A Case Study in Evaluating the Clinical Utility of Early Warning Systems: HeRO

William E King, MS

Background

HeRO was proven to improve all-cause mortality by 22%,[1] and mortality after infection by 40%,[2] in the largest randomized controlled trial among premature infants, and after eight years of commercialization, HeRO monitoring has been adopted in 50 NICUs throughout the world, having monitored approximately 100,000 babies and saving the lives of a number likely in excess of 500.

Nevertheless, metrics describing the predictive performance of the HeRO Score may not have been adequately described.

No case better illustrates the difficulties in assessing the performance of HeRO than the slide in Figure 1, reprinted (including the caption) from a previous report.[3]

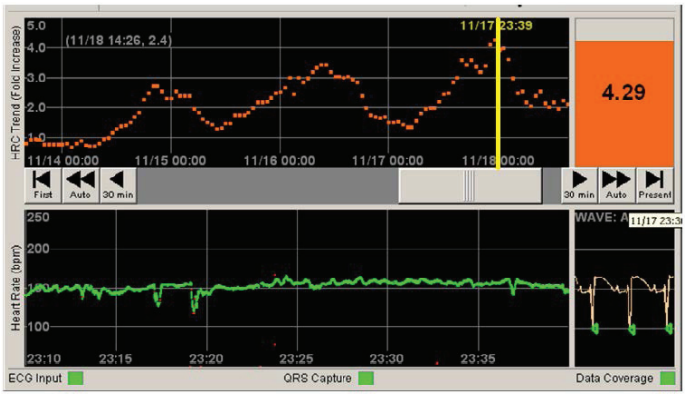

Figure 1. "In this case, the HeRO score acutely spiked from about 1 to progressively higher scores over a 4-day period. At the time indicated by the yellow line, the HeRO score was 4.29, and the associated 30-minute electrocardiogram shows decreased beat-to-beat variability with several superimposed decelerations. At that time, the patient had clinical signs of sepsis and a positive blood culture… Acute increases in HeRO score in the days prior to clinical deterioration may represent opportunities for earlier diagnosis and treatment, leading to better outcomes."

From a clinical standpoint, the HeRO trend in Figure 1 represents the potential for a major improvement in patient care. For three days prior to the diagnosis, the patient's HeRO Score was spiking. Evidence from the RCT would indicate that had this patient been randomized to the HeRO treated group and his or her HeRO Score had been displayed, the diagnosis might have been made at the first or second spike—*days earlier*—rather than during the third spike.[4] The literature is clear that earlier administration of antibiotics is more effective—even a single hour's delay is measurable in mortality.[5,6,7,8] It is this scenario, playing out time after time among the roughly 1500 patients randomized to HeRO-display, that led to the mortality improvement demonstrated in the RCT.

William E King is CEO of Medical Predictive Science Corporation.

Sensitivity, Specificity, and ROC

This was a clear clinical victory for HeRO monitoring, but how is it quantified statistically? Many clinicians are trained to evaluate a potential test based on sensitivity and specificity. The first problem is that the test, the HeRO Score, isn't binary, it's continuous. Every increment in HeRO Score adds risk to the patient—a score of 4.2 is worse than 4.1. But to calculate sensitivity and specificity, a single threshold must be chosen. Since around 10% of HeRO Scores are >2.0, representing at least double the predicted risk, this analysis will utilize that score as the threshold.

There is also a slightly more difficult problem: choosing a time window. Sensitivity and specificity have traditionally been used to measure a single-point-in-time test against the "gold standard" (or "reference standard") condition. HeRO is not a single-point-in-time test—HeRO is updated every hour continuously. But to calculate sensitivity and specificity, a time window, prior to the diagnosis, must be chosen. This decision is as crucial as the HeRO Score threshold because an elevated HeRO Score prior to the chosen time window will be evaluated as a False Positive, whereas inside the time window the same condition will be evaluated as a True Positive. Equally, a low HeRO Score prior to the time window is a True Negative, whereas inside the time window it is a False Negative.

Traditionally, our group has chosen the time window of 24 or 48 hours prior to the diagnosis, but the problem is obvious from Figure 2: under this scenario, the first and second spikes are False Positives. In fact, if we calculate based on the HeRO threshold of 2.0, and a 24-hour window, the points on the graph above yield a sensitivity and specificity of *merely 57% and 58%*, respectively. Yet if we expand our time window to 48 or 72 hours, the troughs in HeRO become False Negatives, and the sensitivity and specificity are not improved.

When the patient status is not clearly "sick" or "well," the analysis should ignore the associated data. During the time period commencing with the positive blood, urine, or CSF culture to 10 days following that culture, the patient status is

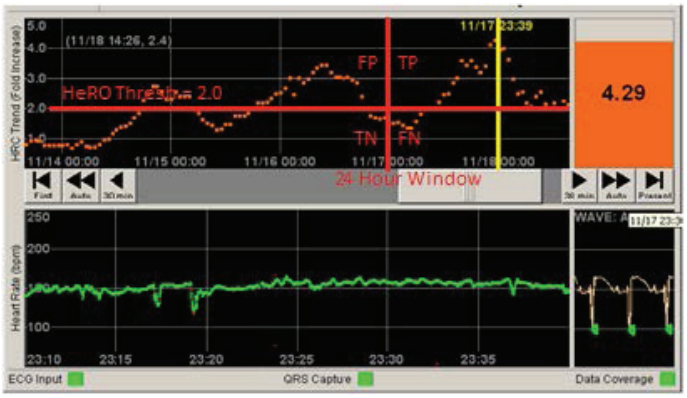

Figure 2. Thresholds defining True Positives, True Negatives, False Positives, and False Negatives, required to calculate Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value, imposed on the slide from Fig 1. Based on the data in this figure, Sensitivity and Specificity are 57% and 58%, respectively.

ambiguous: the patient is likely receiving therapy and recovering. He or she may still be symptomatic or they may have recovered quickly. For the purposes of analysis, these data clearly cannot be labeled as "sick", but neither can they be labeled "well". It is best to ignore the data in this time window from the analysis dataset.

In fact, if we apply these definitions (48-hour window in advance of *Positive* Blood, Urine, or CSF Culture, HeRO threshold of 2.0) to all 1500 VLBW control patients in the RCT, we get a Sensitivity of 44.7% and a Specificity of 86.5%. Sweeping the HeRO threshold over all possible values (0.0 to 7.0) allows us to calculate the Sensitivity and Specificity at each threshold, which yields the area under the curve of the ROC of 0.740.

Our group has reported using these sorts of definitions, despite the clear over-penalization visible in Figures 1 and 2.[9,10,11,12,13,14,15,16] But there are reasonable techniques to minimize these over-penalizations that paint a different picture of the model performance.

Optimizing the Test and Reference Standard

Penalizing the model for "spikes", ie peaks and troughs, can be avoided. Rather than analyze each individual hourly HeRO Score as a separate predictor, we will consolidate all of the HeRO Scores from a 24-hour period into a single number: the maximum HeRO Score. This makes clinical sense: clinicians tend to act based on the highest score over a reasonable period of time, and certainly 24 hours fits within the cyclical rhythm of NICU operation.

By the same token as ignoring data after the diagnosis, we will change the way we deal with the data days in advance of the culture. As seen in Figures 1 and 2, in many cases the HeRO Score will be elevated for days in advance of the culture. But in other cases, it will rise only the day before. From a clinical standpoint, both are acceptable—they both offer the opportunity for earlier diagnosis. How can we reward the model in both cases without penalizing it either?

Since HeRO Scores were statistically significantly different between HeRO-display and control patients in the RCT for at least 7 days prior to culture,[2] there is significant justification to choose a 10-day window *before* (in addition to after) the culture to be ignored from the analysis. So, only the data in the day prior to the day of the culture will be analyzed as "sick" (or more correctly, "about-to-become-sick"), and only data that is at least 10 days removed (either before or after) from a culture will be "well".

The effect of these two changes (analyzing only the highest HeRO Score in a 24-hour period, and ignoring the ambiguous data from well before the culture) has dramatic impact on our metrics of model performance, and the ROC rises from 0.740 to 0.761.

Many techniques have been described to deal with an imperfect reference standard.[17] Consider changing the definition of "sick" from *positive* culture to *any* culture. Certainly, the implications of clinical sepsis among neonates are debatable and beyond the scope of this report.[18] But there is no denying that the presence of a culture (of any result) is a very good surrogate for the clinician's belief at that point in time that the patient might be sick. In fact, our group found that the *presence* of laboratory tests is a better predictor of impending infection than the *results* of those laboratory tests. How can this be? The *presence* of laboratory tests is a surrogate for clinician intuition that the patient is sick, and it turns out that clinician intuition, based on a consideration of the full condition of the patient, is better than the results of the labs at predicting infection. Similarly, the presence of a culture—even a negative one—indicates that clinicians were suspicious of infection at that point in time.

Choosing the definition of "about-to-become-sick" as the presence of *any* culture in the next day, while choosing the definition of "well" as the absence of *any* culture within ±10 days further improves the ROC from 0.761 to 0.789.

This leads us to the approach that we propose as the most appropriate way to examine HeRO's efficacy as a predictor of infection. We choose "about-to-become-sick" patients as those who will have a *positive* culture drawn in the next day, while "well" patients will be those who are not within ±10 days of *any* culture. While it is clear that the patients labeled as sick or well are highly likely to be so, these definitions eliminate the messy middle ground—those patients whose status could be argued one way or another—from the analysis. Applying these definitions to the dataset, ROC improves from 0.789 to 0.821.

Forward looking metrics

But in the context of a continuous early warning system like HeRO, what do sensitivity, specificity, and ROC *actually mean*? In this context, Sensitivity is the probability that the HeRO Score will be above the threshold in the time window we have defined *prior to* the patient's diagnosis. In other words, it looks back in time, once it is already known that the patient is sick. Equally, specificity looks back from points in time where it is known that the patient was healthy.

Is this relevant from a clinical standpoint? It has been stated that "both sensitivity and specificity…are of no practical use when it comes to helping the clinician estimate the probability of disease in individual patients."[19] The point of an early warning system is to tell the clinician, *right now*, what is likely to happen to this patient *in the future*. Indeed, clinicians are prone to a failure of logic known as confusion of the inverse,[20] best illustrated by the scenario described as follows:[21]

"An example of this with regard to sensitivity, consciously chosen in a form that makes the problem clear, would be

Table 1. Summary of metrics of model performance for various types of analysis. Sensitivity, Specificity, PPV, NPV, and Risk Ratio all calculated at a HeRO threshold of 2.0. Dataset consists of the ~1500 VLBW patients randomized to HeRO non-display in the RCT.

Analysis of	Def'n of Sick	Def'n of Well	Sens	Spec	ROC	PPV	NPV	Risk Ratio
Each hourly HeRO Score	Pos Cx within 48 hrs	More than 10 days since Pos Cx	44.7%	86.5%	0.740	9.6%	98.0%	5x
Maximum daily HeRO Score	Pos Cx in the *next* day (up to 48 hrs)	More than ±10 days from Pos Cx	61.0%	79.8%	0.761	4.5%	99.2%	6x
Maximum daily HeRO Score	Any Cx in the *next* day (up to 48 hrs)	More than ±10 days from Any Cx	54.4%	89.0%	0.789	35.0%	94.7%	7x
Maximum daily HeRO Score	Pos Cx in the *next* day (up to 48 hrs)	More than ±10 days from Any Cx	61.0%	89.0%	0.821	12.1%	98.9%	11x

converting the logical proposition *This animal is a dog; therefore it is likely to have four legs* into the illogical proposition *This animal has four legs; therefore it is likely to be a dog.*"

Positive Predictive Value and Negative Predictive Value better address this *clinical problem* of identifying infection, because they both look *forward*. If a patient has a high HeRO Score, the Positive Predictive Value is the probability that he or she will be diagnosed with infection in the subsequent time period. Conversely, Negative Predictive Value starts with a low HeRO Score and represents the probability that the patient will remain healthy over the subsequent time period.

Indeed, it is Sensitivity that is inversely confused by clinicians with PPV (and Specificity with NPV). While Sensitivity, Specificity, and Area Under the Curve (AUC) of the Receiver-Operator Characteristic (ROC) Curve derived from them have their places, Positive and Negative Predictive Values better capture the importance of information provided to the practicing clinician who must make *forward looking* decisions.

But PPV and NPV are biased by a very important factor: the incidence rate of the event in the population. Is a PPV of 10% good? It depends upon the incidence rate. If 8% of patients have the condition, a 10% PPV isn't changing the picture very much. If 1% of the patients have the condition, a 10% PPV indicates that the patient is at 10x risk, and in the case of late onset neonatal sepsis, that would likely push clinicians toward labs, at the least.

When assessing late onset neonatal sepsis, this becomes problematic. When looking at all patients in a NICU, the incidence rate of late onset sepsis might be 2-5%; whereas looking at only VLBW patients, it might be 10-25%; and when looking at ELBW patients, it might be 20-50%.

This problem of incidence rates is even further exacerbated by the definition of sick—what to do (again) about clinical sepsis? Is it appropriate to penalize an early warning system that is elevated prior to a patient who deteriorates, gets sick, but doesn't grow a culture not because there were no bacteria in their blood, but because clinicians failed to draw enough blood *to determine* that there were bacteria in the patient's blood? Is that a True Positive or a False Positive? Whichever is chosen, the incidence rate is altered drastically. In the numerator, there are about 4x as many blood, urine, or CSF cultures as there are positive cultures in this dataset. In the denominator, over one third of patient days are within ±10 days of any culture. When cases are *positive* cultures and controls are ±10 days removed of *positive* culture, the incidence rate is 1.5%; when cases are *any* cultures and controls are ±10 days removed of *any* culture, the incidence rate is 9.8%; finally, when cases are *positive* cultures and controls are ±10 days removed of *any* culture, the incidence rate is 2.4%.

Both of these decisions, patient population and definition of infection, dramatically affect the incidence rate, and thereby the PPV and NPV, rendering both of those values nearly useless for the purpose of comparing model performance. Risk Ratio, however, is the PPV / (1-NPV), and the beauty there is that the incidence rate applies equally to both the numerator and denominator, cancelling itself out. We propose Risk Ratio as the best single metric of performance of an early warning system because it is forward looking *and* robust to changes in incidence rate caused by differing patient populations and definitions of "sick".

Table 1 summarizes the effect of each of the techniques described above on each of the metrics of model performance.

Discussion

Myriad clinical scoring systems, early warning systems, or rapid response systems have been described and implemented, and while some have been successful,[22] most large multicenter randomized controlled trials fail to show benefit;[23,24] and fewer than 15% of ICUs use severity scoring systems.[25] The difference in effect on outcome lies in the difference between *accuracy* and *actionability*. Though this report has spent considerable time describing the accuracy of HeRO to predict infection, many researchers have described models with higher ROC areas for predicting other clinical events that have failed to improve outcomes, whereas HeRO was proven to reduce all-cause in-hospital mortality, mortality after infection, and mortality measured at neurodevelopmental follow up assessments at 18 months. But a model will only improve outcomes if it improves the *actions* taken by clinicians. While something like the APACHE score may have better metrics of performance, it has proven to be simply a regurgitation of what the clinician already knows and is acting upon. HeRO, throughout its development process, was designed to provide *new information*, and we intentionally chose to not utilize demographic and laboratory information *despite the fact that they improved the ROC Area* and other metrics of performance. As opposed to repeating back to the clinicians something that they already know, we distilled HeRO down to new information that *changes clinical actions*. It may be that compromising accuracy for the sake of actionability is the key that led to the improvement in mortality. To quote Amelia Earhart, "The most difficult thing is the decision to act."

This report has been focused on esoteric statistics: ROC areas, predictiveness curves, risk ratios, etc.; but it will finish with this

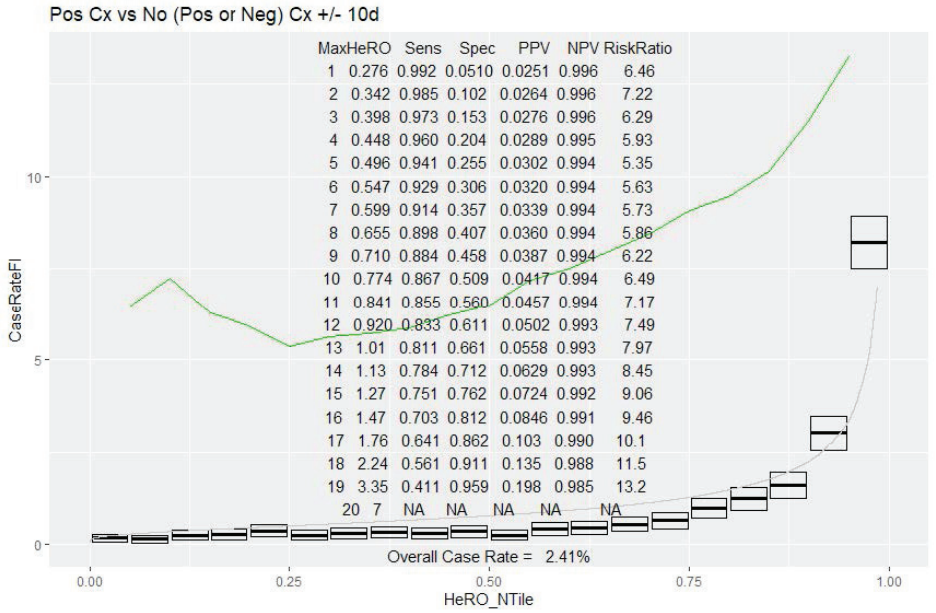

Figure 3. Predictiveness curve comparing highest HeRO Score in a 24-hour period with incidence of positive blood culture in the subsequent 24-hour period as cases, while excluding data within ±10 days of any culture from controls. The smooth gray line represents the HeRO score, i.e. the predicted risk of the event. The thick black lines represent the observed risk for patients in the ventile (20 equisized groups), with the box representing the 95% confidence intervals. The overlapping of the black boxes over the gray line indicates that the model is well calibrated, even in spite of the inherent bias caused by choosing the highest HeRO score in any 24-hour period. Finally, the green line, although plotted on the same scale, represents the Risk Ratio at the given threshold (e.g. patients above the 18th ventile, or 90th percentile of HeRO Score, are 11.5x more likely to have a positive blood culture in the subsequent 24-hour period than those below the 90th percentile).

final thought: these metrics of performance are what we use to *extrapolate* the impact of a test onto patient outcomes, because it is expensive and time consuming to conduct randomized controlled trials to actually find out. But, do we really care that 0.80 and 0.90 distinguish ROC areas that are "acceptable" from "excellent" from "outstanding",[26] whatever those terms signify? In the context of neonatal infection, do we know what sensitivity is required to improve the timing of antibiotic administration, or what negative predictive value is required to improve antibiotic stewardship? The point is that the performance metrics of HeRO are immaterial because *we ran the RCT*. HeRO monitoring yields *actionable* information that changes clinician behavior, resulting in reduced mortality, reduced length of stay, and improved targeting of antibiotics. The rest are simply the means.

References

1 Moorman JR, Carlo WA, Kattwinkel J, Schelonka RL, Porcelli PJ, Navarrete CT, Bancalari E, Aschner JL, Walker MW, Perez JA, Palmer C, Wagner DP, Stukenborg GJ, Lake DE, O'Shea TM. Mortality benefit of heart rate characteristics monitoring in very low birth weight infants: a randomized trial. Journal of Pediatrics. 2011. Editorial accompanies.

2 Fairchild KD, Schelonka RL, Kaufman DA, Carlo WA, Kattwinkel J, Porcelli PJ, Navarrete CT, Bancalari E, Aschner JL, Walker MW, Perez JA, Palmer C, Lake DE, O'Shea TM, Moorman JR. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. Pediatr Res. (2013) Aug 13. doi: 10.1038/pr.2013.136.

3 Fairchild KD, Aschner JL. HeRO monitoring to reduce mortality in NICU patients. P Research and Reports in Neonatology (2012) Aug 14.

4 Sullivan BA, Grice SM, Lake DE, Moorman JR, Fairchild KD. Infection and Other Clinical Correlates of Abnormal Heart Rate Characteristics in Preterm Infants. J Pediatr. (2014) Jan 9.

5 Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med. 2006 Jun;34(6):1589-96.

6 Clec'h C1, Timsit JF, De Lassence A, Azoulay E, Alberti C, Garrouste-Orgeas M, Mourvilier B, Troche G, Tafflet M, Tuil O, Cohen Y. Intensive Care Med. Efficacy of adequate early antibiotic therapy in ventilator-associated pneumonia: influence of disease severity. 2004 Jul;30(7):1327-33. Epub 2004 Jun 9.

7 Køster-Rasmussen R1, Korshin A, Meyer CN. Antibiotic treatment delay and outcome in acute bacterial meningitis. J Infect. 2008 Dec;57(6):449-54. doi: 10.1016/j.jinf.2008.09.033. Epub 2008 Nov 9.

8 Lodise TP1, McKinnon PS, Swiderski L, Rybak MJ. Outcomes analysis of delayed antibiotic treatment for hospital-acquired Staphylococcus aureus bacteremia. Clin Infect Dis. 2003 Jun 1;36(11):1418-23. Epub 2003 May 20.

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

10 Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. Pediatric Research, 53: 920-926, 2003.

11 Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal heart rate characteristics are associated with neonatal mortality. Pediatric Research, 55:782-788, 2004.

12 Griffin MP, Lake DE, Moorman JR. Heart rate characteristics and laboratory tests in the diagnosis of neonatal sepsis. Pediatrics 115: 937-941, 2005.

13 Griffin MP, Lake DE, Bissonette EA, Harrell FE, O'Shea TM, Moorman JR. Heart rate characteristics: novel physiomarkers to predict neonatal infection and death. Pediatrics 116:1070-4,

2005.
14 Griffin MP, Lake DE, O'Shea TM, Moorman JR. Heart rate characteristics and clinical signs in late-onset neonatal sepsis. Pediatric Research, 61: 222-227, 2007.
15 Lake DE, Fairchild KD, Moorman JR. Complex signals bioinformatics: evaluation of heart rate characteristics monitoring as a novel risk marker for neonatal sepsis. J Clin Monit Comput. (2013) Nov 19.
16 Clark MT, Vergales BD, Paget-Brown AO, Smoot TJ, Lake DE, Hudson JL, Delos JB, Kattwinkel J, and Moorman JR. Predictive monitoring for respiratory decompensation leading to urgent unplanned intubation in the neonatal intensive care unit. Ped Research. (2013) January.
17 Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment Volume: 11, Issue: 50, Published in November 2007 https://doi.org/10.3310/hta11500
18 Klingenberg, Kornelisse, Buonocore, Maier, and Stocker. Culture-Negative Early-Onset Neonatal Sepsis - At the Crossroad Beween Efficient Sepsis Care and Antimicrobial Stewardship. Front Pediatr, 2018.
19 Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr (2006) 96:338–41.10.1111/j.1651-2227.2006.00180.
20 Misunderstandings Involving Conditional Probabilities. Available from: https://www.ma.utexas.edu/users/mks/statmistakes/misundcond.html
21 Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. (2017). Front. Public Health 5:307. doi: 10.3389/fpubh.2017.00307
22 Haegdorens F, Monsieurs KG, De Meester K, Van Bogaert P. An intervention including the national early warning score improves patient monitoring practice and reduces mortality: A cluster randomized controlled trial. J Adv Nurs. 2019 Sep;75(9):1996-2005. doi: 10.1111/jan.14034. Epub 2019 Jun 6.
23 Parshuram CS, Dryden-Palmer K, Farrell C, et al. Effect of a pediatric early warning system on all cause mortality in hospitalized pediatric patients. JAMA. 2018:1–11.
24 Hooper MH, Weavind L, Wheeler AP, Martin JB, Gowda SS, Semler MW, Hayes RM, Albert DW, Deane NB, Nian H, Mathe JL, Nadas A, Sztipanovits J, Miller A, Bernard GR, Rice TW. Randomized trial of automated, electronic monitoring to facilitate early detection of sepsis in the intensive care unit. Crit Care Med. 2012 Jul;40(7):2096-101. doi: 10.1097/CCM.0b013e318250a887. PMID: 22584763
25 Breslow MJ, Badawi O. Severity scoring in the critically ill: part 1–interpretation and accuracy of outcome prediction scoring systems. Chest. 2012 Jan;141(1):245-252. doi: 10.1378/chest.11-0330.
26 DW Hosmer, S Lemeshow. Applied Logistic Regression, 2nd Ed. Chapter 5, John Wiley and Sons, New York, NY (2000), pp. 160-164