Striking the Right Balance—Applying Machine Learning to Pediatric Critical Care Data*

Jenna Wiens, PhD
Division of Computer Science and Engineering
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI

James Fackler, MD
Department of Anesthesiology and Critical Care Medicine
Johns Hopkins University School of Medicine
Baltimore, MD

For George Seurat, it was the balances of dense and hollow, flat and deep, and dark and light as he created A Sunday on La Grande Jatte (1). Balance in critical care will increasingly become that with human and machine. A discussion for later will be how clinicians achieve balance as we work with even conversational robots (2). For this discussion, the balance to be achieved is between human and machine learning (ML).

At the core of most machine learning approaches is some notion of similarity (or distance) between examples. Given such a definition, we can identify examples that are in close proximity and make predictions about one given the others. In healthcare, the ability to automatically pinpoint patients along similar trajectories could help guide clinical decision-making, assuming that similar patients have similar outcomes. In the case of inherited disease, we know that genetic similarity can help predict outcomes (3). But, what can we deduce from observational data collected in the PICU?

This is exactly what Williams et al (4), in their article published in this issue of *Pediatric Critical Care Medicine*, set out to test using electronic health record (EHR) data. The authors applied k-means clustering to vitals and laboratory results extracted from the first 6 hours of 11,384 PICU episodes. Based on the Euclidean distance between examples in an 80D feature space, the authors discovered 10 stable clusters. To investigate the clinical relevance of the resulting clusters, the authors measured the empirical distribution of various outcomes (e.g., mortality).

Fortunately, data collected in the first 6 hours of an ICU admission are predictive of therapy given. Any other conclusion would have implied that we are either 1) treating patients randomly or 2) measuring the wrong things. Now that we have established that EHR data contain measurements that precede (and even predict) clinically meaningful events, we can shift our focus to more clinically relevant tasks. Today's critical care physicians already suffer from information overload (5). So, like Williams et al (4), we should be asking ourselves: how can we best transform these data into "knowledge" that a physician can use and then distill it in such a way that it becomes "useful"?

Figure 3 in (4) illustrates the relationship between cluster assignment and risk of mortality. The authors achieve an area under the curve of 0.77. Although better than random, this is still well below the discriminative performance of other classifiers for the same task (6). This is due, in part, to the fact that Williams et al (4) use an unsupervised learning approach. Such an approach makes no assumptions about the outcome of interest, weighting each feature or input equally when comparing patients. In contrast, supervised learning techniques zero in on the most relevant inputs (assumed by the investigators) with respect to a particular outcome. While described as the proverbial icing on the cake (6), supervised learning can help us make more meaningful patient comparisons.

Supervised or unsupervised, during evaluation of an approach, one must often choose to focus on a specific outcome. Like many others, Williams et al (4) focus on in-hospital mortality and length of stay. These outcomes are by far the most common, in part because they are simple to obtain from the EHR. However, models trained specifically to predict risk of more proximately modifiable conditions (e.g., acute respiratory distress syndrome or sepsis) are arguably more actionable than those trained to predict all-cause mortality.

In addition to thinking carefully about the outcome of interest or task for the ML algorithm, data cleaning and feature extraction are critical. In the series of necessary steps laid out by Williams et al (4), these appear as "step 2" and "step 3." After removing erroneous entries and lumping together vitals obtained through different means, the authors generate an 80D feature vector for each episode. Although "data-driven," such data generation is hardly automated; many human or "expert-driven" decisions go into such processes.

*See also p. 599.

Key Words: big data; human-computer interactions; machine learning

Drs. Wiens and Fackler received funding from Machine Learning for Healthcare (Board Member of this nonprofit, reimbursed for travel expenses to annual meeting). Dr. Fackler also disclosed that he is a Founder of Rubicon Health LLC under a licensing agreement between Rubicon Health and the Johns Hopkins University, and he is entitled to royalty distributions on technology described in this presentation. He is also the founder of and holds equity in Rubicon Health; however, Rubicon Health has neither a product nor revenue at this time.

DOI: 10.1097/PCC.0000000000001578

For example, the authors chose to represent the first 6 hours of each vital as a series of six hourly measurements, as opposed to 12 or 3. Such a decision could either double or halve the relative contribution of the vital measurements to the overall notion of patient similarity. Just like the number of clusters, k, how one measures the distance between points is a modeling decision driven by the choice of features. Furthermore, while multiple measurements are considered over the course of the 6-hour period, by using a Euclidean distance any trends are essentially ignored. As the authors state, if decline in respiratory rate is important, but occurs at different points in time, the patients will look quite different. Similarly, the choice of imputation method for filling in missing values can affect results. Here, missing values were imputed with a carry forward method. However, there is often information in the fact that a measurement was not performed. The authors hint at a completely data-driven approach in which representations are learned. Such techniques, however, rely on massive quantities of data and can hinder interpretability and in turn our ability to glean actionable insights.

Machine learning techniques need data, but the EHR has its limitations. EHR data are intermittent, biased, and censored. Furthermore, there is a lot of data that do not make its way into the EHR but should (e.g., clinical reasoning or nuances of patient symptoms). Such data are at times available in the form of clinician entered text. However, data extraction from notes remains challenging. Data from other sources (e.g., patient-entered or wearables) could help in this context. ICUs with more ambient sensing could potentially deliver more and likely more useful data, enabling more accurate definitions of patient similarity. Finally, these data are likely to be most useful if we have the ability to easily follow patients longitudinally as they interact with different systems. To this end, efforts focusing on not only the collection of but also the sharing of data are crucial.

It is clear that, based on the efforts of Williams et al (4), leveraging EHR data requires both data-driven and expert-driven choices in our ML analyses. As we move on to predict more clinically relevant tasks, learning to strike the right balance between data and experts will be critical in moving the field forward. The objective is to get to a symbiosis as foreseen by Licklider (7) 60 years ago, "The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today." In 2018, machines still have much to learn from humans. There is a vast amount of expert knowledge that could and should be incorporated into machine learning algorithms.

Five specific points as the search for balance is sought:

1) Move beyond clustering analyses to settings in which we aim to directly predict potentially modifiable outcomes.
2) Focus on targets that have the greatest potential to improve clinical outcomes.
3) Augment clinical data with data from in-hospital ambient sensing technology (8), personal wearable and environmental sensors (9), and "patient-entered" data (10).
4) Make all the data available (11) and access data across systems using a federated approach (12) (having recently shown promise (13) in the context of machine learning).
5) Strive for the Licklider (7) balance—human computer symbiosis.

REFERENCES

1. Smith R: Seurat, Drawing His Way to the Grande Jatte. 2007. Available at: https://www.nytimes.com/2007/10/26/arts/design/26seur.html. Accessed March 28, 2018
2. Lightman A: The robot advantage. *In*: Augmented: Life in The Smart Lane. King B (Ed). Singapore, Marshall Cavendish International (Asia) Pte Ltd, 2016
3. van 't Veer LJ, Dai H, van de Vijver MJ, et al: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; 415:530–536
4. Williams JB, Ghosh D, Wetzel RC: Applying Machine Learning to Pediatric Critical Care Data. *Pediatr Crit Care Med* 2018; 19:599–608
5. Pickering BW, Herasevich V, Ahmed A, et al: Novel representation of clinical information in the ICU: Developing user interfaces which reduce information overload. *Appl Clin Inform* 2010; 1:116–131
6. Luo Y, Xin Y, Joshi R, et al: Predicting ICU Mortality Risk By Grouping Temporal Trends From a Multivariate Panel of Physiologic Measurements. Phoenix, AZ, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp 42–50
7. Licklider JCR: Man-Computer Symbiosis. Human Factors in Electronics, IRE Transactions on Human Factors in Electronics. New York, IEEE, 1960, pp 4–11
8. Haque A, Guo M, Alahi A, et al: Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. proceedings of machine learning for healthcare. *J Machine Learning Res* 2017; 68:75–87
9. Shcherbina A, Mattsson C, Waggott D, et al: Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* 2017; 7:3–12
10. Wicks P, Vaughan TE, Massagli MP, et al: Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat Biotechnol* 2011; 29:411–414
11. Despite Privacy Concerns, Israel to Put Nation's Medical Database Online. 2018. Available at: https://www.timesofisrael.com/despite-privacy-concerns-israel-to-put-nations-medical-database-online/. Accessed April 2, 2018
12. Mandl KD, Kohane IS: Federalist principles for healthcare data networks. *Nat Biotechnol* 2015; 33:360–363
13. Chang K, Balachandar N, Lam C, et al: Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc* 2018; 29:1–10