frontiers in Artificial Intelligence

OPINION
published: 21 January 2020
doi: 10.3389/frai.2019.00032

Mysteries, Epistemological Modesty, and Artificial Intelligence in Surgery

Tyler J. Loftus[1], *Gilbert R. Upchurch Jr.*[1], *Daniel Delitto*[1], *Parisa Rashidi*[2] and *Azra Bihorac*[3]*

[1] Department of Surgery, University of Florida Health, Gainesville, FL, United States, [2] Departments of Biomedical Engineering, Computer and Information Science and Engineering, and Electrical and Computer Engineering, University of Florida, Gainesville, FL, United States, [3] Department of Medicine, University of Florida Health, Gainesville, FL, United States

Keywords: surgery, machine learning, artificial intelligence, informed consent, risk prediction, phenotyping, prognostics, decision-making

Life is filled with puzzles and mysteries, and we often fail to recognize the difference. As described by Gregory Treverton and Malcolm Gladwell, puzzles are solved by gathering and assimilating all relevant data in a logical, linear fashion, as in deciding which antibiotic to prescribe for an infection. In contrast, mysteries remain unsolved until all relevant data are analyzed and interpreted in a way that appreciates their depth and complexity, as in determining how to best modulate the host immune response to infection. When investigating mysteries, we often fail to appreciate their depth and complexity. Instead, we gather and assimilate more data, treating the mystery like a puzzle. This strategy is often unsuccessful. Traditional approaches to predictive analytics and phenotyping in surgery use this strategy.

OPEN ACCESS

Edited by:
Nicole C. Kleinstreuer, National Institute of Environmental Health Sciences (NIEHS), United States

Reviewed by:
Shi-Cong Tao, Shanghai Sixth People's Hospital, China

***Correspondence:**
Azra Bihorac
abihorac@ufl.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *19 August 2019*
Accepted: *24 December 2019*
Published: *21 January 2020*

Citation:
Loftus TJ, Upchurch GR Jr, Delitto D, Rashidi P and Bihorac A (2020) Mysteries, Epistemological Modesty, and Artificial Intelligence in Surgery. Front. Artif. Intell. 2:32. doi: 10.3389/frai.2019.00032

WEAKNESSES INHERENT TO TRADITIONAL PREDICTIVE ANALYTICS AND PHENOTYPING

Postoperatively, most patients recover along a clinical trajectory that can be predicted by their physiologic reserve, the severity of the underlying disease process, and the physiologic insult associated with the planned operation. These predictions augment the decision to offer an operation and inform discussions with patients and their caregivers regarding treatment options and prognosis. This process often relies on biased, error-prone individual judgement, especially when decisions are made under time constraints and uncertainty, leading to preventable harm. Decision-support tools are intended to augment this process. Unfortunately, traditional decision-support tools regard postoperative trajectories as puzzles which may be solved by gathering and assimilating relevant data in a logical, linear fashion with parametric regression modeling. Some regression models predict dichotomous outcomes with accuracy similar to a coin toss. For example, in applying six different regression-based prediction models to 1,380 patients undergoing colorectal surgery, Bagnall et al. (2018) found that all six models performed poorly with area under the receiver operating characteristic curve (AUROC) 0.46–0.61 (Bagnall et al., 2018). In these cases, poor model accuracy is often attributed to stochastic, or random, risk.

STOCHASTIC RISK AND EPISTEMOLOGICAL MODESTY

For surgeons who are sometimes wrong but never in doubt, stochastic risk is an uncharacteristic foray into epistemological modesty, or recognition that our knowledge and understanding are limited. However, if what we call stochastic risk is instead risk that we have failed to predict because we are treating mysteries like puzzles and using the wrong prediction tools, then we are exercising ignorance and complacency, not epistemological modesty. Parametric regression models make predictions with logical, linear rules expressed as algorithms; machine and deep learning artificial

intelligence models accurately represent the complex, non-linear associations among inputs and outputs by learning from examples. Because pathophysiology does not consistently conform to additive, linear rules, one might expect that artificial intelligence models would be advantageous.

ADVANTAGES FOR ARTIFICIAL INTELLIGENCE IN PREDICTIVE ANALYTICS AND PHENOTYPING

For some tasks, like predicting mortality among heart failure patients, logistic regression can perform as well or better than certain machine learning methods like regression tree analysis (Austin et al., 2010). For complex tasks like predicting several postoperative complications, artificial intelligence models outperform regression-based techniques and clinician judgement (Bertsimas et al., 2018; Bihorac et al., 2018; Brennan et al., 2019). Bertsimas et al. (2018) developed an Optimal Classification Trees machine learning model to predict mortality and 18 complications following emergency surgery, demonstrating superior accuracy compared with the ACS NSQIP calculator (AUROC 0.92 vs. 0.90). The online and phone application asks users 4–11 questions that are generated in response to prior answers. Manual data entry requires more time and input from providers than an automated model, but obviates requirements for data security and encryption of protected health information from electronic health records (EHR). Bihorac et al. (2018) developed and validated the *MySurgeryRisk* platform with automated EHR data linked to US Census data regarding neighborhood characteristics, using 285 variables to predict eight postoperative complications with AUROC 0.82–0.94 (Bihorac et al., 2018). EHR data feeds the algorithm automatically, obviating manual data search and entry, and overcoming a major obstacle to clinical adoption. In a prospective usability study, algorithm accuracy was significantly greater than physician accuracy in predicting postoperative complications (Brennan et al., 2019). These observations have profound implications for the complex, high-stakes decisions surgeons make when offering an operation and addressing modifiable risk factors, tasks that are currently supported by the National Surgical Quality Improvement Program (NSQIP) Surgical Risk Calculator. If machine learning methods consistently outperform the NSQIP calculator and individual surgeon judgement, then surgeons will face a professional and moral imperative to integrate machine learning in the shared decision-making process of informed consent.

Risk assessments and predictive analytics depend on phenotyping to accurately identify and classify patients, diseases, and complications. Phenotyping is also critically important for identifying candidates for emerging treatments and clinical trial enrollment and standardizing definitions for clinical and research applications. Similar to traditional predictive analytics, traditional phenotyping uses rules expressed as algorithms, gathering data into additive and parametric models, treating classification tasks as puzzles. Results from this approach are highly variable, particularly for complex conditions like frailty. Flaatten and Clegg (2018) demonstrated that frailty phenotyping is highly variable among critically ill patients from different institutions—even when applying a single, validated instrument to each individual cohort—with the incidence of frailty ranging from 13 to 53%, without discernable trends relating frailty to chronological age. Surgeons and their patients need accurate frailty phenotyping to inform the decision to operate, identify patients who may benefit from prehabilitation prior to major elective surgery, and predict the likelihood of postoperative complications and recovery (Barberan-Garcia et al., 2018). Emerging evidence suggests that even relatively common and highly morbid conditions with established international consensus definitions like the acute respiratory distress syndrome (ARDS) have subtypes that impact management strategies and outcomes, but are often unrecognized. Sinha and Calfee (2019) found that combining clinical and biological data can identify hyper- and hypo-inflammatory ARDS phenotypes that have different responses to mechanical ventilation strategies, intravenous fluid management, and medications. Notably, suboptimal identification and classification of ARDS may portend failure to rescue postoperative patients (Ghaferi et al., 2009).

As an alternative to traditional phenotyping methods, deep learning models can autonomously and accurately phenotype according to established definitions. In addition to performing predictive analytics, facilitating clinical trial enrollment, and standardizing definitions for clinical and research applications, deep learning can solve phenotyping mysteries. Unsupervised models learn relationships and concepts from data and identify patterns and clusters, promoting the discovery of new clinically relevant phenotypes. Artificial intelligence has the potential to revolutionize oncologic phenotyping and prognostication. Among patients with pancreatic cancer, circulating tumor cell histopathology independently predicts the timing of disease recurrence as well as overall survival when adjusting for margin status and tumor grade (Poruk et al., 2016). The clinical utility of this observation is subject to the time-consuming and resource-intense nature of performing and interpreting immunohistochemistry. Alternatively, computer vision programs are adept at performing similar tasks, like recognizing skin cancer with greater accuracy than board-certified Dermatologists, producing results in moments (Esteva et al., 2017). This approach could be used to inform decisions regarding systemic therapies for cancer patients. Kather et al. (2019) demonstrated that deep learning can accurately detect tumor (AUC < 0.99) and predict microsatellite instability on hematoxylin and eosin-stained slides (AUC 0.77–0.84), identifying patients who are likely to benefit from immunotherapy. By learning from examples and representing complex, non-linear pathophysiology, machine learning has the potential to augment clinical reasoning in surgery by performing predictive analytics, and disease phenotyping, and decision analysis tasks that are beyond the reach of traditional methods (**Figure 1**).

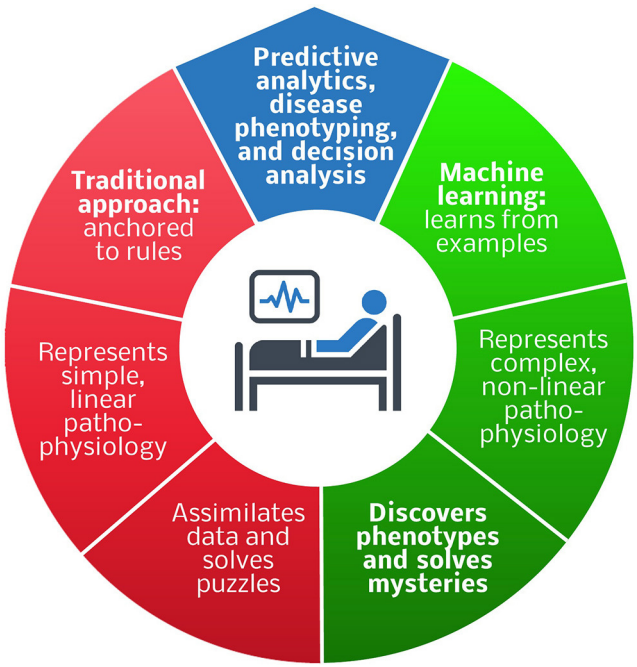

FIGURE 1 | Comparison of traditional and machine learning approaches to predictive analytics, disease phenotyping, and decision analysis for augmented clinical reasoning. Blue region: predictive analytics and disease phenotyping are important tasks in patient-centered decision-making. Red region: traditional rule-based approaches are adequate for representing simple, linear pathophysiology, and solving puzzles, e.g., deciding which antibiotic to prescribe for an infection. Green region: machine learning techniques that learn from examples are preferable for representing complex, non-linear pathophysiology, and solving mysteries, e.g., determining how to best modulate the host immune response to infection.

CHALLENGES AND SOLUTIONS FOR ARTIFICIAL INTELLIGENCE APPLICATIONS IN SURGERY

Despite these advantages, machine learning models have several limitations that must be addressed prior to widespread clinical adoption. Clinicians may be unfamiliar with methods for interpreting machine learning outputs. Conventional methods like Random Forest variants are relatively transparent and easy to interpret, and emerging techniques improve the interpretability of deep learning models, but it remains difficult to ascertain the relative importance of individual model inputs in determining outputs. To improve output interpretability, model self-attention mechanisms can reveal periods during which inputs make significant contributions to outputs, and models can be trained on labeled patient data and then a linear gradient boosting tree so that the model will assign relative importance to patient data input features (Che et al., 2016; Shickel et al., 2019). However, many clinicians also have difficulty interpreting regression outputs like odds ratios, relative risk values, and even simple *p*-values, suggesting that improving statistical fluency is a global objective that is not unique to artificial intelligence modeling (Anderson et al., 2013; Krouss et al., 2016).

Machine and deep learning models perform well, but like regression models, they are fallible. When they fail, they could impact a large number of patients in a short period of time. Therefore, careful monitoring of model outputs and interpretation by astute clinicians is critically important. Artificial intelligence models are capable of providing a proxy measure of how confident they are that their output is accurate, which can alert clinicians to situations in which outputs should not be trusted. This confidence level can be approximated using an activation function on the final layer of a machine learning model with a softmax function that maps network activations to (0,1), with lower values suggesting lower confidence that predicted probabilities match true probabilities, and higher values suggesting higher confidence. Notably, a model may be uncertain of its predictions even when the softmax output is high (Gal, 2016). Alternatively, the predicted probabilities of machine learning models may be calibrated with reliability curves, producing confidence scores rather than distributions of possible outputs (Guo et al., 2017).

In addition, ethical challenges may arise when models fail and liability is distributed among computer programs, their developers, and the clinicians using the programs. Surgeons, data scientists, informatics experts, and ethicists must work together to address these challenges by improving model transparency, optimizing model accuracy, and establishing a framework to assign liability for errors. Initial prospective implementation of artifical intelligence models in surgery should occur on a small scale under close monitoring, consistent with guidelines regarding the Software as Medical Device (SaMD) category created by the US Food and Drug Administration and the International Medical Device Regulators Forum. As technologies continue to improve over time and involved parties commit to thoughtful and sober implementation of these technologies, the safety and efficacy of artificial intelligence healthcare applications will continue on an upward trajectory.

Finally, it seems unlikely that capitalizing on these advantages will be as simple as switching from basic regression-based to machine learning models. Clinical integration of machine learning will require not only extensive medical domain knowledge founded in basic and translational research, but also informatics expertise, multidisciplinary collaboration, and skillful application of implementation science.

CONCLUSIONS

True epistemological modesty recognizes that continued reliance on individual judgement and traditional predictive analytics and phenotyping may lead to preventable harm. It is irresponsible to attribute these failings to mysterious pathophysiology and stochastic processes without deploying new technologies that capture the depth and complexity of underlying pathophysiology and improve phenotyping and predictive accuracy. Thoughtful clinical integration of artificial intelligence

has the potential to transform surgical care by augmenting the decision to operate, informing discussions with patients and their caregivers regarding treatment options and prognosis, predicting treatment response to emerging and experimental treatments, and addressing other unsolved mysteries in surgery.

AUTHOR CONTRIBUTIONS

TL and DD contributed to conceptual design, performed the literature review, and drafted the manuscript. GU, PR, and AB contributed to conceptual design, interpreted the literature, and made critical revisions.

FUNDING

AB was supported by R01 GM110240 from the National Institute of General Medical Sciences (NIGMS) and by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the NIGMS. PR was supported by CAREER award, NSF-IIS 1750192, from the National Science Foundation (NSF), Division of Information and Intelligent Systems (IIS) and by NIH NIBIB R21EB027344-01. TL was supported by a post-graduate training grant (T32 GM-008721) in burns, trauma, and perioperative injury from the NIGMS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Anderson, B. L., Williams, S., and Schulkin, J. (2013). Statistical literacy of obstetrics-gynecology residents. *J. Grad. Med. Educ.* 5, 272–275. doi: 10.4300/JGME-D-12-00161.1

Austin, P. C., Tu, J. V., and Lee, D. S. (2010). Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J. Clin. Epidemiol.* 63, 1145–1155. doi: 10.1016/j.jclinepi.2009.12.004

Bagnall, N. M., Pring, E. T., Malietzis, G., Athanasiou, T., Faiz, O. D., Kennedy, R. H., et al. (2018). Perioperative risk prediction in the era of enhanced recovery: a comparison of POSSUM, ACPGBI, and E-PASS scoring systems in major surgical procedures of the colorectal surgeon. *Int. J. Colorectal Dis.* 33, 1627–1634. doi: 10.1007/s00384-018-3141-4

Barberan-Garcia, A., Ubre, M., Roca, J., Lacy, A. M., Burgos, F., Risco, R., et al. (2018). Personalised prehabilitation in high-risk patients undergoing elective major abdominal surgery: a randomized blinded controlled trial. *Ann. Surg.* 267, 50–56. doi: 10.1097/SLA.0000000000002293

Bertsimas, D., Dunn, J., Velmahos, G. C., and Kaafarani, H. M. A. (2018). Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery Risk (POTTER) Calculator. *Ann. Surg.* 268, 574–583. doi: 10.1097/SLA.0000000000002956

Bihorac, A., Ozrazgat-Baslanti, T., Ebadi, A., Motaei, A., Madkour, M., Pardalos, P. M., et al. (2018). MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann. Surg.* 269, 652–662. doi: 10.1097/SLA.0000000000002706

Brennan, M., Puri, S., Ozrazgat-Baslanti, T., Feng, Z., Ruppert, M., Hashemighouchani, H., et al. (2019). Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: a pilot usability study. *Surgery* 165, 1035–1045. doi: 10.1016/j.surg.2019.01.002

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. (2016). Interpretable deep models for ICU outcome prediction. *AMIA Annu. Symp. Proc.* 2016, 371–380.

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Flaatten, H., and Clegg, A. (2018). Frailty: we need valid and reliable tools in critical care. *Intensive Care Med.* 44, 1973–1975. doi: 10.1007/s00134-018-5404-5

Gal, Y. (2016). *Uncertainty in Deep Learning*. Department of Engineering, University of Cambridge. Available online at: http://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf (accessed July 13, 2019).

Ghaferi, A. A., Birkmeyer, J. D., and Dimick, J. B. (2009). Complications, failure to rescue, and mortality with major inpatient surgery in medicare patients. *Ann. Surg.* 250, 1029–1034. doi: 10.1097/SLA.0b013e3181bef697

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. (2017). *On Calibration of Modern Neural Networks*. arXiv: 1706.04599v2 [cs.LG]. Available online at: https://arxiv.org/pdf/1706.04599.pdf (accessed July 13, 2019).

Kather, J. N., Pearson, A. T., Halama, N., Jager, D., Krause, J., Loosen, S. H., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* 25, 1054–1056. doi: 10.1038/s41591-019-0462-y

Krouss, M., Croft, L., and Morgan, D. J. (2016). Physician understanding and ability to communicate harms and benefits of common medical treatments. *JAMA Intern. Med.* 176, 1565–1567. doi: 10.1001/jamainternmed.2016.5027

Poruk, K. E., Valero, V. 3rd, Saunders, T., Blackford, A. L., Griffin, J. F., Poling, J., et al. (2016). Circulating tumor cell phenotype predicts recurrence and survival in pancreatic adenocarcinoma. *Ann. Surg.* 264, 1073–1081. doi: 10.1097/SLA.0000000000001600

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., and Rashidi, P. (2019). DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci. Rep.* 9:1879. doi: 10.1038/s41598-019-38491-0

Sinha, P., and Calfee, C. S. (2019). Phenotypes in acute respiratory distress syndrome: moving towards precision medicine. *Curr. Opin. Crit. Care* 25, 12–20. doi: 10.1097/MCC.0000000000000571

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.