Bringing machine learning to the bedside:
focusing clinical decision support with predictive modeling

Department of Internal Medicine Grand Rounds

Mujeeb A. Basit, MD, MMSc

Assistant Professor of Internal Medicine

Division of Cardiology

University of Texas Southwestern Medical Center

January 18, 2019

This is to acknowledge that Mujeeb A. Basit, MD, MMSc has disclosed that he does not have any financial interests or other relationships with commercial concerns related directly or indirectly to this program. Dr. Basit will not be discussing off-label uses in his presentation.

Cardiologist Mujeeb Basit, MD, MMSc is an expert in constructing and monitoring complex clinical decision support systems. Dr. Basit earned his undergraduate degree in computer science and worked at the Human Genome Project and the Dallas Heart Study prior to pursuing a medical degree. After completing cardiology fellowship, he earned a clinical informatics fellowship where he worked closely with national experts on clinical decision support and the use of machine learning to identify anomalous behaviors prior to them impacting clinical care. Dr. Basit joined the faculty at the University of Texas Southwestern Medical Center at Dallas in 2016, where he now serves as Assistant Professor of Internal Medicine and is Associate Chief Medical Informatics Officer. Dr. Basit has an interest in clinical process and outcomes improvement with the use of advanced minimally invasive decisions support. In addition to helping patients, Dr. Basit enjoys his role mentoring young physicians-in-training and teaching them the value of understanding informatics in day to day clinical use.

Overview

1. Machine learning algorithms segment or sort complex groups. They are in most cases divided into three major groups: supervised, semi-supervised, and unsupervised models.

2. These systems raise ethical questions about their source data, biases, and implementation.

3. Appropriate use of algorithms in clinical decision support can create enormous value and help guide resources to where they can be used best.

4. Medical professionals must lead the effort to incorporate these models into clinical care in a safe and moral manner. The synergy between clinicians and machine learning will create enormous value to patients and the health system.

Educational Objectives

1. Identify various machine learning algorithms and how they are organized

2. Understand how to assess quality of machine learning algorithms and the models created by them

3. Describe approaches to implement these methods in real-time clinical decision support

Introduction

Machine learning (ML) affects many parts of daily life. From search results, shopping, news feeds and movie recommendations, it has changed daily activities. Regression, either logistic or linear, is a form of ML. Other ML algorithms are slightly more complex functions designed to perform similar operations to their regression cousins. As regression is divided into categorical and continuous outcome variables, ML can be further divided into supervised and unsupervised models. Supervised models are used when the outcome variable is known and unsupervised when it is not. Another dimension to divide these models is by type of algorithm such as: regression, instance-based, regularization, decision tree, Bayesian, clustering, association rule learning, artificial neural network, deep learning, dimensional reduction, ensemble, natural language process, and computer vision. There are other specialty algorithms that do not fall into any of these categories.[1]

Fundamentally these algorithms are mathematical functions that, although complex, return information that adds value to decision making. Examples commonly used include sepsis risk prediction, thirty-day readmission risk, acute in hospital decompensation, and no-show probability.[2] The creation of these models follows traditional steps. Data is collected, cleaned and analyzed, creating a model file which can be used to generate additional predictions for future data. Modern electronic health record systems (EHRs) now have the capacity to directly import these model files and generate a new patient-specific prediction score in near real-time. This has significantly reduced the amount of development effort and cost needed to bring a model to production.

At UT Southwestern, medical informaticists working with Health System Information Resources (HSIR) have focused on creating and implementing stable understandable models in the domains of clinical risk identification and operational improvement. The five-step process to bring a model to production after approval by clinical decision support (CDS) governance is:

1. Model creation and statistical validation
2. EHR build in silent mode for prospective prediction evaluation
3. Evaluate where in the clinical process the CDS intervention should be
4. CDS build, implementation, and testing
5. Continuous performance monitoring

Early experience with this process has shown success in creating ML based CDS tools. These tools still require hard work in building good CDS and additional complexity around asynchronous alerting. The interaction between clinicians and more intelligent EHRs shows promise in creating better healthcare systems of care.

The Learning Problem

Advances in computational power, memory and fast storage along with increased amounts of complex structured and semi-structured data have created a perfect formula enabling computers to perform complex learning tasks thought extremely difficult even a few years ago. Starting with the defeat of Gary Kasparov by Deep Blue in 1997,[3] the superiority of human thinking was challenged. The initial promise of artificial intelligence to solve many of the world's problems and replace humans was quickly dashed as early algorithms failed to generalize. Many of us are still traumatized by Microsoft's early efforts in AI with Clippid (also known as Clippy), the paper clip shaped office assistant. ML algorithms and the research applications of them in healthcare have been available for decades but they have not reached the bedside. Many barriers have been identified for this slow adoption with a lack of ML education during medical training[4] being a key factor.

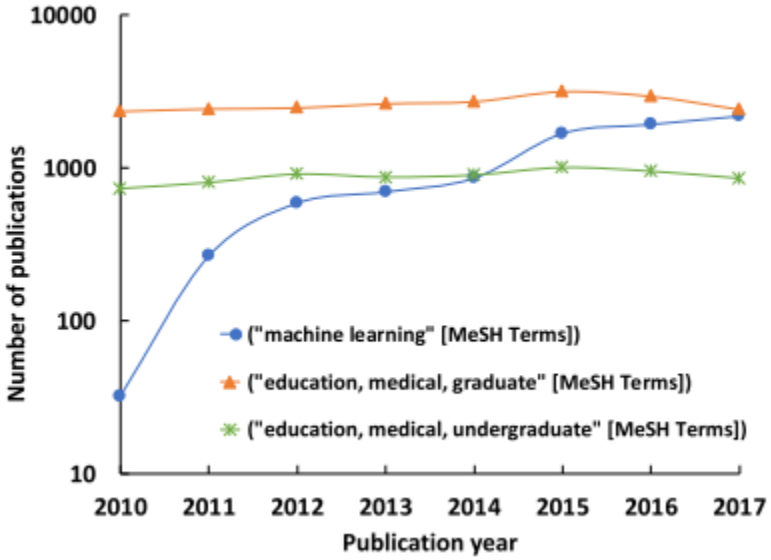


Figure 1. Published papers using "machine learning" as MeSH terms.[4]

Other barriers include the technical complexity in building real-time interfaces to operational databases for externally run ML systems, poor curation of medical data within the EHR, difficult to understand model outcomes, and ethical challenges of broadly using highly personal data to develop these models[5]. Despite these barriers, commercial funding in healthcare AI has continued to increase leading to a growth of startups focused on solving medicine's many challenges.[6,7]

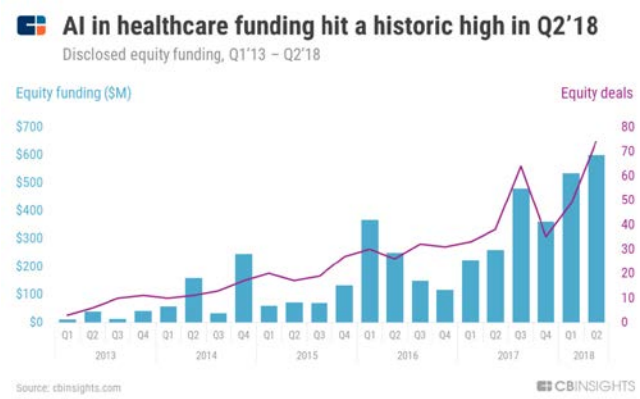


Figure 2. AI equity funding continues to grow.[6]

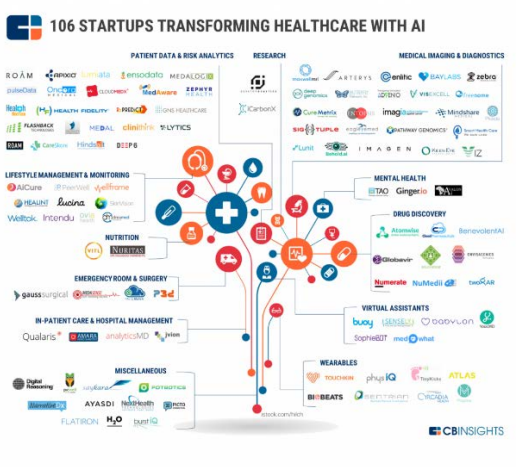


Figure 3. The number of startup "transforming healthcare with AI" continues to grow.[7]

The "solutions" are coming and it becomes incumbent on us as clinicians to determine what tools are best for our patients and our systems of care. The three strategies for accomplishing this are education, good CDS design, and systems to monitor ML performance prospectively.

ML Algorithm Taxonomy

ML algorithms can be categorized using various methods. One very common one is to divide models into whether the outcome or predicted variable is known, partially known, or unknown. The respective types are supervised learning where the outcome is known, semi-supervised where the outcome is partially known, and unsupervised learning where the outcome is unknown but there is strong suspicion of clusters. Another separate category is reinforcement learning. In this framework, the model or agent is repetitively trained on a given environment (i.e. chess, go, driving simulator, etc.) and is rewarded/reinforced for good performance. The agent iteratively improves until a desired state of performance is reached.[8,9]

Table 1. Supervised, semi-supervised and unsupervised learning simplified examples.[10]

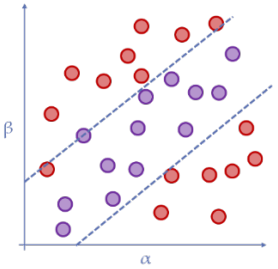
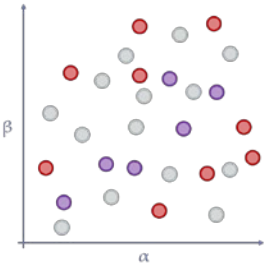
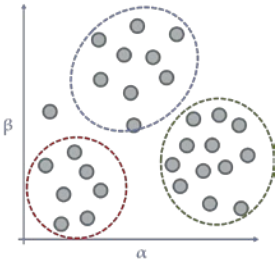
Supervised Learning	Semi-Supervised Learning	Unsupervised Learning	Reinforcement Learning
			

Table 2. Example algorithms for each ML taxonomy type.

	Supervised	Semi-Supervised	Unsupervised	Reinforcement
Discrete	Logistic Regression K-nearest neighbor Naïve-Bayes Support vector machines	Heuristics approaches	Association analysis Hidden Markov Model	Game AI Robot Navigation Real-time Decisions
Continuous	Linear regression Decision trees Random forest	Transductive or inductive learning	Clustering and Dimensional Reduction SVD, PCA, K-means	

Within these types of models there are many algorithms with specific strengths and weaknesses. Discussing most of these models in any detail is beyond the scope of this talk. There are several sources for learning about these models including MIT ML course, Weka, R community and Python community (links in research links section below). Two very common models necessary to know because of their continued success and popularity are random forest and deep learning.

Random forest is an ensemble learning method for classification or regression. It randomly divides a training set both in the number of observations as well as selection of samples or rows to multiple weak classification models. This is known as bootstrap sampling. These classifiers are individually modeled followed by a bootstrap aggregating phase. During aggregation the many weak voters come together to create a stronger single result either by simple majority vote or other more complex methods.[2,11] Deep learning is a type of neural network with multiple complex layers. It is similar to the function of the human brain. It uses large labeled data sets to train these networks with many rounds of training until appropriate recognition levels are reached.[12] This process requires extensive computational time and memory. Until recently deep learning methods were theoretical or in the domain of super computers. The availability of very powerful graphical processing units at reasonable prices has made these models more accessible. Deep learning neural networks are commonly used for image recognition.[13,14]

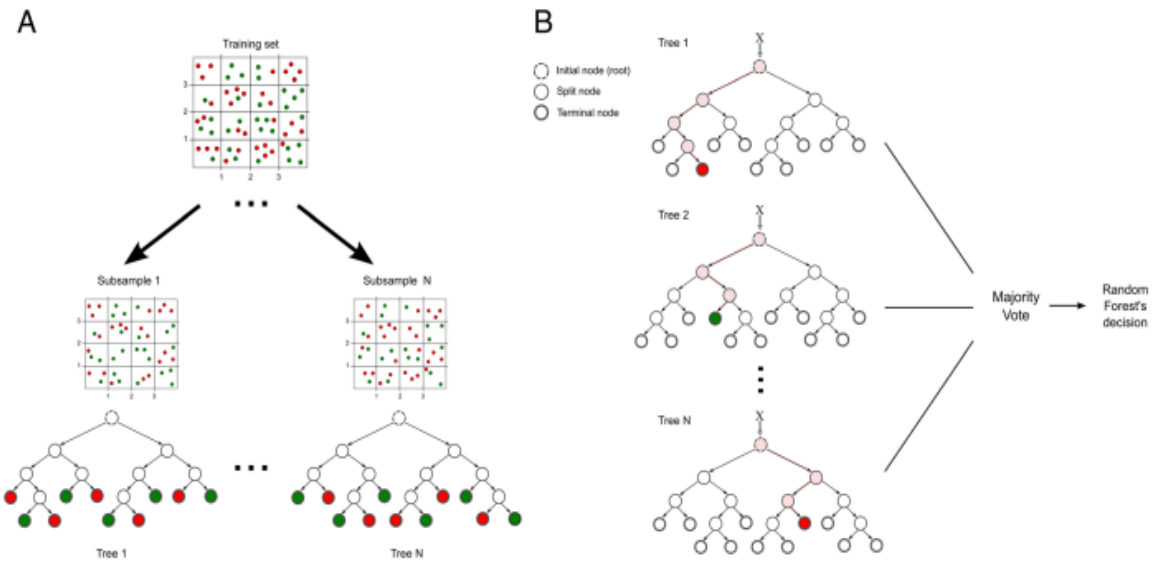


Figure 4. Example of training and classification processes using random forest.[15]

Assessing Models

Sensitivity, specificity, mean absolute error, and coefficient of determination (R^2) are commonly taught and appreciated ways to understand performance of many models. With ML it becomes a little more complicated given the dynamic nature of model creation. To reduce random chance and over fitting a few additional statistical techniques are needed. With large amounts of data, cross validation can be accomplished by dividing the data into three parts consisting of training data (50%), validation data (25%) and test data (25%). Training data is used to construct the model, the validation data is used to determine performance and error during development, and the test data is retained till the end of model creation to assess generalizability. Unfortunately, data is usually scarce and more advanced K-fold cross validation is needed where k is an integer usually less than 12, frequently 5 or 10. In K-fold cross validation, the data is randomly divided into k groups or folds of approximately equal samples in each fold.

Figure 5. With a large amounts of data simple validation can be used to divide the data into three parts. The training data (50-60%) is used to develop the model. The validation data (20-25%) is used to determine performance and error during development. The test data (20-25%) is held until the end to assess generalizability of the model.

Figure 6. When large amounts of data is not available k-fold cross validation can be used. Five fold cross validation is shown here. In fold 1, the first sample is held for validation and the other four sets are used for training. In the other folds different samples are used for validation. The cumilative error for k-fold cross validation is the average error of all the folds.

The most common form of assessment is area under the curve of the receiver operating characteristic (ROC) curve. An AUC of 0.8 is considered good and 0.5 random chance. Additional assessment values including Log-loss and F1 score are also valuable.[16,17]

	Actual Yes	Actual No
Predicted Yes	True Positive	False Positive
Predicted No	False Negative	True Negative

Table 3. Confusion matrix used to calculate many model performance statistics.

Figure 7. ROC curve used to calculate AUC values.

Ethics

Ethical issues on using data outside of a clinical mandate, and generalizing models built on biased source data will continue to plague ML, as it does general research. Most large longitudinal datasets continue to under enroll minorities, and datasets built on EHR data reflect the societal injustices inherit to our current medical model.[5] This was highlighted in the 2016 ProPublica research and article “Machine Bias”. They demonstrated how machine learning when trained on a historically biased dataset continued that bias prospectively in predicting likelihood of future crime.[18] This type of predictive failure could happen in many clinical cases and requires continuous supervision to prevent it from happening. Vayena et al. propose a

three category approach to understanding the ethical and regulatory concerns around ML shown in the below figure.[5,19]

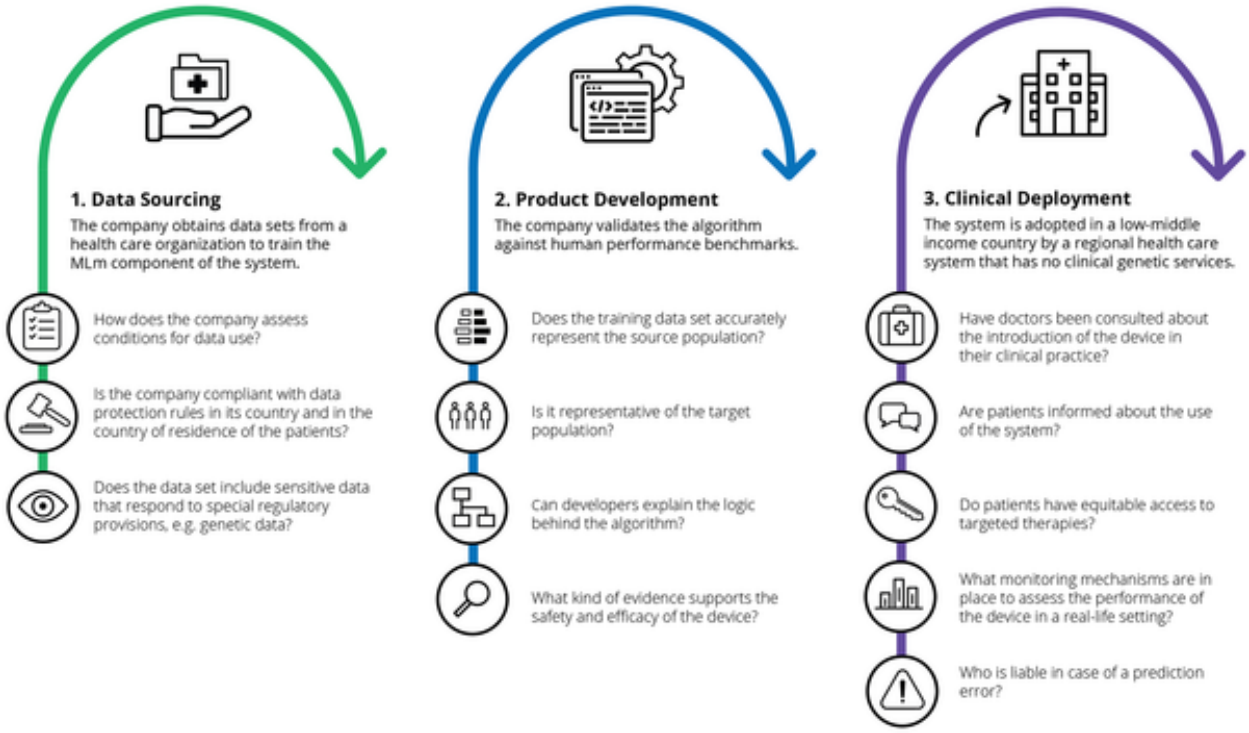


Figure 8. Key questions to ask at each stage of development and implementation of ML based tools.[19]

Use within Clinical Decision Support

Building the model is only start of a complex analysis and design process needed to create effective CDS. Shortliffe and Sepúlveda propose six well thought-out criteria for good ML based CDS. These are: 1. “Black boxes are unacceptable”, 2. “Time is a scarce resource”, 3. “Complexity and lack of usability thwart use”, 4. “Relevance and insight are essential”, 5. “Delivery of knowledge and information must be respectful”, and 6. “Scientific foundation must be strong”.[20] We strive to meet these ambitious criteria through scientific rigor, efficient CDS, sophisticated and mostly automatic use of finite state machines, and governance.

We have developed a methodology and structure using Epic EHR based Care Paths to implement and compare new predictive models with our current method for prediction and detection. By assigning predetermined thresholds from a predictive model to the transitions within the care paths, patients are easily segmented into various risk strata with which to deliver specific CDS. The ease of exchanging the underlying model or clinical decision rule without affecting how the end user interacts with the alerts makes this methodology quite novel.

We have implemented two sepsis care paths consisting of 3 risk states (Low, Moderate, High), a treatment state (Administered Antibiotics), and three disposition states (Admit Floor, Admit ICU, Discharge). The two care paths differ only by the transitions between the various risk states where one care path uses our hospital's rule-based method, modeled after SIRS criteria, while the other uses the Epic Sepsis Predictive Model. This method has allowed us to prospectively validate the discriminatory capacity of the Epic Sepsis Predictive model and compare it to our current rule-based method for detecting Sepsis.

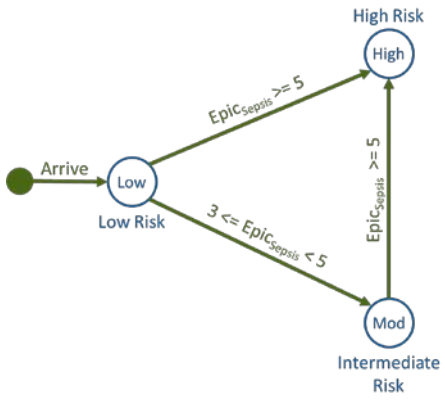


Figure 9. Risk triangle state diagram using Epic ML based score.

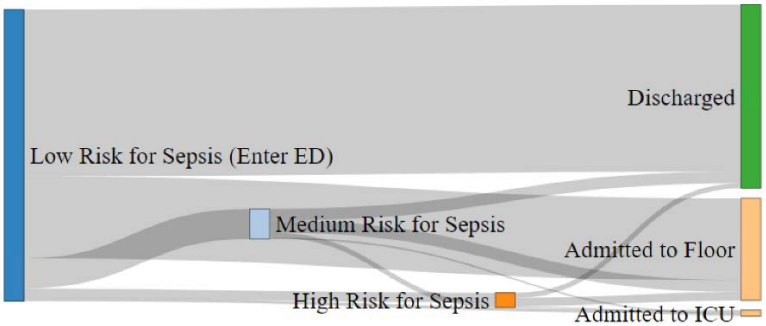


Figure 10. Sankey diagram shows sepsis risk transition in William P. Clemens Jr. University Hospital Emergency Department during silent evaluation period.

By modeling state transitions directly in the EHR, monitoring becomes significantly easier. With the state model many interesting questions become possible to calculate in near real-time. How many patients currently in the ED are at high risk for sepsis? How long does it take to identify patients who are at high risk for sepsis? What state takes the longest for patients who develop sepsis? Initial data demonstrate the value of this by reducing time to detection of high-risk patients by over 60 minutes.

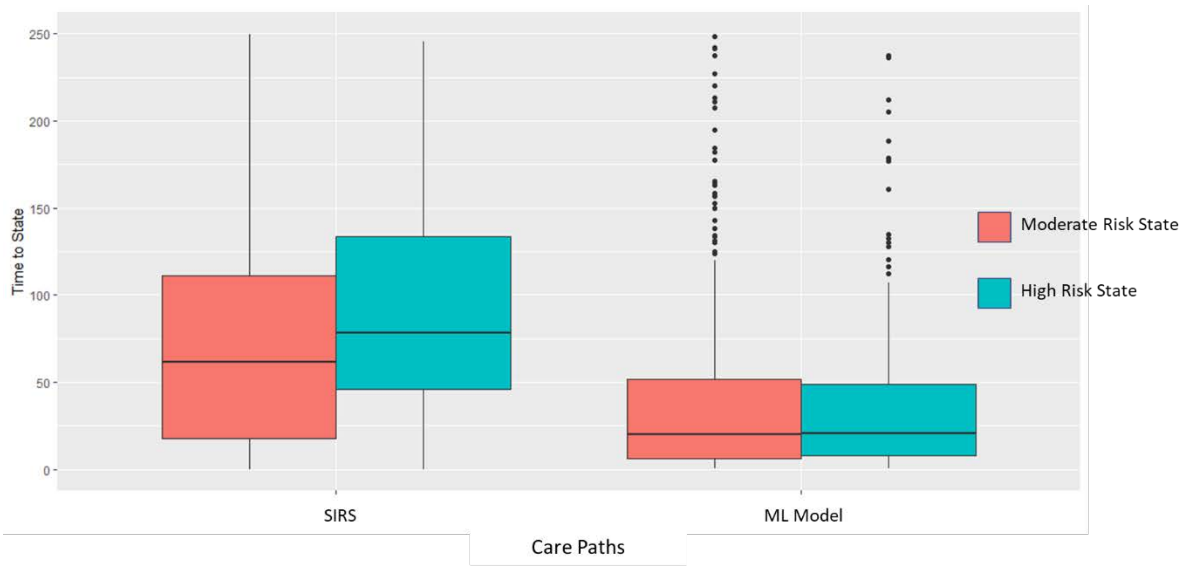


Figure 11. Data from silent evaluation period gives indication of the advantage of ML in earlier detection of high risk sepsis.

Full evaluation of the sepsis predictive model is currently underway but shows initial promise. In addition to sepsis, no-show prediction has completed early evaluation and is currently undergoing intervention design. Several other ML models are currently in early development. Although ML adds greater discrimination it does not reduce the design effort needed to bring high value CDS to production. Handling asynchronous state transition, when a risk score changes without any clinician being logged into the patient's chart, makes ML based CDS design and implementation more challenging than a rules-based system. Implementing ML based CDS at William P. Clemens Jr. University Hospital follows a multi-step process and is described in the below figure. Early involvement of a clinical informaticist is strongly recommended for bringing any ML model to production.

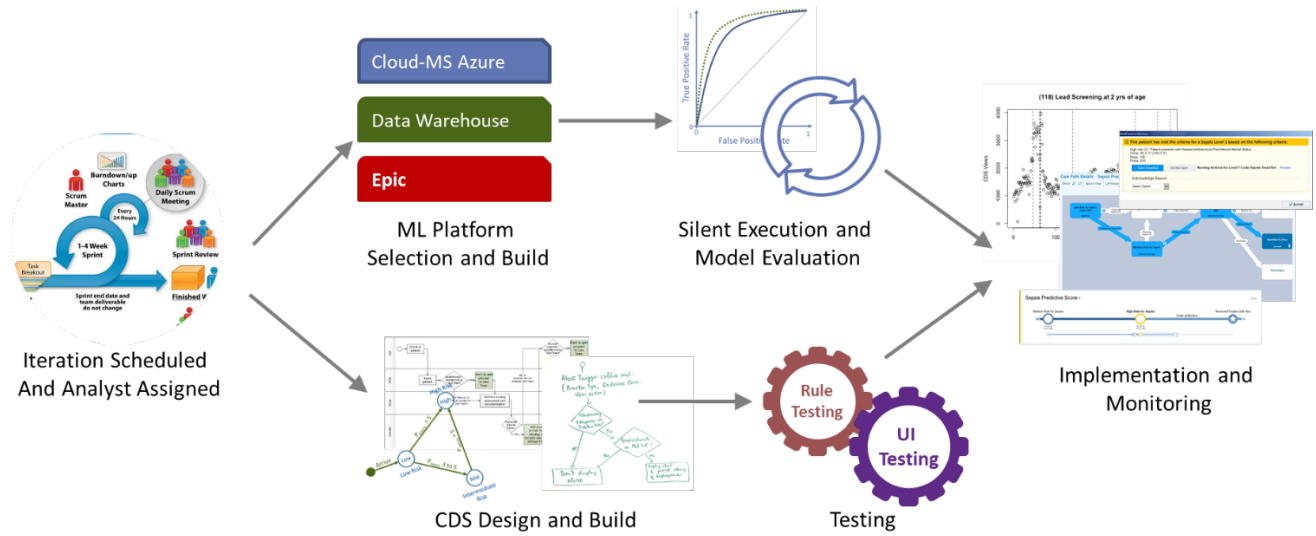


Figure 12. Process needed to bring ML based CDS to the bedside.

Conclusion

Machine learning has been around for 20 years and computation power, memory, and data availability has recently made it possible to implement these algorithms in EHRs for real-time CDS. The volume of data available for each patient is growing rapidly as data sharing across institutions becomes easier via health information exchanges. The amount of data will continue to be amplified by patient generated data from smart devices with increasingly powerful health sensors. We are approaching or are past the point where clinicians are no longer able to read the entirety of the patient chart. ML based patient summarization and data aggregation systems will no longer be a luxury but a necessity for good patient care.

The job of evaluating and monitoring these systems will be the responsibility of clinicians led by clinical experts and informaticists. Achieving this will require increased understanding of these models by all clinicians. Understanding of these models and increasing their adoption will require a change in education at all levels of clinical training. Sensitivity, specificity, PPV, NPV, and other traditional statistical assessment will continue to be valuable but need to be enhanced with AUC, ROC, F1 and Log-Loss in the context of these slightly more advanced models.

Models can be categorized as supervised, semi-supervised or unsupervised. Within these categories exist many different algorithms, each with their specific strengths and weaknesses. Deciding on data organization and specific model selection will require the help of a specialist. UTSW HSIR has built a framework for implementing complex models directly in the EHR and monitoring their performance in real-time. Using care paths with a finite state machine-based framework adds operational and long-term monitoring efficiencies hard to achieve otherwise. Early data shows ML based sepsis prediction reduces the number of alerts and speeds time to reach moderate and high-risk states, without adversely impacting predictive power.

Many fear ML models are coming to replace the clinician, but they are coming to enhance our effectiveness and reduce the amount of low value work. Many groups are working on or have functional systems to ease frustrating tasks such as billing, problem-based summarization, and discharge summary writing. They will not create a complete result as an experienced physician, but they are very close to creating a third-year medical school caliber result. ML is a tool and it is up to the clinician to accept its results and find value in it. It will not replace us but be our partner in improving the care of our patients.

> I conclude that though the individual physician is not perfectible,
> the system of care is, and that the computer will play a major part
> in the perfection of future care systems.
>
> ~ Clem McDonald, MD NEJM 1976[21]

References

1. Deo, R. C. Machine Learning in Medicine. *Circulation* **132**, 1920–1930 (2015).
2. Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* **23**, 89–109 (2001).
3. Tomayko, J. E. (James E. . Behind Deep Blue: Building the Computer that Defeated the World Chess Champion (review). *Technol. Cult.* (2003). doi:10.1353/tech.2003.0140
4. Kolachalama, V. B. & Garg, P. S. Machine learning and medical education. *npj Digit. Med.* **1**, 54 (2018).
5. Char, D. S., Shah, N. H. & Magnus, D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
6. The AI Industry Series: Top Healthcare AI Trends To Watch. Available at: https://www.cbinsights.com/research/report/ai-trends-healthcare/. (Accessed: 29th December 2018)
7. 106 Artificial Intelligence Startups In Healthcare. Available at: https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/. (Accessed: 29th December 2018)
8. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
9. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine* (2017). doi:10.1109/MSP.2017.2743240
10. Uchida, S. Image processing and recognition for biological images. *Development Growth and Differentiation* (2013). doi:10.1111/dgd.12054
11. Svensson, C.-M., Hübler, R. & Figge, M. T. Automated Classification of Circulating Tumor Cells and the Impact of Interobsever Variability on Classifier Training and Performance. *J. Immunol. Res.* **2015**, 1–9 (2015).
12. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* (2015). doi:10.1038/nature14539
13. Hinton, G. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA* **320**, 1101 (2018).
14. V, G., L, P. & M, C. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410
15. Machado, G., Mendoza, M. R. & Corbellini, L. G. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Vet. Res.* **46**, 85 (2015).

16. Giger, M. L. Machine Learning in Medical Imaging. *J. Am. Coll. Radiol.* (2018). doi:10.1016/j.jacr.2017.12.028

17. Chen, J. H. & Asch, S. M. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *N. Engl. J. Med.* **376**, 2507–2509 (2017).

18. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine Bias. *ProPublica* (2016). doi:http://dx.doi.org/10.1108/17506200710779521

19. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLOS Med.* **15**, e1002689 (2018).

20. Shortliffe, E. H. & Sepúlveda, M. J. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* **320**, 2199 (2018).

21. McDonald, C. J. Protocol-Based Computer Reminders, the Quality of Care and the Non-Perfectibility of Man. *N. Engl. J. Med.* **295**, 1351–1355 (1976).

Research Links

An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. http://www-bcf.usc.edu/~gareth/ISL/

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. https://web.stanford.edu/~hastie/ElemStatLearn/

MIT ML Online course: Rohit Singh, Tommi Jaakkola, and Ali Mohammad. *6.867 Machine Learning.* Fall 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Weka: https://www.cs.waikato.ac.nz/~ml/weka/

R Statistical software: https://cran.r-project.org/