treatment of a disease.

Ara Shaikh, A., Nirmal Doss, A., Subramanian, M., Jain, V., Naved, M., & Khaja Mohiddin, M. (2022). Major applications of data mining in medical. *Materials Today: Proceedings*, 2300-2304. https://doi.org/10.1016/j.matpr.2021.11.642

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Dreisbach C et al, 2019/USA	In this systematic review, we aimed to synthesize the literature on the use of natural language processing (NLP) and text mining as they apply to the extraction and processing of symptoms in patient-authored electronic text (ePAT)	Text Mining	Understanding the role that ePAT plays in health communication and real-time symptom assessment, through the use of NLP and text mining, is critical for a patient-centered healthcare system

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, *125*, 37-46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Shakibaei Bonakdeh et al, 2022/Australia	Summarize empirical studies on the value of electronic health records for hospital care	Data mining and electronic health records	Since the results of secondary studies are often consulted by health care managers and policy makers, researchers should be aware of the broad impacts of research bias on fundamental decisions that they can cause

Shakibaei Bonakdeh, E. (2022). Interpretation Bias Toward the Positive Impacts of Digital Interventions in Health Care. Comment on "Value of the Electronic Medical Record for Hospital Care: Update From the Literature." *Journal of Medical Internet Research*, *24*(3), e37208. https://doi.org/10.2196/37208

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Reis et al, 2017/Brazil	The purpose of this study was to evaluate the cost-effectiveness of using electronically available inpatient data systems, health information exchange, or standards to support interoperability between systems.	Data mining and system interoperability	Hospital information systems, along with information sharing, have the potential to improve clinical practice by reducing staff errors or incidents, improving automated harm detection, monitoring infections more effectively, and enhancing continuity of care during information transfers

Reis, Z. S. N., Maia, T. A., Marcolino, M. S., Becerra-Posada, F., Novillo-Ortiz, D., & Ribeiro, A. L. P. (2017). Is there evidence of cost benefits of electronic medical records, standards, or interoperability in hospital information systems? overview of systematic reviews. In *JMIR Medical Informatics* (Vol. 5, Issue 3). JMIR Publications Inc. https://doi.org/10.2196/medinform.7400

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Kariotis et al, 2022/Australia	objective to conduct a scoping review to explore the impact of electronic health records on information practices in mental health settings and explore how confidential information, data standardization, and therapeutic relationships are managed when using electronic health records in mental health settings	Data Mining in Electronic Health Records	As electronic health records are increasingly considered critical to modern healthcare systems, and healthcare decision makers must consider how medical record data can better reflect the complexity and sensitivity of information practices and workflows in the mental health context.

Kariotis, T. C., Prictor, M., Chang, S., & Gray, K. (2022). Impact of Electronic Health Records on Information Practices in Mental Health Contexts: Scoping Review. In *Journal of Medical Internet Research* (Vol. 24, Issue 5). JMIR Publications Inc. https://doi.org/10.2196/30405

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Sarwar et al, 2023/Australia	This study can serve as a primer for researchers to understand the use of EHRs for data mining and analysis purposes	Data Mining in Electronic Health Records	This review provides a comprehensive discussion of current methods for dealing with data types and challenges in health record EHR applications, which can be used as guidelines for future data mining studies

Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., & Cavedon, L. (2023). The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. In *ACM Computing Surveys* (Vol. 55, Issue 2). Association for Computing Machinery. https://doi.org/10.1145/3490234

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Jothi et al, 2015/Malasia	In this study, we reviewed the various works involved in this field in terms of method, algorithms and results	Data Mining in Electronic Health Records	This review article and consolidated the reviewed articles in line with the disciplines, model, tasks and methods.

Jothi, N., Rashid, N. A., & Husain, W. (2018). Data mining in healthcare - A review. *Procedia Computer Science*, *72*, 306-313. https://doi.org/10.1016/j.procs.2015.12.145

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Dubromel et al, 2020/France	The main objective was to identify the criteria already used to evaluate the organizational impact of health innovations	Data Mining in Electronic Health Records and Health Innovations	This study suggests the first steps in developing a validated MCDA method for assessing the organizational impact of healthcare innovations

Dubromel, A., Duvinage-Vonesch, M.-A., Geffroy, L., & Dussart, C. (2020). Organizational aspect in healthcare decision-making: a literature review. *Journal of Market Access & Health Policy*, *8*(1), 1810905. https://doi.org/10.1080/20016689.2020.1810905

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Pooja et al, 2019/India	Discover useful and important information effectively from the huge amount of data available through data mining techniques	Data Mining in Electronic Health Records	Review different predictive and descriptive data mining techniques that have been applied to voluminous health records to aid in the decision-making process

Pooja, H., & Jagadeesh M.P., P. (2019). A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records. *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 51-55. https://doi.org/10.1109/ICATIECE45860.2019.9063623

Author/Year/Country	Purpose/Delineation of Studies	Evaluation Focus	Study Findings
Islam M, et al, 2018/USA	In this study, the authors produced a review of the systematics on health analytics using data mining and big data.	Data Mining in Electronic Health Records	The use of human-generated data is prevalent, considering the wide adoption of the Electronic Medical Record in clinical care. However, data-based analysis from websites and social media has increased in recent years. The lack of prescriptive analysis in practice and the integration of domain expertise into the decision-making process emphasize the need for future research.

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare*, *6*(2), 54. https://doi.org/10.3390/healthcare6020054

5 CONCLUSIONS

This review showed clear evidence of the prospective value of electronic health record management and expert diagnostic support in hospitalization conditions, impossible to be only manually ascertained. Knowing the data mined, can, in addition to important clinical, care safety and economic advantages can offer improvements in quality of care across studies (**Wu W. et al, 2021**). Finally, the findings suggest in most cases, implementations of expert systems applications, interfaced as decision support with electronic medical records. In the meantime, the use of secondary data studies prevailed over primary data studies.

REFERENCES

Ara Shaikh, A., Nirmal Doss, A., Subramanian, M., Jain, V., Naved, M., & Khaja Mohiddin, Md. (2022). Major applications of data mining in medical. *Materials Today: Proceedings*, *56*, 2300-2304. https://doi.org/10.1016/j.matpr.2021.11.642

Drayton-Brooks, S. M., Gray, P. A., Turner, N. P., & Newland, J. A. (2020). The use of big data and data mining in nurse practitioner clinical education. *Journal of Professional Nursing*, *36*(6), 484-489. https://doi.org/10.1016/j.profnurs.2020.03.012

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, *125*, 37-46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

Dubromel, A., Duvinage-Vonesch, M.-A., Geffroy, L., & Dussart, C. (2020). Organizational aspect in healthcare decision-making: a literature review. *Journal of Market Access & Health Policy*, *8*(1), 1810905. https://doi.org/10.1080/20016689.2020.1810905

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare*, *6*(2), 54. https://doi.org/10.3390/healthcare6020054

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data mining in healthcare - A review. *Procedia Computer Science*, *72*, 306-313. https://doi.org/10.1016/j.procs.2015.12.145

Kariotis, T. C., Prictor, M., Chang, S., & Gray, K. (2022). Impact of Electronic Health Records on Information Practices in Mental Health Contexts: Scoping Review. *Journal of Medical Internet Research*, *24*(5), e30405. https://doi.org/10.2196/30405

N.P., J., & Aruna, R. (2022). Big data analytics in health care by data mining and classification techniques. *ICT Express*, *8*(2), 250-257. https://doi.org/10.1016/j.icte.2021.07.001

Pooja, H., & Jagadeesh M.P., P. (2019). A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records. *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 51-55. https://doi.org/10.1109/ICATIECE45860.2019.9063623

Reis, Z. S. N., Maia, T. A., Marcolino, M. S., Becerra-Posada, F., Novillo-Ortiz, D., & Ribeiro, A. L. P. (2017). Is There Evidence of Cost Benefits of Electronic Medical Records, Standards, or Interoperability in Hospital Information Systems? Overview of Systematic Reviews. *JMIR Medical Informatics*, *5*(3), e26. https://doi.org/10.2196/medinform.7400

Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., & Cavedon, L. (2023). The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Computing Surveys*, *55*(2), 1-40. https://doi.org/10.1145/3490234

Shakibaei Bonakdeh, E. (2022). Interpretation Bias Toward the Positive Impacts of Digital Interventions in Health Care. Comment on "Value of the Electronic Medical Record for Hospital Care: Update From the Literature." *Journal of Medical Internet Research*, *24*(3), e37208. https://doi.org/10.2196/37208

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering*, *2018*, 1-9. https://doi.org/10.1155/2018/4302425

Vimal, S., Rho, S., & Pelusi, D. (2023). Special Issue on Artificial Intelligence Empowered Big Data Analytical Patterns for Medical Applications. *Neural Processing Letters*, *55*(1), 1-1. https://doi.org/10.1007/s11063-023-11158-8

Wu, W.-T., Li, Y.-J., Feng, A.-Z., Li, L., Huang, T., Xu, A.-D., & Lyu, J. (2021). Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research*, *8*(1), 44. https://doi.org/10.1186/s40779-021-00338-z