Neurosurgical Forum

BROCA'S AREA

The embedded biases in hypothesis testing and machine learning

Ghaith Habboub, MD,¹ Matthew M. Grabowski, MD,¹
Michael L. Kelly, MD,² and Edward C. Benzel, MD¹

¹Department of Neurosurgery, Cleveland Clinic, Cleveland; and

²Department of Neurosurgery, The Metrohealth System, Cleveland, Ohio

The Development of the Therapeutic Illusion

In his *New Yorker* piece “Overkill,” Atul Gawande describes the essence of the problem related to overutilization of healthcare resources in the US by stating, “An avalanche of unnecessary medical care is harming patients physically and financially.”³ A significant portion of this overutilization phenomenon is related to physicians succumbing to what has been described as the “therapeutic illusion,” as defined in David Cassarett’s *New England Journal of Medicine (NEJM)* article titled “The Science of Choosing Wisely—Overcoming the Therapeutic Illusion.”¹ In this perspective piece, he reasons that “the outcome of virtually all medical decisions is at least partly outside the physician’s control, and random chance can encourage physicians to embrace mistaken beliefs about causality.” He goes on to address the endgame of therapeutic illusion—confirmation bias—by stating, “Once a treatment is under way, physicians (and patients) tend to look for evidence that is having some kind of positive effect.” This illusion of success is very problematic, because it leads to an exaggerated eagerness to treat, which affects future care-related decisions via the introduction of unnecessary, costly, and sometimes harmful medical care.

The Origin of Hypothesis Testing

In his book *The Logic of Scientific Discovery*, Karl Popper describes how science is based on disconfirming hypotheses, as opposed to confirming them, while also pointing out that it can be easy to derive a conclusion if the researcher is looking for it. Popper termed the latter “pseudoscience.”⁵ This foundational fact about the process of scientific investigation represents a paradigm shift from Popper’s strategy of disconfirming the null hypothesis,

to what is routinely practiced in clinical research today, i.e., jumping to confirm an alternative hypothesis, as opposed to rejecting the null hypothesis. We term the latter “conclusion-based research.” With it, the researcher creates a hypothesis and through deductive reasoning, or top-down logic, attempts to prove his/her theory. Conversely, an approach utilizing inductive reasoning, or bottom-up logic, has the advantage of exploring for the truth without relying on a preconceived notion (exploratory inquiry), a methodology we define as “process-based research.”

Reaching conclusions from an incomplete set of observations relies on the implementation of one or more reasoning strategies. Deductive reasoning is used to derive conclusions from implicit information provided in the premises; for example, cervical decompression improves functional outcomes in patients with cervical myelopathy. This logic is sound as long as all patients with cervical myelopathy have stenosis and cervical decompression adequately addresses the stenosis. Inductive reasoning, however, supports conclusions from premises even though the conclusion can go beyond what the premises state, such as degenerative myelopathy is associated with stenosis and decompression adequately addresses the stenosis, therefore decompression improves outcome in metastatic spinal disease, causing myelopathy. Abductive reasoning uses the simplest possible explanation to reach a conclusion. It is often used in day-to-day clinical decision-making. With this form of reasoning, a hypothesis is made and then tested. If accurate, the reasoning process is complete. If not accurate, a new hypothesis is generated and tested. This process is repeated until the hypothesis is found to be accurate. This “Occam’s razor” approach can be used when multiple conclusions exist or in machine learning when multiple algorithms provide similar information. Despite being the foundation of traditional logic, both inductive and deductive reasoning suffer from intrinsic flaws. Deductive reasoning relies on the preconceived notion embedded in the method, while inductive reasoning can establish unreal correlations between variables. Perhaps, with the right tools, we can employ inductive reasoning logic by using more advanced inductive reasoning methods and thus remove preconceived notions.

An example of this occurred in April 2016, when two prospective, randomized trials were published in the *NEJM*. Each attempted to discern whether the addition of lumbar fusion to standard laminectomy produces superior outcomes in patients with degenerative spondylolisthesis.^{2,4} The articles, while appearing to address the same question, had widely disparate methodologies and therefore yielded contradictory findings, with one showing superior outcomes with the addition of lumbar fusion, while the other not showing such superiority. The primary focus of the Swedish trial was all patients with 1-2-level lumbar stenosis.² However, their subgroup analysis addressed the presence or absence of spondylolisthesis, which provided an opportunity to compare these two articles. Of note, these two articles were published in the same issue of the *NEJM* for the purpose of comparison.

This example illustrates an undesirable effect: that medical/surgical studies, while they seem internally valid, are designed in a way that likely confirms the initial biases of those engaged in conclusion-based research. The scrutinizing eye will note that this limits the generalizability and obscures the actual correlations and cause-and-effect relationships. But to the ordinary reader, these results can be misinterpreted and applied to future practice and literature generation. This has been observed frequently in both the surgical and nonsurgical literature. So, is there another strategy for discovering or predicting the truth, or at a minimum, guiding our investigative inquiries in such a way that they minimize or avoid such human biases?

Machine Learning as an Alternative

Machine learning is a rapidly emerging field in artificial intelligence that utilizes data-driven algorithms. It involves the construction and application of statistical algorithms that learn and make observations from existing data, and then creates a predictive model based on that data. With advances in computer-processing capability, data storage, and networking, these computer-based algorithms can perform the intricate and extremely complex mathematical operations of classification or regression on immense amounts of data to detect intricate and potentially previously unknown patterns in that data. To accomplish this, the researcher only needs to identify the categories of the outcome or the numeric outcome in regression tasks. The data is then processed in iterative cycles composed of data input, generating predictions, and error calculation and minimization, again with the intent of creating the most accurate predictive model. This removes many harmful human biases by letting the computer “decide” what is important within the parameters set, and leaves minimal room for researcher assumptions regarding an association or cause-and-effect relationship in the generation of a model.

Process- and Conclusion-Based Research

We elaborate on the terms conclusion-based and process-based research as follows: conclusion-based research uses deductive reasoning (the tool) through misusing hypothesis testing (the method) to prove a preconceived assumption, while process-based research uses inductive

reasoning (the tool) through a variety of methods, including machine learning (the method), to explore and attempt to determine the conclusion (Fig. 1A). Rules are embedded in the input of conclusion-based research, while such rules are generated in the output of process-based research. Recognizing that both methods can drift from truth, we believe relying more on data and less on rules might decrease our human intrinsic biases.

Limitation of Machine Learning

While less human bias is observed in machine learning, new types of biases emerge. Machine-learning algorithms are able to form unrealistic connections between variables due to the model's high capacity. Such connections are translated into the algorithm memorizing the training data, a process described as overfitting. Relying on limited metrics (e.g., accuracy) and not appropriately validating the data heavily contributes to overfitting. In addition, such algorithms are data hungry, and providing low volumes of data can easily facilitate model overfitting. Other limitations also exist, including machine-learning fairness, model generalizability, and model drifting.

The Human Monitoring of Machine-Learning Biases

Human supervision can provide another level of reassurance when utilizing machine learning. The simplified machine-learning model cycle is as follows: data mining, model training, and model deployment. At each of these steps, human checkpoints can be added to minimize machine biases. Machine-learning interpretability can help us convert the described black box algorithm (i.e., machine learning) into a transparent model. Interpretability methodologies can help us understand why the algorithm is making such a decision and thus provide an insertion point for adding human checks into the decision process (Fig. 1B).

Conclusions

Limitations exist in both hypothesis testing and machine learning. Substituting human biases in hypothesis testing with machine biases in machine learning is evident in the recent literature. Humans, despite good intentions, are biased regarding their decision-making process. However, relying solely on machine logic, while minimizing human bias, can shift us into a machine-biased world. Understanding the limitations of machine learning and adding checkpoints can provide a safety measure against these biases.

We did not intend for this article to function as a review of machine learning. Rather, we intended to provide a brief background regarding the philosophy behind standard statistical analysis and machine learning. In addition, we have emphasized the limitations of these methods and have provided a foundation for readers to critique them.

Considering existing concerns regarding the limitations of conclusion-based research in our modern scientific climate, judicial use of machine learning offers a powerful tool that can support our search for the truth and aid in our medical decision-making, while limiting or eliminating the undue influence of conscious and unconscious

FIG. 1. A: Portrayal of the iterative means by which process-based research uses existing evidence of truth (answers) and data to reach the truth. While less affected by human choices, both conclusion-based and process-based research can deflect from the truth. **B:** Description of a simplified lifecycle of a machine-learning model and checkpoints to assess model performance per each step.

biases. Perhaps then we begin to understand the harmful and costly overtreatment of our patients that plagues our medical environment today.

References

1. Casarett D: The science of choosing wisely—overcoming the therapeutic illusion. *N Engl J Med* 374:1203–1205, 2016
2. Försth P, Ólafsson G, Carlsson T, Frost A, Borgström F, Fritzell P, et al: A randomized, controlled trial of fusion surgery for lumbar spinal stenosis. *N Engl J Med* 374:1413–1423, 2016
3. Gawande A: Overkill: an avalanche of unnecessary medical care is harming patients physically and financially. What can we do about it? *New Yorker*. May 4, 2015 (<https://www.newyorker.com/magazine/2015/05/11/overkill-atul-gawande>) [Accessed March 9, 2020]
4. Ghogawala Z, Dziura J, Butler WE, Dai F, Terrin N, Magge SN, et al: Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis. *N Engl J Med* 374:1424–1434, 2016
5. Popper KR, Weiss G: The logic of scientific discovery. *Phys Today* 12:53–54, 1959

Disclosures

The authors report no conflict of interest.

Correspondence

Ghaith Habboub: habboug@ccf.org.

INCLUDE WHEN CITING

DOI: 10.3171/2020.2.FOCUS191016.

©AANS 2020, except where prohibited by US copyright law