Advances in Medicine and Engineering Interdisciplinary Research, 2024, Volume 2, Issue 4
DOI: 10.32629/ameir.v2i4.2813
ISSN Online: 2972-3825
ISSN Print: 2972-3833

Predictive Models in Health Based on Machine Learning

Javier Mora Pineda

1. Department of Cardiac, Vascular and Thoracic Surgery, Clínica Las Condes, Santiago, Chile.
2. ECMO Unit, Clínica Las Condes, Santiago, Chile.
3. Clinical Data Science Unit, Clínica Las Condes, Santiago, Chile.

Abstract: Finding causality in medicine is of great interest in research, in order to generate interventions that treat or cure the disease. Most classical statistical models allow association to be inferred, and only a few designs are able to demonstrate cause and effect with an adequate methodology and solid evidence. Evidence-based medicine supports its findings in models that go from a hypothesis to search for data to prove or rule it out. This also applies to the development of predictive models to be reliable and to produce impact in clinical practice. The large amount of data stored in electronic health records and greater computational power mean that machine learning techniques can play a preponderant role in the development of new predictive analysis and recognition of unknown patterns with these modern computational models. These models, along with changing the view from data to information, are increasingly being incorporated into daily clinical practice, providing greater precision and speed for supporting decision making. The intent of this review is to provide theoretical bases and evidence of how these modern computational techniques of machine learning have allowed to achieve better results and they are being widely used. This article will review the most relevant aspects of health data science in Latin America.

Key words: data sciences; machine learning; artificial intelligence; predictive models; clinical decision support systems; cardiac surgery; data

1. Introduction

When, in 1854, John Snow made a critical analysis of the behavior of the cholera pandemic in London, he never thought that he would lay the foundations of modern epidemiology. Snow, with an enlightened mind, not only associated many cause-effect bio-demographic variables to his investigation, but also plotted them on a map (Fig. 1). This made it possible to clearly visualize the distribution of cholera deaths around the nearest contaminated water pump in Broad Street and convinced the authorities to take action and begin to combat the pandemic efficiently [1].

Based on this model, other health data began to be explored in different scenarios, with the same logic of finding explanations and/or associations between different phenomena and health. The systematized recording of data and their subsequent analysis began to take on importance. Mathematical and statistical models are gradually being applied, making it possible to formulate hypotheses and test their veracity or reject them in relation to a problem.

It is also well known that in scientific health research, both in basic and clinical sciences, important findings and discoveries have occurred unexpectedly or unplanned in the context of a research model for another purpose. A classic

example is the discovery of penicillin. This could be analogous to what is known in data mining as "letting the data speak," i.e., using analytical data processing to provide new information (grouping or clustering, classification, distribution, etc.) that was not previously considered or not considered at all.

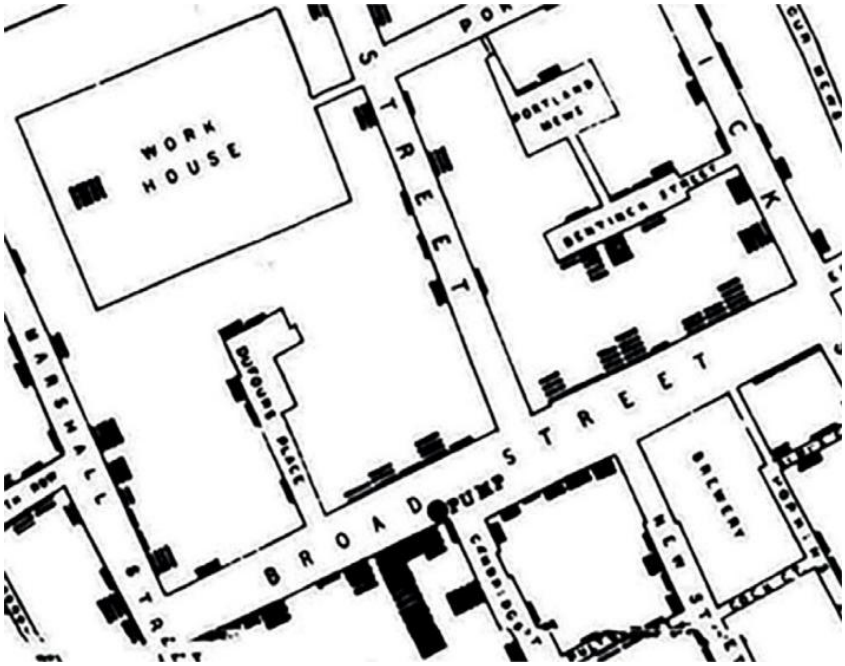

Figure 1. Map made by John Snow of cholera deaths in the Broad Street area. The pump is located at the intersection of Broad and Cambridge Street. The black bars correspond to deaths. The brewery (Brewery) and work house (Work House) are also noted. Source: Cerda J and Valdivia G1.

In classical biostatistics (as the methodological underpinning of evidence-based medicine), the essential model consists of setting out a hypothesis and an appropriate research method that best satisfies the evidence that can be obtained, and once this is established, to seek the data that will allow the model to be satisfied in order to answer the hypothesis or research question. Incidentally, there is also a high percentage of misinterpretation of results, such as statistical significance in some publications [2] (Fig. 2).

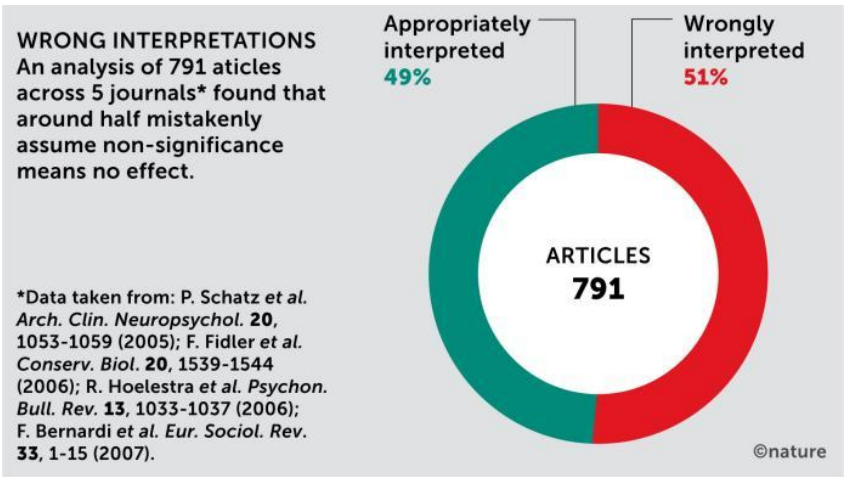

Figure 2. Inadequate interpretation of statistical significance in published articles. Source: Amrhein V. et al. [2]

The paradigm shift provided by machine learning and data science, especially in processing large amounts of data, is a new aspect that is different from biostatistics. Biostatistics applies computational models "from data" to obtain new information related to the research field; The definition or determination of the problem features (prediction or automatic

classification) that affect the sought answers is provided by the machine learning model itself, not just by researchers, thereby minimizing bias and achieving better results than traditional methods. These techniques have important applications in medicine and in the field of health in general; an important utility has been demonstrated in clinical decision support systems [3].

In machine learning, complete clinical records are inputs for learning algorithms. This data does not necessarily have to be structured or tabular, and can be input with images, free text, audio, video, time series, etc. The resulting models can subsequently be used to help healthcare professionals to pinpoint diagnoses for future new patients (new data). In this way, the examination and diagnosis of a patient can be faster, more accurate and reliable.

In this review, we intend to provide some theoretical basis and evidence of how these modern computational techniques of machine learning have allowed not only to reach better results, but also how and why they are being increasingly used in clinical practice with real-time or near-real processing, of the immense volume of data we obtain from different sources of clinical records, biomarkers, physiological parameters of patients and healthy people (with specific sensors or smart watches), allowing to predict health situations with increasing accuracy and speed, facilitating timely intervention; immediately for emergencies and preventively for regular check-ups.

2. Machine Learning versus Statistics

Based on general data science concepts discussed in a previous article in this journal [4] and in the book by researchers at the University of Chile, "A Look at the Data Age " [5], this review focuses on the development of the most commonly used models for the task of predicting a final outcome in the medical field. Of those concepts, it is appropriate to highlight Mitchel's concept of machine learning: "a computer program is said to learn from an experience E with respect to some type of task T with a performance measure P, if its performance on the task T as measured by P improves with the experience E" [6].

One way of expressing the methodological difference between statistics and machine learning is that the former involves testing hypotheses, while the latter involves the task of building knowledge from data and storing it in some form of computation, such as mathematical models, algorithms or any other computational form that can help determine patterns or predict results, and that can also be part of other models that require algorithms already trained for a specific purpose (transfer learning) (Fig. 3).

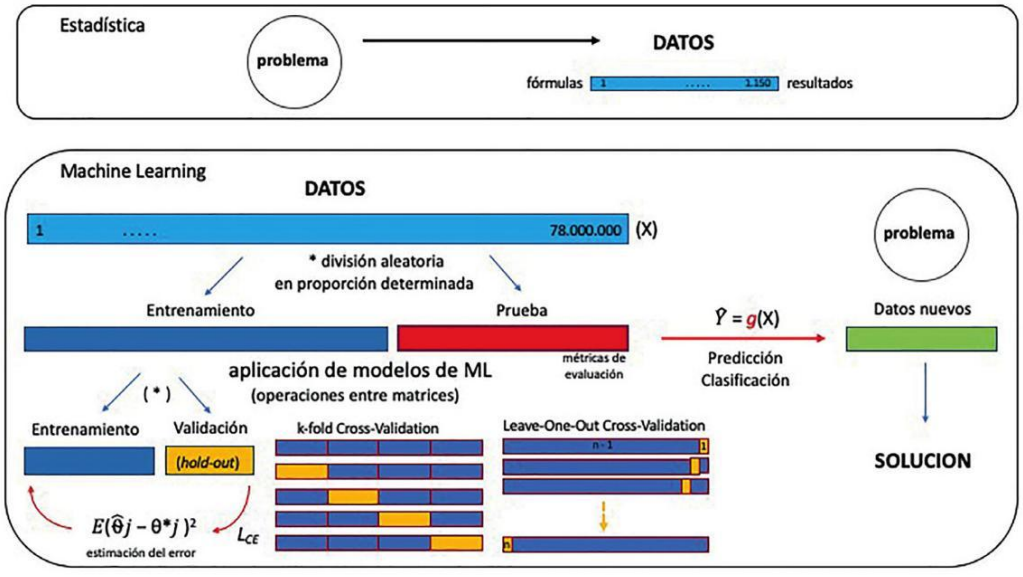

Figure 3. Diagram of the data paradigm in statistics and machine learning techniques.

It is, however, pertinent to make it clear that statistics is not exclusive to data science and machine learning, since a wide variety of statistical techniques are applied in these computational models, especially during the evaluation of the performance of each one and when comparing them with each other, in a dynamic and iterative way. In a classification task, for example, the result is a probability distribution of the various possibilities. In addition, the concept of data science includes statistics as one of its disciplines, as schematized in Fig. 4.

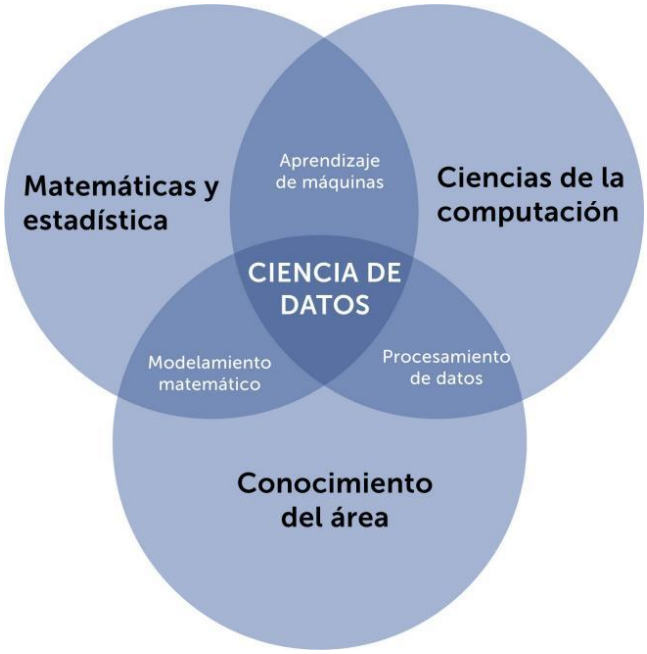

Figure 4. Diagram of the Data Science concept and the disciplines that bring it together. Used with permission of the authors. From "A Look at the Data Age", Jocelyn Dunstan, Alejandro Maass, Felipe Tobar, Editorial Universitaria 2022. ISBN 9789561128989 [5].

In the past, methods have been developed in parallel for both statistics and machine learning; four statisticians published in 1980 the book "Classification and Regression Trees", whose statistical techniques have been widely adapted by computer science researchers to improve classification performance and to make a procedure computationally well-organized [7].

In this context, it is interesting to analyze why these models "learn" from the data, without necessarily having to program or apply explicit instructions, formulas or rules step-by-step and to the entire data set (the data paradigm shift in medicine, enunciated in the introduction).

In a case-control study of traditional biostatistics, for example, data are recorded in both groups and then statistical formulas are applied to the total data set of each group to obtain a final result that is used to analyze the problem under study. A similar situation occurs in the branches of a prospective double-blind randomized trail, where statistical inferences are also made on the data set at one time.

In machine learning models, on the other hand, the data set is randomly divided into 3 groups: training, validation and test. The first group is the input of the algorithm (one of the algorithms used according to the task to be solved), with which it "learns" to solve the task (basically, it assigns certain values to certain constants according to the functions used); with the second group,l we evaluate how well it solves the task with those "learned" values (or in other words, how much error it makes in the task), so as to have the opportunity to adjust some parameters of the algorithm (fine-tuning) that can be run again with the training group to improve the metrics of the validation group; and finally, with the test group we see the results of how the algorithm will behave with new data in the future (generalization) (Fig. 3).

One of the main objectives of using machine learning methods concerns classification and prediction. The terms classification and prediction have been used interchangeably [8], but, in general, prediction is adopted for a "binary" outcome (present or absent, yes or no, is or is not, etc.) while classification is used to determine 1 outcome within a set of 3 or more possible outcomes depending on the task to be solved. Classification or prediction can be used to extract models describing classes or groups of relevant data versus others or to predict future trend of new data from the same problem [9]. Both tasks have the ability to generalize over a data set, which means the ability to identify new results with new data from the previous data used in training the algorithm. In turn, it is important to note that, when solving classification and prediction problems, they are sensitive to missing or missing data from a repository [10].

3. Applications of Predictive Models with Machine Learning

The main machine learning predictive models applied in medicine have been: neural networks (both fully connected, convolutional and recurrent networks), support vector machines, decision trees, random forest, linear regressions, Bayesian models (naive Bayes), and nearest neighbors, mainly. In particular, convolutional networks, proposed by Yann LeCun in 1989, are machine learning algorithms based on the functioning of the visual cortex of the animal eye, which allows the computer to "see" [11, 12], with a multilayer neural network architecture in which the image is divided into receptive fields or segments of the image that pass through convolutional layers (mathematical operation that makes the integral of the product of two functions, or signals, in a third function). This allows it to extract different characteristics from the initial image in each layer, in order to classify it later. In medicine, this technique has made it possible to process various types of images to predict a diagnosis, from biopsies to X-rays, scans, magnetic resonance imaging and photos of different pathologies.

The various machine learning approaches are based on the identification of strong data associations, but free of a theoretical foundation (so-called "black box" models). The confusion makes it a substantial leap in causal inference to identify modifiable factors that may actually have an impact on altering results. As is well known, association does not imply causation. Many predictions from these models can be highly accurate, especially in cases where the likely outcome is already obvious to the clinician. The last mile of clinical implementation ultimately becomes the truly critical task of predicting events, requiring early intervention to influence nursing decisions, outcomes, and health outcomes [11]. Prediction techniques using machine learning increasingly offer decision support tools for risk management and early warning [12].

In neuroscience, Liu, from Taiwan University Hospital, developed a model for predicting brain death using neural networks called EANN AAN [13]. Rughani compared predictive neural network models used in neurosurgery with the ability to predict survival with linear regression models and with that of surgeons, obtaining significantly better results [14]. Güler developed a predictive model with neural networks to assign severity in traumatic brain injury, with an accuracy of 91% [15]. Decision trees and linear regression have been used to predict low severity outcomes based on a dichotomized Glagow scale with up to 80% accuracy [16].

In psychiatry, a predictive model based on clinical data using deep neural networks had a result of 0.73 and 0.67 area under the ROC curve for predicting generalized anxiety disorder and major depressive disorder, respectively [17]. Some of the most discussed issues in machine learning, such as bias and fairness, data cleaning, and model interpretation, may be unfamiliar in the use of neuroimaging-based predictive models in psychiatry. Similarly, diagnostic research and brain-based feature selection for psychiatric intervention are modern issues that have been found to be appropriate to address with predictive machine learning models [18].

In genetics and molecular biology, researchers have developed and evaluated several new predictive models of

evolutionary information in DNA transcription using Support Vector Machine techniques to determine the binding sites of genetic material (specifically, "binding residue sequences") of DNA and RNA. Their findings showed that these classifiers had 77.3% sensitivity and 79.3% specificity for DNA binding residue prediction, and 71.6% sensitivity and 78.7% specificity for RNA binding site predictions, important for estimating possible genetic mutations. This finding demonstrated that the Support Vector Machine classifier was better and more accurate than existing models for predicting molecular binding sites in gene transcription [19].

In ophthalmology, one study compared three machine learning models for predicting the need for surgical intervention in primary open-angle glaucoma using clinical records, with multivariate logistic regression in this work being the most effective for discriminating patients with progressive disease requiring surgery (area under the ROC curve 0.67) [20].

In clinical management, applications to reduce the no-show rate of a patient to a medical appointment (consultation or examination) by predicting their scheduling behavior have been of great importance, since this absenteeism amounts to between 10 to 20% worldwide, and up to 30% in some hospitals in China [21]. With techniques based on deep neural networks (deep learning or representation learning), it has been possible to reduce the no-show patient rate by 6 to 14% and increase effective patient attendance, optimizing the use of resources and time spent in healthcare centers. These predictive models are fed by automated communication techniques of text messages, delivering in turn adequate and accurate information to patients in a timely manner to avoid delays and absenteeism, and can also automate the rescheduling of free spaces generated by patients or predicted by the system. In Chile, important research has been developed in this area by the group of the Mathematical Modeling Center of the University of Chile, studying and applying these technologies in three public health centers [22]; also, through solutions based on neural networks, a Chilean start-up has managed to improve access to health care in public and private hospitals by optimizing the management of scheduling [23].

The problem of avoiding unplanned rehospitalization after patient discharge has been studied using neural networks. In 2007, U.S. Medicare reported 17% of such hospital readmissions within 30 days of discharge, with an estimated 76% potentially preventable, representing $15 billion in Medicare expenditures at the time, and also as a measure of quality of care [24]. These are some of the reasons why it is interesting to find methods that can predict the risk of readmission. The 2020 Clinica Las Condes published a study that for the first time in the literature reported results of this task with a model, in Spanish, of highly unstructured clinical data; the work yielded an area under the ROC curve of 0.76 for prediction, using the neural network model called Long Short Term Memory (LSTM) [26]. Demographic data, reasons for consultation, procedures, diagnoses, prescriptions, and clinical notes from 9 years of records were used as inputs [25]. This result was very similar to that published at the time with clinical records in English [27], which gives it great value and has motivated us to continue working along these lines, seeking to improve the metrics of the model and its testing in sub-areas of the hospital environment.

In 2021, a library of the Python programming language (the most widely used in data science and machine learning) was specially developed for predictive models in healthcare: PyHealth [28]. It consists of special modules for data preprocessing, predictive modeling (with 30 models available) and evaluation, in such a way as to diagram complex medical data models in a simple way and with few lines of code.

4. Future Projections

Research in computer science and health data science continues to improve the accuracy of clinical predictions, but even a perfectly fine-tuned prediction model may not translate into better clinical care. Even a very accurate prediction of the outcome of a health problem does not tell you what to do or how to do it if you want to change that outcome; in fact,

you cannot even assume that it is possible to change the predicted outcomes.

Predictive models based on machine learning currently provide great support to the clinical knowledge and experience of professionals. To reduce subjectivity, many expert systems have been created to encode and combine medical knowledge. Predictive methods can be integrated into these systems, contributing to reduce bias and subjectivity, while potentially providing new medical knowledge in different areas of medicine, or in different geographic areas where it is applied (fitting a machine learning model with "local" data can be more efficient than applying a formula made with population from other countries). Therefore, the increasingly accurate prediction of a patient's possible outcomes during the different stages of the care process poses a constant and evolving challenge in healthcare. In addition, the ethical considerations of the different applications of artificial intelligence and machine learning are currently the subject of significant discussion and analysis in the specialized scientific community, an issue that also extends to the general population's perception of these "new" technologies applied to healthcare.

Performing outcome prediction studies of clinical problems is very relevant nowadays because they can help the medical team to make more accurate decisions with machine learning techniques that are easy to implement and low cost. The evolution of these models and the emergence of new ones in the future will undoubtedly provide increasingly accurate and dynamic predictions in daily clinical practice.

With robustness and scalability in mind, the PyHealth library is constantly evolving to achieve best practices for development, testing, and interactive integration to enable the increasing application of algorithms and other widely used Python libraries to clinical data [28].

5. Conclusions

Prediction is nothing new in medicine. From risk scores to guide anticoagulation (CHADS2) and high cholesterol drug use (ASCVD), to risk stratification of intensive care unit patients (APACHE), among many others, data-driven clinical predictions are routine in clinical practice. Today, in combination with modern machine learning techniques, various sources of clinical data allow us to quickly or in real time generate prediction models for thousands of similar clinical tasks or problems. In addition, it has been possible to include in the development of these models, different types of unstructured data that previously were not possible to process as part of a predictive model with traditional methods (e.g., large amounts of free text widely used in clinical records, direct images from their capture, computer vision, etc.). In healthcare, time to diagnosis is a crucial factor in patient prognosis. The potential applicability of this approach to the use of clinical data is substantial, both for critical early warning systems and for high-precision diagnostic imaging, to optimize clinical appointment management, among others.

Although predictive algorithms obviously cannot eliminate uncertainty in medical decision-making, they can improve or optimize the allocation of resources in medical care, both human and physical. For example, the Pulmonary Embolism Severity Index (PESI) is a widely validated predictive model of the risk of death from this condition [29] and others prioritize patients awaiting liver transplantation in an appropriate (if not fair) manner by means of the Model for End-stage Liver Disease (MELD) score. Early warning systems that would have taken years to develop using traditional techniques can now be quickly deployed and optimized using real-world data and on an ongoing basis. Similarly, deep learning neural networks are able to routinely recognize pathological images with a high accuracy previously thought impossible.

I believe it is also necessary to emphasize that these models, like so many others in the different areas of machine learning, are not infallible and, like any human, are wrong in their results a percentage of the time. In fact, the metrics used to train and evaluate the performance of an algorithm under development are mainly based on "minimizing error" (such as Root Mean Square Error (RMSE), cross-entropy, among others). In contemporary transformation, we are experiencing the

process of integrating more and more of these technologies into daily clinical practice, and people tend to expect them to be error free simply because they are executed by modern so-called intelligent machines or computers, which may be mistakenly thought to be not 100% accurate.

It seems irrelevant to debate whether predictive machine learning models will become "smarter" than healthcare professionals themselves. It is indisputable that in healthcare, as in all other industries, domain experts are a fundamental and irreplaceable link in the data science working model, making sense of and modulating both the creation and clinical use of these algorithms (Fig. 4). Undoubtedly, the final result of these tools can increasingly resemble or approach "human behavior", integrating several predictive models simultaneously (even choosing which ones to use in a given case) and in a very short time, to make complex decisions (of diagnosis or treatment) almost automatically or in real time. However, we must not forget that in their most basic logic they are nothing more than mathematical operations of matrices and high-dimensional tensors that do not think, do not feel, and do not have the intangible human interaction and virtues ("clinical eye") that no machine has replaced or will replace, from the work of Alan Turing to the present day and in the future.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Cerda J, Valdivia G. John Snow. La epidemia de cólera y el nacimiento de la epidemiología moderna. [John Snow, the cholera epidemic and the foundation of modern epidemiology]. *Rev Chil Infectol*. 2007;24(4):331-334. doi: 10.4067/S0716-10182007000400014

[2] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature.* 2019;567(7748):305-307. doi: 10.1038/ d41586-019-00857-9

[3] Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J Med Syst*. 2017;41(4):69. doi: 10.1007/s10916-017-0715-6

[4] Mora J. Proyecciones de la ciencia de datos en la cirugía cardíaca. [Projections of data science in cardiac surgery]. *Rev Med Clin Condes*. 2022;33(3):294-306. doi: 10.1016/j.rmclc.2022.05.007

[5] Dunstan J, Maass A, Tobar F. Una Mirada a la Era de los Datos. Ed. Universitaria 2022.

[6] Mitchell TM. *Machine Learning*. McGraw-Hill. 1997.

[7] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Routledge; 2017.

[8] Carreno A, Inza I, Lozano JA. Eventos raros, anomalıas y novedades vistas desde el paraguas de la clasificacion supervisada. IX Simposio de Teoría y Aplicaciones de la Minería de Datos. 2018:925-930. https://sci2s.ugr.es/caepia18/proceedings/docs/CAEPIA2018_paper_192.pdf

[9] de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022;5(1):2. doi: 10.1038/s41746-021-00549-7.

[10] Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J Med Syst*. 2017;41(4):69. doi: 10.1007/s10916-017-0715-6

[11] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Back propagation applied to handwritten zip code recognition. *Neural Comput*. 1989;1(4):541-551.

[12] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib, Michael A. (ed.). The handbook of brain theory and neural networks (Second ed.). *The MIT press*. 1995:276-278.

[13] Liu Q, Cui X, Abbod MF, Huang SJ, Han YY, Shieh JS. Brain death prediction based on ensembled artificial neural networks in neurosurgical intensive care unit. *J Taiwan Inst Chem Eng*. 2011;42(1):97-107. https://doi.org/10.1016/j.jtice.2010.05.00

[14] Rughani AI, Dumont TM, Lu Z, Bongard J, Horgan MA, Penar PL, Tranmer BI. Use of an artificial neural network to predict head injury outcome. *J Neurosurg*. 2010;113(3):585-90. doi: 10.3171/2009.11.JNS09857

[15] Güler , Gökçil Z, Gülbandilar E. Evaluating of traumatic brain injuries using artificial neural networks. *Expert Syst. Appl.* 2009;36(7):10424-10427. Doi: 10.1016/j.eswa.2009.01.036

[16] Low D, Kuralmani V, Ng SK, Lee KK, Ng I, Ang BT. Prediction of outcome utilizing both physiological and biochemical parameters in severe head injury. *J Neurotrauma*. 2009;26(8):1177-82. doi: 10.1089/ neu.2008.0841

[17] Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci Rep*. 2021;11(1):1980. doi: 10.1038/s41598-021-81368-4

[18] Tejavibulya L, Rolison M, Gao S, Liang Q, Peterson H, Dadashkarimi J, et al. Predicting the future of neuroimaging predictive models in mental health. *Mol Psychiatry*. 2022. doi: 10.1038/s41380-022-01635-2

[19] Ma X, Wu J, Xue X. Identification of DNA-binding proteins using support vector machine with sequence information. *Comput Math Methods Med*. 2013;2013:524502. doi: 10.1155/2013/524502

[20] Baxter SL, Marks C, Kuo TT, Ohno-Machado L, Weinreb RN. Machine Learning-Based Predictive Modeling of Surgical Intervention in Glaucoma Using Systemic Data From Electronic Health Records. *Am J Ophthalmol*. 2019;208:30-40. doi: 10.1016/j.ajo.2019.07.005

[21] Fan G, Deng Z, Ye Q, Wang B. Machine learning-based prediction models for patients no-show in online outpatient appointments. *Data Sci Manage*. 2021;2:45-52. doi: 10.1016/j.dsm.2021.06.002

[22] Ramírez H, Dunstan J, Montenegro H. Soluciones tecnológicas, basadas en técnicas matemáticas avanzadas de aprendizaje de máquinas. FONDEF IDeA I+D ID19I10271

[23] Modelo de IA a través de chatbot logra mejorar la agenda de citas médicas en Chile. Salud Digital, Fundacion Carlos Slim. https://saluddigital.com/es/noticias/modelo-de-ia-a-traves-dechatbot-logra-mejorar-la-agenda-de-citas-medicas-en-chile/

[24] Covacevich Stipicich T. Exploring representations of ICD codes for patient readmission prediction (Doctoral dissertation, Pontificia Universidad Catolica de Chile (Chile)). 2021. https://repositorio.uc.cl/xmlui/handle/11534/58535

[25] Fierro C, Pérez J, Mora J. Predicting unplanned read missions with highly unstructured data. 2020. *Workshop paper at AI4AH*, ICLR 2020. doi: 10.48550/arXiv.2003.11622

[26] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. doi: 10.1162/neco.1997.9.8.1735.

[27] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. doi: 10.1038/s41746-018-0029-1

[28] Zhao Y, Qiao Z, Xiao C, Glass L, Sun J. PyHealth: A Python Library for Health Predictive Models. arXiv:2101.04209 https://doi.org/10.48550/ arXiv.2101.04209

[29] Wells PS. Integrated strategies for the diagnosis of venous thromboembolism. *J Thromb Haemost*. 2007;5 Suppl 1:41-50. doi: 10.1111/j.1538-7836.2007.02493.x