In press: Journal of Electrocardiology

A crossroads in predictive analytics monitoring for clinical medicine

Full Institutional Affiliations

J Randall Moorman, M.D.

University of Virginia School of Medicine, Department of Medicine, Division of Cardiovascular Medicine

Complete Mailing Address for Presenting and Corresponding Author

J Randall Moorman, M.D.

1215 Lee Street

Box 800158

Charlottesville VA, 22908

USA

Email: rm3h@virginia.edu

Telephone: 434-982-3367

A crossroads in predictive analytics monitoring for clinical medicine

J Randall Moorman, M.D.

University of Virginia

ABSTRACT

A new goal for medical informatics is to develop robust tools that integrate clinical data on a patient in order to estimate the risk of imminent adverse events. This new field of *predictive analytics monitoring* is growing very quickly. Its claims, however, can be vulnerable when clinicians fail to use the best mathematical and statistical tools, when quantitative scientists fail to grasp the nuances of clinical medicine, and when either fails to incorporate knowledge of physiology. Its potential, though is clear: we can provide more effective clinical decision support and make better predictive analytics monitoring tools if we apply principles learned from physiology and mathematics to the right problems in clinical medicine.

A crossroads in predictive analytics monitoring for clinical medicine

J Randall Moorman, M.D.

University of Virginia

The pace quickens in the discussion of how data can be used to improve health. The appearance of the Electronic Health Record, with its vast warehouse capabilities, at around the same time as the ideas of Big Data and the renaissance of Deep Learning methods led naturally to the concepts of Personalized and Precision Medicine. It is an appealing vision to a clinician that all of a patient's data can be matched up against all the previous patients and all the published literature in such a way as to direct clinical care, especially in hospital and ICU medicine. Here, the focus is on the short-term, for example, the problem of patients that deteriorate under our noses. Unlike cancer, where the diagnosis often lies in the genes, the blueprints, it is the timeline that matters at the hospital and ICU bedside. Clinicians can agree that their own sixth senses are good for detecting early phases of deterioration, but we cannot be at every bedside all of the time. Another appealing vision is that all of the patient's data can be analyzed in real-time, and subtle changes can be identified in time to call the clinician before things go seriously wrong.

Of course, when I say "sixth sense," I am referring to the kind of predictive analytics monitoring that clinicians do in their heads all the time. This is data-based – we gather and assess patient information that is objective, and then process it using our experiential known patterns of abnormality. Thus, our intuition is really our data-based internal predictive analytics engine.

The current excitement is related but distinct – we now seek to train computers to do the same thing, and to present risk estimates to us. In my view, there are two approaches to developing automated predictive analytics monitoring for bedside use, and we are at a crossroads.

One approach is the old-fashioned one. Learn the physiological signatures of illness and learn how to detect them automatically from the real-time monitor. This requires deep knowledge of physiology and clinical

medicine as well as mathematical techniques to analyze the time series data. The more the principles are known, the fewer data are required. We put a man on the moon without any precedent data (and with a fraction of the computing power on your phone today), but we did have a good understanding of the laws of motion and gravity put forth by Newton in the 17th century. Another familiar and apt example of this old-fashioned predictive analytics is weather forecasting, which is based on systems of equations that describe how weather behaves.

The other is the new way. Use large computers to develop complex, many-layered neural networks that discover features in the data and learn how to recognize patterns. This requires only a large data set, the larger the better, but no knowledge whatsoever of the underlying physical system. Thus, the emerging paradigm is for statistical pattern recognition that does not take into account what is known about physiology and medicine. Here, I wish to challenge the unopposed and uncritical rise of this approach into widespread clinical use, and to temper and clarify the notion of predictive analytics monitoring by emphasizing and incorporating fundamental principles.

Road #1: Application of Deep Learning to the EHR

Early this year, Google published an important paper on predictive analytics in hospital patients that took the new way.[1] They studied 216K admissions of 114K adult patients to the hospitals of the University of Chicago and University of California – San Francisco, and extracted a very, very large number of data points – 47B, in all. The enormity arises from parsing each aspect of a data entry into separate *tokens.* Their example is that a single drug order contains multiple tokens - name, dose, ingredients, and so on. Moreover, they incorporated the entire EHR at the time of admission – on average, each patient already had nearly 150K tokens of information before the admission began, which added another 80K tokens.

The biggest bulk of data, though, were individual words of text from the notes, each of which was made a token. I found this very appealing. From the beginning of medical school onward, we are told to take a good history from the patient. “Listen to your patient,” said Sir William Osler, who knew a thing or two, “He is telling

you the diagnosis". The Google team incorporated every word from the entire EHR from its first encounter with the patient forward. Talk about taking a history!

They proceeded to use three state-of-the-art Deep Learning methods: one based on recurrent neural networks (Long Short-Term Memory, LSTM), one on an attention-based time-aware neural network model (TANN), and one on a neural network with boosted time-based decision stumps. The final risk estimate was an ensemble of the model outputs.

The outcomes were clinically relevant – in-hospital death (rate 2.3%), unplanned 30-day readmissions (12.9%), length of stay more than 7 days (approximately the top quartile, 23.9%). They compared the Deep Learning to conventional statistical models once per 12 hours starting 24 hours prior to admission and ending at discharge. A great strength of their study design was that they optimized the conventional models – National Early Warning Score, HOSPITAL score and Liu score – with newly-fit coefficients and sometimes with new variables. Another strength of the study was their capability to attribute the predictive ability to inputs. In their example, a woman with metastatic cancer who died in-hospital, it was interesting to see what played a role. It was terms including metastatic breast cancer, R lung malignant effusion, R lung empyema, R lung Pleurx (their hospital's brand name of chest tube), vancomycin, metronidazole, and others. Their results, measured as ROC areas, were very good. They were: in-patient mortality 0.94, readmission 0.76, and long LOS 0.86, all measurably better than the standard models even after their optimizations.

Is anything the matter with this approach? Perhaps not. The mathematics are straightforward and draw from conventional techniques like logistic regression, entropy estimation, maximum likelihood estimation, and calculus. But it is initially off-putting to know that no clinical or physiological insight went into the Deep Learning models. And there might be a new Deep Learning exercise required at every hospital, given the differences in practice with, for example, brand names of chest tubes. This would lead to problems – the benefit of Deep Learning over more shallow Machine Learning techniques is only evident when the data sets are very, very large, much larger than a single hospital might hold. This limits its use to large hospital systems, and, for the same reason, to outcomes that do not need clinician input. I suspect this was the reason that

Google chose readmissions, length of stay and death rather than the more nuanced diagnoses of sepsis or patient deterioration leading to ICU transfer. There would be no place for Deep Learning in early detection of events that are clinically nuanced and require time-consuming clinician effort for identification of individual cases with which to train the models.

This is one of the best results of Deep Learning in the field of EHRs. A recent extensive review [2] noted only a few successes, and accurately recounted the challenges of assembling enormous data sets in an age of preserving patient privacy from cyberattacks, of allowing clinicians to understand and interpret the model risk estimates, and having a sufficient number of clinically-adjudicated outcomes to make the expensive computational effort of Deep Learning worthwhile.

This excellent example of how Deep Learning can be applied to predictive analytics monitoring is very different from the classical method, an example of which now follows.

Road #2. Early detection of sepsis in premature infants

Sepsis is difficult to diagnose, and is particularly pernicious in premature infants.[3] The clinical problem is that the patient often weighs less than a kilogram, is immobile and mute, and there is no characteristic physical finding in the earliest, most treatable stages of the illness. Practitioners and the literature are consistent only in that the best indicator of sepsis in the premature infant is that the mother or yesterday's nurse feels that the baby is different today. Moreover, premature infants are much more than usually vulnerable to sepsis, as their immune systems are immature and their barriers to infection are reduced. In fact, the skin is often invaded by tubes and catheters, portals of entry for infecting organisms. As a result, sepsis in premature infants is common, reaching 25%.[4] The consequences of neonatal sepsis are dire. Mortality doubles, and there are often lifelong central nervous system deficits, principally delayed cognitive development.[5] Since, clearly, early treatment is best, neonatal clinicians are proactive when it comes to sepsis. At any moment, a quarter or more of premature infants in a neonatal ICU will be receiving antibiotic therapy for suspected or proven sepsis.

We approached the problem from the standpoint of what we knew about the time course of sepsis. While the clinical presentation might be sudden, the illness is not instantaneous, like a gunshot wound. The core process was taken to be systemic inflammation, and we were influenced by the work of Roger Bone, an intensivist in Chicago. In 1991[6], he crystallized the notion of the systemic inflammatory response syndrome, the idea that death from sepsis is not due to bacteria (they are wiped out early in the antibiotic course) but rather to an anarchic host immune response, one that brings down organs in its attempts to deal with the insult. Much work and thought has followed, and these are now mainstream ideas. From our viewpoint, the salient points were that multiple organs were responding, so the heart was a reasonable thing to look at, and that the time course was hours, not minutes. We were influenced by the new literature on heart rate variability and by the application of chaos theory to heart rate patterns. We made long recordings of EKG in the NICU, and examined RR interval time series, plots of the times between heart beats.[7,8] Quickly, we made a discovery:[9] the heart rate pattern changed during the hours prior to diagnosis. Instead of the familiar pattern of constant ups and downs on fast timescales and slow, there was nearly a fixed rate – a flat line in the RR interval time series – punctuated by short-lived decelerations. We were struck, after reflection, by the similarity to the abnormal heartbeat pattern of distress that many of us have seen first-hand while monitoring the heart rate pattern of the fetus during labor.

Whatever its origins, we considered this phenomenon to be of potential use in early detection of sepsis in the premature infant. We sought mathematical tools that captured the phenotype that we might employ in an automated way for all of the patients in our unit. We found that conventional measures or heart rate variability that had been developed in sick adults were not useful. There, the phenomenon of interest was simply reduced variability. While the baseline of reduced variability lent itself to these metrics, the decelerations inflated the overall variability, and conventional measures were blinded. Thus, we developed or optimized measures specifically for the task of identifying a time series that was often flat but occasionally spiked in only one direction. We soon found out that the canonical measures of standard deviation and the power spectrum were no different in the two examples. This is at first surprising, but the reason is that the decelerations inflate the standard deviation[9] and thus the power spectrum[10] (recall Parseval's Theorem, which leads to the finding that the area under the power spectrum is equal to the variance). Thus, we found that the math toolbox for

heart rate variability analysis in adults[11] was not useful.

As a result, we developed or adopted new RR interval time series measures that would detect the difference. First, we measure the width of the distribution by its *standard deviation*. Decelerations led many but not all abnormal records to have high standard deviation – these were effectively detected by the standard deviation. Second, we directly measured the weights of the two halves of the histogram by measuring separately the mean squared differences from the median. While similar to the third standardized moment, or skewness, the new measure allows separate consideration of the accelerations and decelerations. We called these values R1 and R2, and defined *sample asymmetry* as R2/R1.[12] Finally, we measured the *entropy* of the time series, a measure of its irregularity or uncertainty. The precept is that the highest entropy possible occurs with Gaussian random numbers. Thus, a distribution with a lower value of entropy is more non-Gaussian. Our finding[13] was that the most abnormal neonatal heart rate records were the most non-Gaussian, and had the lowest value of an entropy measure that Richman and I devised that we called *Sample Entropy (SampEn)*.[14] Entropy estimation is not just central to the neonatal sepsis story, but represents the kind of mathematics that clinicians would profit from familiarity with. Sample Entropy and multiscale entropy are very widely used – hundreds of citations yearly, each – and, while not exactly trivial, can be readily understood by everyone. That understanding brings with it more informed implementation and, more importantly, more informed interpretation. For example, we figured out that the low value of SampEn that was so useful in detecting sick babies had, in fact, nothing to do with reduced irregularity or uncertainty, but instead resulted from the non-stationarities – in our case, the heart rate decelerations[13]. Thus, while every paper that uses entropy estimation ends with conclusions about dynamics, some should not. Instead, the investigators would do well to think through why the entropy value came back low. Sometimes, as we described, it is due to non-stationarities, another mathematical matter altogether.

We used a standard biostatistical tool, multivariable logistic regression, to relate these measures – standard deviation, R1, R2 and SampEn – to the clinical events of neonatal sepsis, as determined by individual chart review of hundreds of NICU patients. We emerged with a clinical tool that provided neonatal clinicians with an estimate of risk that the patient would be ill the next day[15], obtained FDA 510(k) clearance, and began to show

a display.

Soon, we heard a success story. An infant was doing well, but the HRC index rose. They went to look at the baby but saw nothing wrong. They did the usual lab tests, and all were normal. But, because of what they knew about the predictive analytics monitoring, they also cultured the blood. The lab called in 12 hours to say that the blood was growing *Serratia marcescens*, a gram-negative organism. *Serratia* sepsis in the NICU is a dire illness, with nearly 40% mortality[16]. They looked again at the baby, who still looked well. They began antibiotics and the baby never got ill. The interpretation is that the illness was diagnosed early enough for antibiotics to be curative, and the high-mortality systemic inflammatory response syndrome never started.

To measure the impact of predictive analytics monitoring in the NICU, we performed the largest individually randomized controlled trial ever in neonatology. We randomized 3000 very low birth weight infants (<1500g), half of them extremely low birth weight (<1000g) in 9 NICUs cared for by >100 doctors and >1000 nurses. Babies were randomized to have the display shown, or not. There was no mandated intervention. The result was that infants with predictive analytics monitoring had decreased mortality, from 10.2% to 8.1% (p<0.05).[17] The survival benefit was more marked in ELBW infants (18% VS 13%) and those with sepsis (20% VS 12%).[4]

New as it is, the classical approach to predictive analytics monitoring draws from three rich traditions with long histories – physiology, mathematics, medicine. Its workers focus on fundamental themes: how living organisms work, the laws that the universe obeys, and how doctors and nurses take care of patients. The overview is this. From understanding how excitable tissues work in the brain, heart and lungs, we can know what to look for when illnesses distort them. From engineering and mathematics, we can know how to extract quantitative information from the signals that we can acquire from our patients. From clinical medicine, we can know which illnesses present themselves slowly enough for recognition and treatment to do some good.

Which road do we take?

How do we pick between these strategies, the classical and the new? While I do not think we need to pick one at the cost of excluding the other, I believe a limitation of the Deep Learning approach is that there is no attempt to seed the machine with things that we know to be useful. Were we to approach diagnosis of neonatal sepsis using Deep Learning (and we should), would we throw away what we know about abnormal heart rate characteristics of reduced variability and transient decelerations? Should we not include standard deviation, sample asymmetry and sample entropy as predictors? Such a highly specific question can, of course, be addressed directly through experiment. I suspect that our patients will profit more if we include everything we think we know about diagnosing illness into the mix along with all the unsorted bits of data, many of which we are convinced cannot be of any use.

We talk about prediction, but our goal is not to predict illnesses such as infection. Rather, we seek to diagnose them early in their course. What we are predicting is the clinical diagnosis, which can only come hours after the true inception of the illness. We wish to identify the earliest changes, and to make the earliest possible detection. Obviously, some kinds of illnesses are better suited to this exercise than others. Hence, a very great deal of clinical thinking needs to take place before any such investigation begins.

My point-of-view is that knowledge of the fundamentals of the relevant physiology, mathematics, and clinical medicine will produce better predictive analytics monitoring that will benefit more patients. Standing at the bedside of a patient that I have my doubts about, I would rather have a predictive analytics monitoring model fashioned in this classical way that targets this kind of patient and the illness and pathophysiology I am worried about than a more universal model targeting something else like 30-day readmission even if it were trained on more patients.

Bibliography and References Cited

1. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. *arXiv preprint arXiv:1801.07860*. 2018.

2. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):10.1098/rsif.2017.0387. doi: 20170387 [pii].

3. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. *The Lancet*. 2017.

4. Fairchild KD, Schelonka RL, Kaufman DA, et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr Res*. 2013;74(5):570-575. doi: 10.1038/pr.2013.136 [doi].

5. Stoll BJ, Hansen NI, Adams-Chapman I, et al. Neurodevelopmental and growth impairment among extremely low-birth-weight infants with neonatal infection. *JAMA*. 2004;292(19):2357-2365. doi: 292/19/2357 [pii].

6. Bone RC. The pathogenesis of sepsis. *Ann Intern Med*. 1991;115(6):457-469.

7. Aghili AA, Rizwan-uddin, Griffin MP, Moorman JR. Scaling and ordering of neonatal heart rate variability. *Phys Rev Lett*. 1995;74(7):1254-1257. doi: 10.1103/PhysRevLett.74.1254 [doi].

8. Griffin MP, Scollan DF, Moorman JR. The dynamic range of neonatal heart rate variability. *J Cardiovasc Electrophysiol*. 1994;5(2):112-124.

9. Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*. 2001;107(1):97-104.

10. Chang KL, Monahan KJ, Griffin MP, Lake D, Moorman JR. Comparison and clinical application of frequency domain methods in analysis of neonatal heart rate time series. *Ann Biomed Eng*. 2001;29(9):764-774.

11. Camm AJ, Malik M, Bigger JT, et al. Heart rate variability: Standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*. 1996;93(5):1043-1065.

12. Kovatchev BP, Farhy LS, Cao H, Griffin MP, Lake DE, Moorman JR. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. *Pediatr Res*. 2003;54(6):892-898. doi: 10.1203/01.PDR.0000088074.97781.4F [doi].

13. Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol*. 2002;283(3):R789-97. doi: 10.1152/ajpregu.00069.2002 [doi].

14. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*. 2000;278(6):H2039-49.

15. Griffin MP, O'Shea TM, Bissonette EA, Harrell FE,Jr, Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr Res*. 2003;53(6):920-926. doi: 10.1203/01.PDR.0000064904.05313.D2 [doi].

16. Stoll BJ, Hansen N, Fanaroff AA, et al. Late-onset sepsis in very low birth weight neonates: The experience of the NICHD neonatal research network. *Pediatrics*. 2002;110(2 Pt 1):285-291.

17. Moorman JR, Carlo WA, Kattwinkel J, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: A randomized trial. *J Pediatr*. 2011;159(6):900-6.e1. doi: 10.1016/j.jpeds.2011.06.044 [doi].