The Future of Medicine and the Logic of Data Management – Data Discrimination Problems

Ozar P. Mintser[a]

[a]**Shupyk National Healthcare University of Ukraine**

ABSTRACT

Considered problems of medical data management. It is emphasized that this process has the greatest potential to ensure accurate and cost-effective patient care, as well as knowledge transfer during medical education. Analyzing the main areas of professional application of data: integration of big data, working with data of various types (from batch to stream) and their transformation for further use, we note that all the listed areas do not have a clear interpretation and recommendations for their use.

The purpose of the study was to substantiate the strategy of a safe and effective medical data management system and the logic of creating completely open data systems. This strategy will allow streamlining the data flows of instrumental and laboratory studies and ensure the delivery of their big data directly to medical institutions or medical research centers.

Conclusions: 1. There is a need for a comprehensive real-time medical data management system that will allow physicians, patients, and external users to enter their medical and lifestyle data into the system. 2. The inclusion of big data analytics will help to better predict or diagnose diseases, and accordingly help in the development of an effective plan for the prevention of complications and treatment of the disease. 3. Scaling communication with each patient in real time is possible only with the help of artificial intelligence. Manual intervention simply cannot serve thousands of users at the same time, each in their own way, addressing each individual directly. This conclusion should be especially emphasized when creating a perfect model of family medicine organization. 4. Transferring information is at the heart of developmental biology, and thus it is imperative that we can form a logical and structured approach to the healthcare language. If, as it appears, information theory has much offer to biology, further advances will depend on its integration. Given that these fields share many terms with developmental biology, effective collaboration may necessitate redefining the meaning attached to signaling, communication and information, in the context of the biology and medicine. Biosemiotics is inherently concerned with the language and rules of signals and codes in biological systems. It combines many ideas from diverse areas including systems theory, information theory and linguistics and may offer us a new perspective on the classification and meaning of biological and medical signaling.

KEYWORDS

Big data, medical data discrimination, data analytics, eHealth, electronic medical records, health care, medical information management, metadata, metatechnologies, infonomics, internet of things

2. Introduction

Health data management is tasked not only with organizing medical data, but also with integrating and analyzing it to make patient care more efficient and to obtain information that can improve medical outcomes while protecting data privacy and security.

As medical practices continue to use more sophisticated electronic health record systems, the need for effective health data management is increasing. Health data management is the process used to record, store, protect and analyze the data you receive from various sources. Effective healthcare data management allows healthcare professionals to develop a comprehensive view of a patient's condition.

Effective data management practices are vital to keeping up with the vast amount of data that healthcare facilities generate each month. They also help to deliver more personalized care, more effective communication and regulatory compliance.

Good data management benefits the patient, providers and insurers, and has far-reaching implications for the health of the entire population. In particular, health data management helps to: Create a complete picture of a patient's health by integrating data collected from various sources; increase patient engagement with predictive analytics and faster diagnosis based on available data; improve health problem tracking and predict complications; provide data for making effective business decisions that will increase the efficiency of the medical institution; and, finally, adhere to standardized care plans to improve disease management.

The purpose of the study was to substantiate the strategy of a safe and effective medical data management system and the logic of creating completely open data systems. This strategy will allow streamlining the data flows of instrumental and laboratory studies and ensure the delivery of their big data directly to medical institutions or medical research centers.

3. Results

Our main focus in this publication is the problem of data discrimination, also called "algorithmic discrimination".[1,2] The last is defined as "the bias that occurs when predefined types of data or data sources are intentionally or unintentionally treated differently than others." The use of big data can make discrimination more common.[3] Big data analytics is defined as a set of advanced digital technologies (such as data mining, neural networks, deep learning, profiling, automated decision making, and scoring systems) designed to analyze large data sets in order to identify patterns, trends, and associations related to the behavior of patients, play an increasingly important role in our everyday life - automatic registration for doctor's appointments, electronic registration of those wishing to learn to ensure continuous professional development. All of these problems are influenced by computers and algorithms, not humans. Thus, data analysis technologies increasingly tap into people's sensitive personal characteristics, their daily activities, and their future capabilities. Therefore, it is not surprising that today big data technologies and their applications are carefully studied in order to analyze and understand the new ethical and social problems of big data, especially related to the privacy and anonymity of data, informed consent, epistemological problems (first of all, the question of the validity of the obtained knowledge that is not subject to any doubts; from this point of view, the concept of the norm and conformity to this norm with the mandatory distinction between the actually existing and the proper) and more conceptual problems, such as changing the concept of personal identity through profiling or analysis, can be considered key "datafication", "information" society, etc.[4]

Another reason for data discrimination is the processing of information. Because the data scientist needs to translate the problem into formal computer code, the choice of target variable and class labels is a subjective process. Another algorithmic cause of discrimination is related to data bias in the model. For the development of automation of the intelligent data analysis model, appropriate sets for training are required, since the training to make classifications is carried out taking into account the given examples.

Accordingly, if the training data is contaminated with discriminatory or biased cases, the system will consider them valid examples from which to learn and reproduce the discrimination in its own results. This contamination can occur due to historically biased datasets or due to manual assignment of class labels by data collectors. An additional problem with training data can be data collection bias or sampling bias. Data collection bias can be the under-representation of certain groups and/or protected classes in a data set, which may result in unfair or unequal treatment, or the over-representation in a data set. In this way, "disproportionate attention is paid to a protected class group, and increased attention, in turn, may lead to a greater likelihood of observing a targeted violation." In this context, the phenomenon of "overfitting" was mentioned, where "models can become too specialized or specific to the data used for training", and instead of finding the best possible decision rule, they simply learn the most appropriate rule. The training data thus causes it to shift. Another possible algorithmic cause of discriminatory results is proxies for protected characteristics (e.g., race, gender, age). These include the creation of a negative vicious circle where certain inputs in the data set cause statistical deviations that are learned and fixed by the algorithm in a self-fulfilling cycle of cause and effect.

"Error of omission" is another form of cost function misspecification that occurs when terms that cause discrimination are ignored or not accounted for in the model. Simply put, this means that the model does not take into account the differences in how the algorithm classifies protected and unprotected classes.

Data processing and reasoning techniques are often biased toward "middle" or dominant groups. This is especially observed when conducting surveys. The whole concept of testing reinforces this, as average results (such as the frequency with which a certain effect is observed) are applied to the rest of the audience. Even when testing against carefully segmented mailing lists, the ultimate definition of "success" is determined by data decisions based on averages. Working with averages prioritizes a generalized "ideal" client that, at best, can only reflect some of the user's preferences.

In today's strategy, data decision-making algorithms must be constantly self-learning: thanks to the use of artificial intelligence (AI), it is possible to react and interact with each individual visitor in real time. We emphasize that the ability to scale communication with each person in real time is possible only with the help of AI. Manual intervention simply cannot serve thousands of users at the same time, each in their own way, addressing each individual directly. This conclusion should be taken into account when creating a perfect model of family medicine organization.

An important aspect of this chapter is the analysis of security opportunities for fair data mining. Many papers describe algorithmic decision-making as a “black box” system in which the algorithm's inputs and outputs are visible, but the internal process remains unknown, leading to a lack of transparency about the methods and logic behind evaluation and assessment. Predictive systems are becoming especially important. The reasons for the opacity of automated decision-making are numerous: first, algorithms can use huge and highly complex data sets that cannot be interpreted by researchers, who often lack the necessary computer science knowledge to understand algorithmic processes; second, automated decision-making may inherently surpass human understanding because algorithms do not use the theories or contexts that exist in normal human decision-making; and finally, algorithmic processes of firms or companies may be subject to intellectual property rights or subject to trade secret provisions.[2,5] If there is no transparent information about how the algorithms and processes work, it is almost impossible to evaluate the fairness of the algorithms or to detect discriminatory patterns in the system.[5]

Human bias is identified as another major obstacle to fair data collection, and human subjectivity underlies the design of data mining algorithms, as decisions about which attributes to consider and which to ignore are subject to human interpretation and will inevitably reflect implicit or explicit meanings of their attributes.

Thus, there is quite a large number of reasons for data discrimination. Their correct analysis does not yet have a corresponding algorithm. On the other hand, communication is at the heart of developmental

biology, and it is therefore essential that we can develop a logical and structured biomedical approach to the language of health care. We have proposed the use of the ideas of systemic biomedicine, in particular the logic of signaling principles for timely warning about incorrect use of data.

Conventionally, the study of informational signaling has focused on the linear, and is concerned with *components* of pathways, ostensibly neglecting the context and networks within which these components function. With the advent of systems biomedicine, the conceptualization of cell signaling has evolved from the relatively simple linear cascades of previous decades, towards an appreciation of signaling as interplay between highly complex and context-dependent modules of activity.[6] These larger and more complex systems have necessitated bioinformatics, the mathematical modelling of cell signaling networks as systems.

Modeling of signal networks enables verification of theoretical molecular mechanisms, highlights 'molecular hubs' and allows an appreciation of pathways, in the context of other operational networks.[7] Crucially, the latter confers an appreciation of the emergent properties of positive or negative-feedback signaling networks, namely biostability and ultra-sensitivity, or adaptation and desensitization, respectively. Remarkably, modeling of the non-linear dynamics of cell signaling networks also reveals striking similarities between signaling networks of the cell and telecommunications.

Cognitive neuroscience was an early adopter of mathematical modelling, following from the recognition that human consciousness represents an emergent property of networks rather than a tangible property of signal transduction itself.[8] Indeed, recent work modeling signal transmission in neural networks revealed that the 'transmitting' neuron transmits a signal as a modulation of delay time, while the 'receiving' cell 'decodes' the signal by tracking the delay time, in a striking resemblance to the principles of spread spectrum technique, employed in wireless communications.[9] Modeling networks of cell signaling in development has furthered the appreciation of such parallels.

Biological and medical semiotics is inherently concerned with the language and rules of signals and codes in biological and medical systems. It combines many ideas from different fields, including systems theory, information theory, and linguistics. Accordingly, biomedical semiotics can offer for us a new perspective on both the classification and meaning of biological and medical signaling, and on errors associated with data discrimination.[10,11]

3. Conclusions

1. There is a need for a comprehensive real-time medical data management system that will allow physicians, patients, and external users to enter their medical and lifestyle data into the system.

2. The inclusion of big data analytics will help to better predict or diagnose diseases and predict diseases, and accordingly help in the development of an effective plan for the prevention of complications and treatment of the disease.

3. Scaling communication with each patient in real time is possible only with the help of artificial intelligence. Manual intervention simply cannot serve thousands of users at the same time, each in their own way, addressing each individual directly. This conclusion should be especially emphasized when creating a perfect model of family medicine organization.

4. Transferring information is at the heart of developmental biology, and thus it is imperative that we can form a logical and structured approach to the healthcare language. If, as it appears, information theory has much offer to biology, further advances will depend on its integration. Given that these fields share many terms with developmental biology, effective collaboration may necessitate re-defining the meaning attached to signaling, communication and information, in the context of the biology and medicine. Biosemiotics is inherently concerned with the language and rules of signals and codes in biological systems. It combines many ideas from diverse areas including systems theory, information theory and linguistics and may offer us a new perspective on the classification and meaning of biological and medical signaling.

References and notes

[1] Balsa AI, McGuire TG, Meredith LS. Testing for statistical discrimination in Health Care. *Health Services Research.* 2005;40(1):227–252. doi: 10.1111/j.1475-6773.2005.00351.x

[2] Nong P, Williamson A, Anthony D, Platt J, Kardia S. Discrimination, trust, and withholding information from providers: Implications for missing data and inequity. *SSM – Population Health*. 2022;18:101092. doi: 10.1016/j.ssmph.2022.101092

[3] Favaretto M, De Clercq E, Elger BS. Big Data and discrimination: Perils, promises and solutions. A systematic review. *Journal of Big Data.* 2019;6(1). doi: 10.1186/s40537-019-0177-4

[4] Rahm E, Hong HD. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering.* 2000;(23):3–13.

[5] Zarsky T. The trouble with algorithmic decisions. *Science, Technology, & Human Values*. 2015;41(1):118–132. doi: 10.1177/0162243915605575

[6] Jordan JD, Landau EM, Iyengar R. Signaling networks: the origins of cellular multitasking. *Cell*. 2000;103(2):193–200. doi: 10.1016/s0092-8674(00)00112-4

[7] Eungdamrong NJ. Modeling Cell Signaling Networks. *Biology of the Cell*. 2004. doi: 10.1016/s0248-4900(04)00077–2

[8] Capra F. The Hidden Connections: A Science for Sustainable Living. London: HarperCollins; 2002.

[9] Xu M, Wu F, Leung H. A biologically motivated signal transmission approach based on stochastic delay differential equation. *Chaos: An Interdisciplinary Journal of Nonlinear Science.* 2009;19(3):033135. doi: 10.1063/1.3227642

[10] George BJ, Brown AW, Allison DB. Errors in statistical analysis and questionable randomization lead to unreliable conclusions. *Journal of Paramedic Science.* 2015;6:153–154.

[11] Platt CC, Nicholls C, Brookes C, Wood I. Classification of cell signalling in tissue development. *Cell Communication & Adhesion.* 2011;18(1-2):9–17. doi: 10.3109/15419061.2011.586755