VIEWPOINT

Unintended Consequences of Machine Learning in Medicine

Federico Cabitza, PhD
Department of Informatics, University of Milano-Bicocca, Milan, Italy; and IRCCS Istituto Ortopedico Galeazzi, Milan, Italy.

Raffaele Rasoini, MD
Centro Studi Medicina Avanzata, Florence, Italy.

Gian Franco Gensini, MD
Centro Studi Medicina Avanzata, Florence, Italy.

Over the past decade, machine learning techniques have made substantial advances in many domains. In health care, global interest in the potential of machine learning has increased; for example, a deep learning algorithm has shown high accuracy in detecting diabetic retinopathy.[1] There have been suggestions that machine learning will drive changes in health care within a few years, specifically in medical disciplines that require more accurate prognostic models (eg, oncology) and those based on pattern recognition (eg, radiology and pathology).

However, comparative studies on the effectiveness of machine learning–based decision support systems (ML-DSS) in medicine are lacking, especially regarding the effects on health outcomes. Moreover, the introduction of new technologies in health care has not always been straightforward or without unintended and adverse effects.[2] In this Viewpoint we consider the potential unintended consequences that may result from the application of ML-DSS in clinical practice.

Reducing the Skills of Physicians

A major issue related to incorporation of ML-DSS in medicine could be overreliance on the capabilities of automation. Although the phenomenon of overreliance on technology could be tempting to users in the short term for the convenience and efficiency of automated aids, in the long term these tools can lead to the related phenomenon of deskilling[3] (ie, the reduction of the level of skill required to complete a task when some or all components of the task are partly automated, and which may cause serious disruptions of performance or inefficiencies whenever technology fails or breaks down). This process can affect physicians' ability to derive informed opinions on the basis of detectable signs, symptoms, and available data.

For example, in a study of 50 mammogram readers, there was a 14% decrease in diagnostic sensitivity when more discriminating readers were presented with challenging images marked by computer-aided detection.[4] Another study of 30 internal medicine residents showed that the residents exhibited a decrease in diagnostic accuracy (from 57% to 48%) when electrocardiograms were annotated with inaccurate computer-aided diagnoses.[5] Further research is needed to better understand whether the overreliance on ML-DSS that could outperform or perform as well as human observers could also cause a subtle loss of self-confidence and affect the willingness of a physician to provide a definitive interpretation or diagnosis.

Focus on Text and the Demise of Context

Machine learning technologies also can lead to focusing more on what can be rendered as text (ie, data) at the expense of other elements that are more difficult or impossible to easily describe. Relying on ML-DSS requires considering digital data as reliable and complete representations of the phenomena that these data are supposed to render in a discrete and trustworthy form. This may be a problem when the clinical context is not represented, particularly if physicians lose awareness of the existence of clinical elements that are not included in the clinical record.

Such lack of information may lead to partial or misleading interpretations of ML-DSS diagnostics and therapeutic or prognostic outputs. It also could lead to reduced interest in and decreased ability to perform holistic evaluations of patients, with loss of valuable and irreducible aspects of the human experience such as psychological, relational, social, and organizational issues. These factors may not be incorporated into any ML-DSS because of their qualitative and complex nature, yet are fundamental to individualized care beyond diagnostic and therapeutic categories.

An example in which context mattered and lack of its inclusion resulted in a technically valid but misleading machine learning prognostic model was the use of mortality risk prediction to make decisions about whether to provide treatment on an inpatient or outpatient basis for 14 199 patients with pneumonia.[6] In that setting, an ML-DSS suggested considering patients with pneumonia and asthma to be at a lower risk of death from pneumonia than patients with pneumonia but without asthma. This indication surprised the researchers involved, who nevertheless ruled out that asthma could be a protective factor in patients with pneumonia. However, machine learning models do not apply explicit rules to the data they are provided, but rather identify subtle patterns within those data.

There were 2 causes for the algorithm being correct, but producing a counterintuitive and dangerous output. First, at the hospitals hosting this study, patients with a history of asthma who presented with pneumonia were usually admitted directly to intensive care units to prevent complications; this led to patients with pneumonia and asthma having better outcomes than patients diagnosed with pneumonia and without a history of asthma, with an approximately 50% mortality risk reduction (with mortality rates of 5.4% vs 11.3%, respectively). Second, this contextual information could not be included in the ML-DSS, and thus the algorithm "correctly misinterpreted" the presence of asthma as a protective variable. Failing to include difficult to represent factors into medical decision making may lead to other similar contextual errors, and overreliance on ML-DSS may enhance the odds of the occurrence of these types of errors when contextual factors cannot be easily integrated.

Corresponding Author: Federico Cabitza, PhD, University of Milano-Bicocca, 1 Piazza dell'Ateneo Nuovo, Milan, Italy 20126 (fcabitza@gmail.com).

Intrinsic Uncertainty in Medicine

Machine learning–based decision support systems bind empirical data to categorical interpretation. Potential unintended consequences arising from this approach may be related to the formalization into a decision model of the mapping between the physical signs that a physician can evaluate and their "right" class as identified by observers. In medical practice, observers often do not agree with each other about diagnostic findings and outcome evaluation. This observer variability is related not only to interpretive deficiencies, but also to an intrinsic ambiguity in the observed phenomena.[7]

However, the intrinsic uncertainty of medical observations and interpretations that are part of input to "optimize" machine learning models is not usually considered. As a result, the extent that reliability and accuracy of machine learning performance is affected by observer variability can be underrated; this has been shown to negatively affect the performance of the most common machine learning models. For example, interobserver variability in the identification and enumeration of fluorescently stained circulating tumor cells was observed to undermine the performance of ML-DSS supporting this classification task.[8]

Users and designers of ML-DSS need to be aware of the inevitable intrinsic uncertainties that are deeply embedded in medical science. Further research should be aimed at developing and validating machine learning algorithms that can adapt to input data reflecting the nature of medical information, rather than at imposing an idea of data accuracy and completeness that do not fit patient records and medical registries, for which data quality is far from optimal.

The Need to Open the Machine Learning Black Box

A further issue involves the nature of machine learning algorithms, which are often referred to as "black box models," whereby the rationale for the outputs generated is inscrutable not only by physicians but also by the engineers who develop them. The preceding case about management of patients with pneumonia[6] is a relevant example. In that setting, different machine learning models for risk prediction were evaluated to choose the most accurate one. The identification of the intensive care acting as a confounder could be considered in virtue of the model that had its classification rules explicit; however, the other models did not permit such post hoc analysis.

Because purely accuracy-driven performance metrics are now pushing toward more opaque models like artificial neural networks, as in the study of referable diabetic retinopathy,[1] similar subtle shortcomings of ML-DSS may be difficult or impossible to prevent or detect. To alleviate the tension between accuracy and interpretability, research is being conducted to have ML-DSS automatically provide explanations, and to offer physicians rich interactive visualization tools to explore the implications of potential exposure variables. Despite the utility of these technology improvements, their availability will not relieve physicians from acquiring stronger skills in assessing the value of machine learning–based aids in practice.

Conclusions

It is likely that machine learning applications will soon transform some sectors of health care in ways that may be valuable but may have unintended consequences. Use of ML-DSS could create problems in contemporary medicine and lead to misuse. The quality of any ML-DSS and subsequent regulatory decisions about its adoption should not be grounded only in performance metrics, but rather should be subject to proof of clinically important improvements in relevant outcomes compared with usual care, along with the satisfaction of patients and physicians.

A prudent attitude toward research on unintended consequences could help reduce the odds of negative consequences. Moreover, if such consequences occur despite these efforts, research could help manage and reduce the related effects of these consequences.

ARTICLE INFORMATION

Published Online: July 20, 2017. doi:10.1001/jama.2017.7797

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Additional Contributions: We thank Rich Caruana, PhD (Microsoft Research and Cornell University, Ithaca, New York), for helpful comments and for sharing relevant unpublished results, and Camilla Alderighi, MD (Centro Studi Medicina Avanzata, Florence, Italy), for helpful comments and assistance with the literature survey. No compensation was received for their contribution.

REFERENCES

1. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.

2. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care. *J Am Med Inform Assoc*. 2004;11(2):104-112.

3. Hoff T. Deskilling and adaptation among primary care physicians using two work innovations. *Health Care Manage Rev*. 2011;36(4):338-348.

4. Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions. *Med Decis Making*. 2013;33(1):98-107.

5. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error. *J Am Med Inform Assoc*. 2003;10(5):478-483.

6. Caruana R, Lou Y, Gehrke J, et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer International Publishing AG; 2015:1721-1730.

7. Dharmarajan K, Strait KM, Tinetti ME, et al. Treatment for multiple acute cardiopulmonary conditions in older adults hospitalized with pneumonia, chronic obstructive pulmonary disease, or heart failure. *J Am Geriatr Soc*. 2016;64(8):1574-1582.

8. Svensson CM, Hübler R, Figge MT. Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance. *J Immunol Res*. 2015; 2015:573165.