Background

HeRO was proven to improve all-cause mortality by 22%, and mortality after infection by 40%, 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.

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. The literature is clear that earlier administration of antibiotics is more effective—even a single hour’s delay is measurable in mortality.

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.

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,7 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.

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

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 any 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
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”?, 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


