PREMATURE INFANTS

Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants

Suchi Saria,[1] Anand K. Rajani,[2] Jeffrey Gould,[2] Daphne Koller,[1]* Anna A. Penn[2]*

(Published 8 September 2010; Volume 2 Issue 48 48ra65)

Physiological data are routinely recorded in intensive care, but their use for rapid assessment of illness severity or long-term morbidity prediction has been limited. We developed a physiological assessment score for preterm newborns, akin to an electronic Apgar score, based on standard signals recorded noninvasively on admission to a neonatal intensive care unit. We were able to accurately and reliably estimate the probability of an individual preterm infant's risk of severe morbidity on the basis of noninvasive measurements. This prediction algorithm was developed with electronically captured physiological time series data from the first 3 hours of life in preterm infants (≤34 weeks gestation, birth weight ≤2000 g). Extraction and integration of the data with state-of-the-art machine learning methods produced a probability score for illness severity, the PhysiScore. PhysiScore was validated on 138 infants with the leave-one-out method to prospectively identify infants at risk of short- and long-term morbidity. PhysiScore provided higher accuracy prediction of overall morbidity (86% sensitive at 96% specificity) than other neonatal scoring systems, including the standard Apgar score. PhysiScore was particularly accurate at identifying infants with high morbidity related to specific complications (infection: 90% at 100%; cardiopulmonary: 96% at 100%). Physiological parameters, particularly short-term variability in respiratory and heart rates, contributed more to morbidity prediction than invasive laboratory studies. Our flexible methodology of individual risk prediction based on automated, rapid, noninvasive measurements can be easily applied to a range of prediction tasks to improve patient care and resource allocation.

INTRODUCTION

Early, accurate prediction of a neonate's morbidity risk is of significant clinical value because it allows for customized medical management. The standard Apgar score has been used for more than 50 years to assess neonatal well-being and the need for further medical management. We aimed to develop a modern tool akin to an "electronic" Apgar assessment that reflects a newborn's physiological status and is predictive of future illness severity. Such an improvement in neonatal risk stratification may better inform decisions regarding aggressive use of intensive care, need for transport to tertiary centers, and resource allocation, thus potentially reducing the estimated $26 billion per year in U.S. health care costs resulting from preterm birth (*1*). Gestational age and birth weight are highly predictive of death or disability (*2*) but do not estimate individual illness severity or morbidity risk (*3*). These perinatal risk factors, in addition to laboratory measurements, have been incorporated into currently used algorithms for mortality risk assessment of preterm infants (*4*–*6*). These algorithms, however, predict mortality rather than morbidity (*3*). They also rely on invasive testing and require extraction of data from multiple sources to make a risk assessment.

Although it has been recognized that changes in heart rate characteristics (*7*) or variability (*8*) can suggest impending illness and death in a range of clinical scenarios, from sepsis (*9*) in intensive care patients to fetal intolerance of labor (*10*), the predictive accuracy of a single parameter is limited. Intensive care providers observe multiple physiological signals in real time to assess health, but certain informative patterns may be subtle. To achieve improved accuracy and speed of individual morbidity prediction for preterm neonates, we developed a new probability score (PhysiScore) based on physiological data obtained noninvasively after birth along with gestational age and birth weight. Two recent advances enabled the use of multiple complex physiological signals for this purpose: the digitization of medical records, which allows linking of real-time physiological signals with later outcomes, and the increasing sophistication of machine learning and pattern recognition algorithms, which allows optimization of PhysiScore in an automated, unbiased manner. We evaluated PhysiScore's use for predicting overall morbidity and mortality, specific risk for infants with infection or cardiovascular and pulmonary complications, and a combination of complications associated with poor long-term neurodevelopment and compared its performance to standard scoring systems in a preterm neonatal cohort.

RESULTS

PhysiScore development based on patient characteristics and morbidities

To develop our prediction tool, we studied a total of 138 preterm neonates that were 34 weeks gestational age or less and <2000 g in weight without major congenital malformations and with baseline characteristics and morbidities as shown in Table 1. Mean birth weight was 1367 g at an estimated mean gestational age of 29.8 weeks, placing these infants at significant risk of both short- and long-term complications.

Patients were then classified as high morbidity (HM) or low morbidity (LM) on the basis of their illnesses. The HM group was defined as any patient with major complications associated with short- or long-term morbidity. Short-term morbidity complications included culture-positive sepsis, pulmonary hemorrhage, pulmonary hypertension, and

[1] Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [2] Department of Pediatrics, Division of Neonatal and Developmental Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA.
*To whom correspondence should be addressed. E-mail: apenn@stanford.edu (A.A.P.); koller@cs.stanford.edu (D.K.)

Table 1. Baseline and disease characteristics of the study cohort. SGA, small for gestational age; NOS, not otherwise specified.

Category	
Subjects (*N*)	138
Birth weight (g)	1367 ± 440
Gestational age (weeks)	29.8 ± 3
Gender, female	68
Apgar score at 5 min	7 ± 3
SGA (≤5th percentile)	7
Multiple gestation	
Total	46
Twins	20
Triplets	6
Respiratory distress syndrome	112
Pneumothorax	10
Bronchopulmonary dysplasia	
Total	29
NOS*	2
Mild	12
Moderate	5
Severe	10
Pulmonary hemorrhage	2
Pulmonary hypertension	3
Acute hemodynamic instability	11
Retinopathy of prematurity (ROP)†	
Total	25
Stage I	9
Stage II	12
Stage III	4
Intraventricular hemorrhage (IVH)‡	
Total	34
Grade 1	19
Grade 2	7
Grade 3	3
Grade 4	5
Posthemorrhagic hydrocephalus	6
Culture-positive sepsis	11
Necrotizing enterocolitis	
Total	8
Stage 1	2
Stage 2	4
Stage 3	2
Expired	4

*Infants with oxygen requirement at 28 days for whom oxygen requirement was not known at 36 weeks after menstrual age. †ROP is counted by the most severe stage in either eye during the hospitalization. ‡IVH is counted by the most severe grade in either cerebral hemisphere by Papile classification.

Table 2. Performance summary with AUCs.

	Apgar	SNAP-II	SNAPPE-II	CRIB	PhysiScore
Predicting high morbidity	0.6978	0.8298	0.8795	0.8509	0.9151
Infection	0.7412	0.8428	0.9087	0.8956	0.9733
Cardiopulmonary	0.7198	0.8592	0.9336	0.9139	0.9828

acute hemodynamic instability. Long-term morbidity was defined by moderate or severe bronchopulmonary dysplasia (BPD), retinopathy of prematurity (ROP) stage 2 or greater, intraventricular hemorrhage (IVH) grade 3 or 4, and necrotizing enterocolitis (NEC) on the basis of the strong association of these complications with adverse neurodevelopmental outcome. Death was also included in the long-term morbidity group. Most infants in the HM category had short- and long-term complications affecting multiple organ systems. Infants with only common problems of prematurity such as mild respiratory distress syndrome (RDS) and patent ductus arteriosus (PDA) without major complications were classified as LM.

Probabilistic score for illness severity

We developed a method to estimate the probability that an infant would be in the HM category on the basis of physiological signals recorded in the first 3 hours of life plus gestational age and birth weight. This time period was selected for analysis because it is less likely to be confounded by medical interventions and provides prediction early enough in the infant's life to be useful for planning therapeutic strategy.

First, we processed the physiological signals (heart rate, respiratory rate, and oxygen saturation) that were recorded for all infants for the first 3 hours after birth. Mean values plus baseline and residual variability signals (capturing both short- and long-term variability) were calculated for heart and respiratory rates. Mean oxygen saturation and the ratio of hypoxia (oxygen saturation <85%) to normoxia over the 3-hour span were calculated.

We then defined the probability for illness severity with a logistic function that aggregated individual risk features as

$$P(\mathrm{HM}|v_1,v_2,\ldots,v_n) = \left(1+\exp\left(b+w_0{}^*c+\sum_{i=1}^{n} w_i{}^*f(v_i)\right)\right)^{-1} \quad (1)$$

where n was the number of risk factors and $c = \log P(\mathrm{HM})/P(\mathrm{LM})$ was the a priori log odds ratio. The ith characteristic, v_i (physiological parameter, gestational age, or weight) was used to derive a numerical risk feature $f(v_i)$ via nonlinear Bayesian modeling (detailed in Materials and Methods). The score parameters b and w were learned from the training data for use in prospective risk prediction. The parameter w_i represents the weight of the contribution of the ith characteristic to the computed probability score, with higher weight characteristics having a greater effect.

PhysiScore is a probability score that ranges from 0 to 1, with higher scores indicating higher morbidity. PhysiScore is calculated by integrating the following 10 patient characteristics into Eq. 1: mean heart rate, base and residual variability; mean respiratory rate, base and residual variability; mean oxygen saturation and cumulative hypoxia time; gestational age and birth weight. Each of these patient characteristics carries a specific learned weight, as denoted by w in Eq. 1. Plotting the receiver operating characteristic (ROC) curve (Fig. 1A) and associated area under the curve (AUC) values (Table 2) shows that PhysiScore exhibits good discriminative ability for prediction of morbidity and mortality risk and compares it to other risk assessment tools. Specifically, PhysiScore was compared to the Apgar score, long used as an indicator for the base physiological state of the newborn (*11*), as well as to extensively validated neonatal scoring systems that require invasive laboratory measurements [Score for Neonatal Acute Physiology-II (SNAP-II) (*5*), SNAP Perinatal Extension-II (SNAPPE-II) (*5*), and Clinical Risk Index for Babies (CRIB) (*6*)]. For making predictions with the Apgar score, we constructed a model as in Eq. 1 using the 1- and 5-min Apgar scores as the only two inputs; this combined model outperformed either of the two Apgar scores when used in isolation. PhysiScore (AUC 0.9197) performed well across the entire range of the ROC curve and significantly better (P < 0.003) (*12*) than all four of the other comparison scores (Table 2). PhysiScore's largest performance gain occurred in the high-sensitivity/specificity region of the ROC curve. Setting a user-defined threshold based on desired sensitivity and specificity allows optimization for individual settings. For example, in our neonatal intensive care unit (NICU), a threshold of 0.5 achieves a sensitivity of 86% at a specificity of 95% for HM, as seen in Fig. 1A (inset). Alternately, the use of a lower threshold would improve sensitivity at the expense of specificity.

We added the values obtained from laboratory tests to determine the magnitude of their contribution to risk prediction beyond the PhysiScore alone (Fig. 1B), incorporating parameters included in standard risk prediction scores (for example, SNAPPE-II): white blood cell count, band neutrophils, hematocrit, platelet count, and initial blood gas measurement of $P\mathrm{aO}_2$ (partial pressure of oxygen, arterial), $P\mathrm{aCO}_2$ (partial pressure of carbon dioxide, arterial), and pH (if available at <3 hours of age). No additional discriminatory power was achieved, suggesting that laboratory information is largely redundant with the patient's physiological characteristics.

To further assess the performance of PhysiScore, we analyzed prediction performance for infants in major morbidity categories. Specifically, we extracted two categories: infection (NEC, culture-positive sepsis, uri-

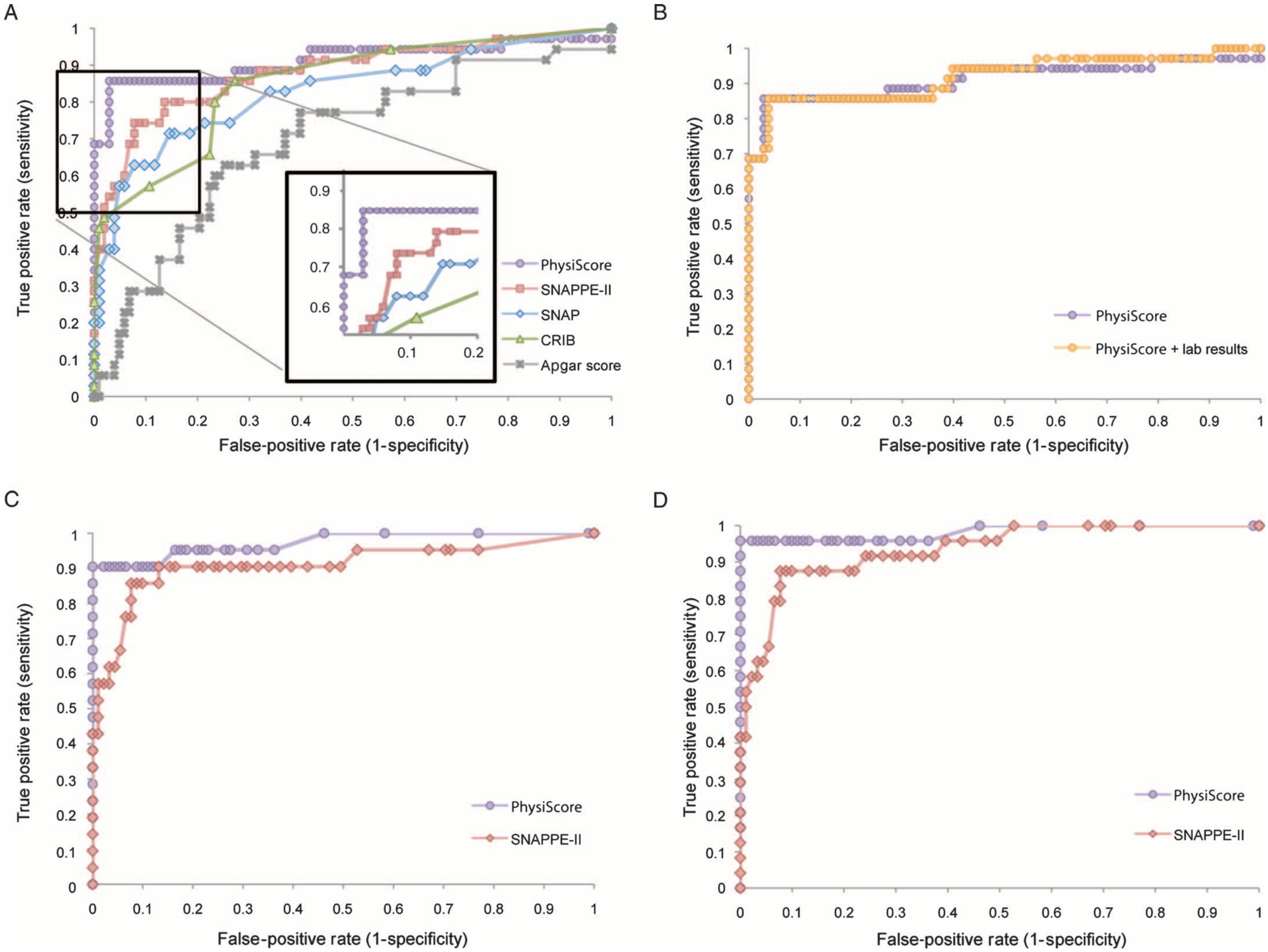

Fig. 1. (**A**) ROC curves demonstrating PhysiScore's performance in predicting high morbidity as it relates to conventional scoring systems. (**B**) PhysiScore's performance with laboratory studies. (**C**) Predictions for infants with infection-related complications. (**D**) Predictions for infants with major cardiopulmonary complications.

nary tract infection, and pneumonia) (Fig. 1C) and cardiopulmonary complications (BPD, hemodynamic instability, pulmonary hypertension, and pulmonary hemorrhage) (Fig. 1D). Plotting data from the infants in the HM category who had a specific complication against data from all infants in the LM category yields ROC curves for discriminative ability for HM infants in these independent morbidity categories (Fig. 1, C and D). Comparison to SNAPPE-II (the best-performing standard score) is also shown; AUCs were calculated for all scoring methods (Table 2) in these specifically defined sets. At a threshold of 0.5, PhysiScore achieves near-perfect performance (infection: 90% sensitivity at 100% specificity; cardiopulmonary: 96% at 100%).

Morbidity is most difficult to predict in patients with isolated IVH, for which all scores exhibit decreased sensitivity. The PhysiScore AUC for any IVH was 0.8092, whereas SNAP-II, SNAPPE-II, and CRIB had AUCs of 0.6761, 0.6924, and 0.7508, respectively. PhysiScore did not identify three infants who had severe IVH (grade 3 or 4) in the absence of any other HM complications. However, most infants who developed IVH can be found on the left side of the ROC, suggesting that PhysiScore offers high sensitivity without significant compromise in specificity (Fig. 2).

Importance of physiological features

Ablation analysis (comparison of model performance when different subsets of risk factors are included) was used to examine the contribution of score subcomponents in predicting HM versus LM. As expected, gestation and birth weight alone achieved reasonable predictive performance (AUC 0.8517). However, these two characteristics are not sufficient for individual risk prediction (*3*). Notably, physiological parameters alone were more predictive than laboratory values alone (AUC, 0.8540 versus 0.7710, respectively). Adding physiological parameters to gestation and birth weight (that is, PhysiScore) increased the AUC to 0.9129, a significantly ($P < 0.01$) (*12*) better prediction than gestation and birth weight alone. Addition of laboratory values and physiologic characteristics did not significantly increase the AUC (0.9197), again suggesting that these parameters are redundant with the laboratory data in morbidity prediction.

Examination of the learned weights (w_i in Eq. 1) of individual physiological parameters incorporated into PhysiScore (Fig. 3A) demonstrated that short-term heart and respiratory rate variability made a significant contribution to the value of the PhysiScore, but long-term variability did not. Thus, short-term variability patterns—often difficult to see by eye, but easily calculated by PhysiScore—carried significant physiological information that long-term variability patterns did not.

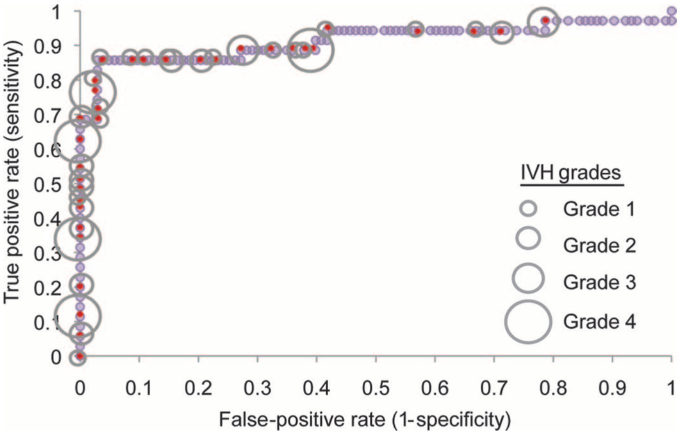

Fig. 2. ROC curve demonstrating the limited sensitivity of PhysiScore in predicting morbidity for infants with IVH. Each circle represents the IVH grade of a preterm neonate overlaid on the respective score.

Only three categories of commonly obtained physiological measurements were required for PhysiScore: heart rate, respiratory rate, and oxygen saturation. From these measures, using Bayesian modeling, we obtained individual curves that convey the probability of HM associated with individually calculated physiological parameters (Fig. 3B).

As expected, a respiratory rate between 35 and 75 breaths per minute had a greater probability of being associated with health, whereas higher or lower rates carried a greater probability of morbidity. A decreased short-term heart rate variability also indicated increased risk, consistent with previous findings linking this parameter to sepsis (*9*). This visual analysis of the nonlinear relationships seen in Fig. 3B also suggests unexpected associations. Short-term respiratory rate variability, not commonly used as a physiological marker, was associated with increased morbidity risk. Unlike residual heart rate variability, its effect was nonmonotonic. Risk curves describing oxygen saturation suggest, respectively, that risk increases significantly with mean saturations less than 92% and prolonged time spent (>5% total time) at oxygen saturations below 85%. Oxygenation is routinely manipulated by physician intervention, suggesting that intervention failure (for example, the inability to keep saturations in a specific range) that allows desaturations lasting for >5% of total time is associated with higher morbidity risk, a threshold that can now be prospectively assessed in clinical trials.

DISCUSSION

We have developed a risk stratification method that predicts morbidity for individual preterm neonates by integrating multiple continuous physiological signals from the first 3 hours of life. This score is analogous to the Apgar score (*11*), in that only physiological observations are used to derive morbidity and mortality predictions. However, the use of time series data combined with automated score calculation yields significantly more information about illness severity than is provided by the Apgar score.

Discriminative capacity

Past efforts have resulted in several illness severity scores that use laboratory studies and other perinatal data to achieve improved discriminative ability over the Apgar score alone. For all of the available neonatal illness scores, much of the discriminative ability comes from gestational age and birth weight. Nevertheless, it is well-recognized that age- and weight-matched neonates may have significantly different morbidity profiles (*3*). The CRIB score uses logistic regression to define six factors and their relative weights in predicting mortality: birth weight, gestational age, congenital malformation, maximum base deficit in the first 12 hours, plus minimum and maximum FiO_2 (fraction of inspired oxygen) in the first 12 hours (*6*). SNAP-II and SNAPPE-II were both derived from SNAP. SNAP uses 34 factors identified by experts as important in the first 24 hours of life (specific laboratory data, minimum and maximum vital sign values, and other clinical signs). The resulting score correlated well with birth weight, mortality, length of stay, nursing acuity, and physician estimates of mortality, but was complex to calculate (*4*). Logistic regression performed on the 34 factors in SNAP identified six variables most predictive of mortality that were recorded in the first 12 hours of life (lowest mean blood pressure, lowest core body temperature, lowest serum pH, multiple seizures, urine output, and FiO_2/PaO_2 ratio); these were retained in SNAP-II. SNAPPE-II is calculated with the same data as SNAP-II, along with the 5-min Apgar score, small for gestational age status, and birth weight. The additional variables present in SNAPPE-II were found to be independent risk factors for mortality (*5*). None of these scores, however, discriminate morbidity risk as well as PhysiScore, which integrates a small set of continuous physiological measures calculated directly from standard vital sign monitors.

An intriguing aspect of our findings is that PhysiScore provides high-accuracy predictions about morbidity risk from limited initial data (only 3 hours), even when such outcomes manifest days or weeks later (for example, BPD or NEC). PhysiScore gives positive weight to loss of short-term heart rate variability, much in the way that fetal heart rate monitoring uses loss of short-term heart rate variability to predict fetal distress and guide delivery management (*13*). PhysiScore additionally identifies short-term respiratory variability as having high predictive value, suggesting that further exploration of this factor in other settings might be warranted. Although the precise source of variability loss—either pre- or postnatally—is unknown, autonomic dysregulation likely plays a role. Whether short-term variability loss causes morbidity or is simply a marker of illness is not clear at this point.

Unlike fetal heart rate monitoring or heart rate spectral analysis (*14*) in the neonate, our approach uses multiple physiological parameters to improve accuracy and provide long-term predictions that extend beyond acute risk. Unlike biomarkers, such predictions are made with data that are already being collected in NICUs. Patient oxygenation, heart rate, and respiratory rate can be automatically processed to compute a score, and a predetermined sensitivity/specificity threshold can be used to make morbidity predictions to guide clinical actions,

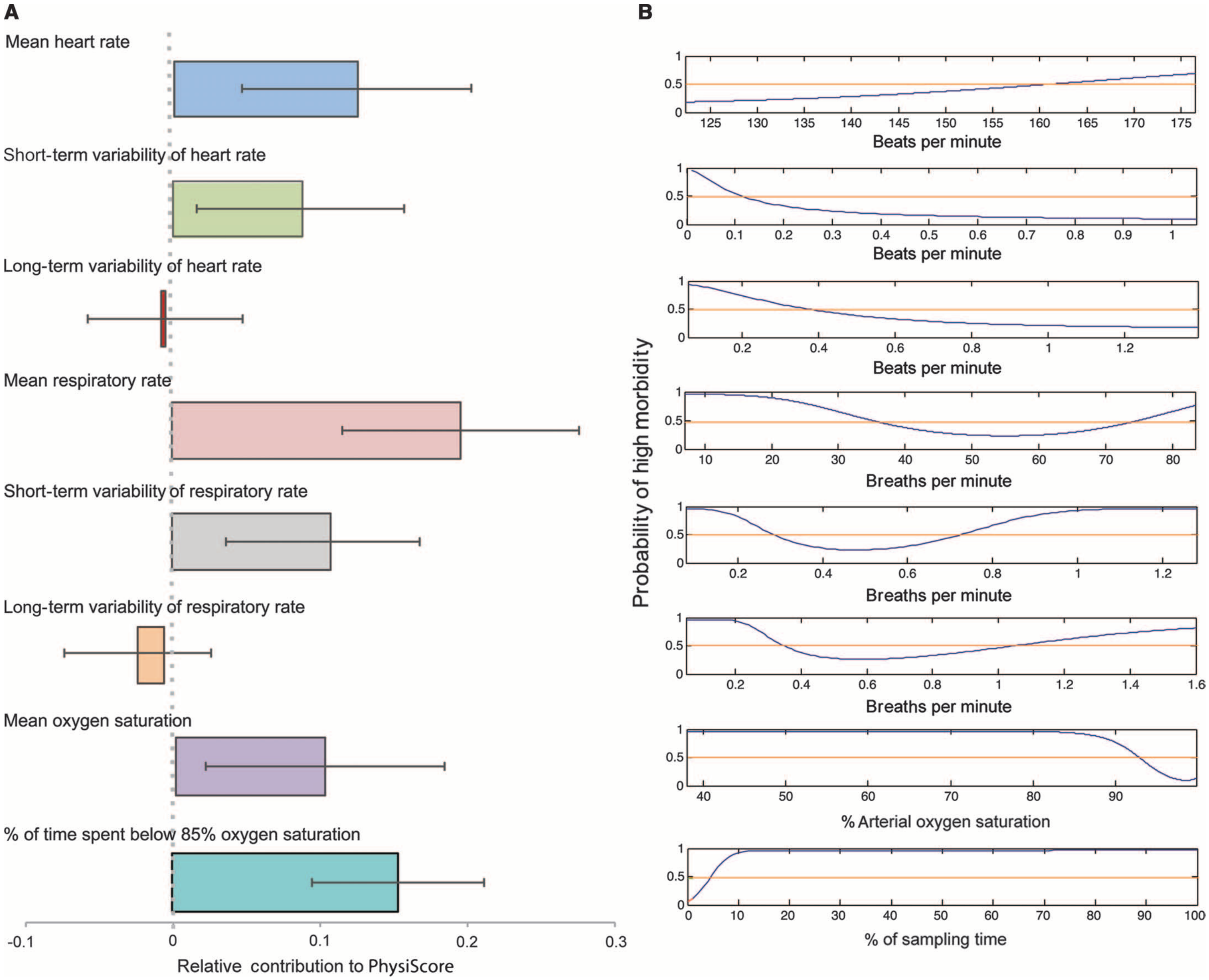

Fig. 3. The significance of different physiological parameters in predicting high morbidity. (**A**) The learned weight (w_i in Eq. 1) for each physiological parameter incorporated in PhysiScore; error bars indicate variation in the weight over the different folds of the cross-validation. (**B**) The nonlinear function associating the parameter with the risk of high versus low morbidity.

thereby removing the need for end-user expertise. When integrated into a bedside monitor, the algorithm would indicate the statistical likelihood that an individual patient is at high risk of major morbidities, allowing real-time use of the PhysiScore calculation. This method of deployment would effectively provide an automated electronic Apgar score, with significantly higher predictive accuracy regarding neonatal morbidity.

The PhysiScore's ability to assess physiologic disturbances before it can be confounded by medical intervention makes it particularly descriptive of initial patient acuity; thus, it is particularly well suited as a tool for quality assessment between NICUs. Identification of a patient's future risk of developing HM complications may be particularly useful for decision-making in primary nurseries to make more informed decisions regarding aggressive use of intensive care, need for transport to higher levels of care, and resource allocation. Such economic, social, and medical advantages should be evaluated in a large-scale clinical trial.

Technical considerations

Although we have a relatively small sample size, analysis methods appropriate to small sample sizes (*15*) were used, and ROC curves were made only for morbidities seen in >10% of our population. Our model, with its automatic factor modeling and selection, requires essentially no parameter tuning, which greatly helps to prevent overfitting in small samples.

In addition, our sample is from a single tertiary care center and was limited to patients born in our institution to ensure that continuous physiological data were available for the first hours of life. Validation in other settings will be required.

Detection of IVH remains elusive in the field of neonatal medicine. Previous work reported that fractal analysis of the original newborn heartbeat may be an early indicator of IVH (*14*), but yielded no better sensitivity than PhysiScore. It is possible that the underlying pathophysiology of IVH is variable (*16*), particularly in infants in whom severe IVH is the only morbidity. Although IVH is usually associated with cardiopulmonary instability, recent literature suggests that there may be genetic predisposition to isolated IVH, potentially limiting the role of antecedent physiological signals before large hemorrhages (*17*). Thus, it is possible that the small number of infants with isolated IVH that were not identified as high risk by PhysiScore represents a distinct subpopulation.

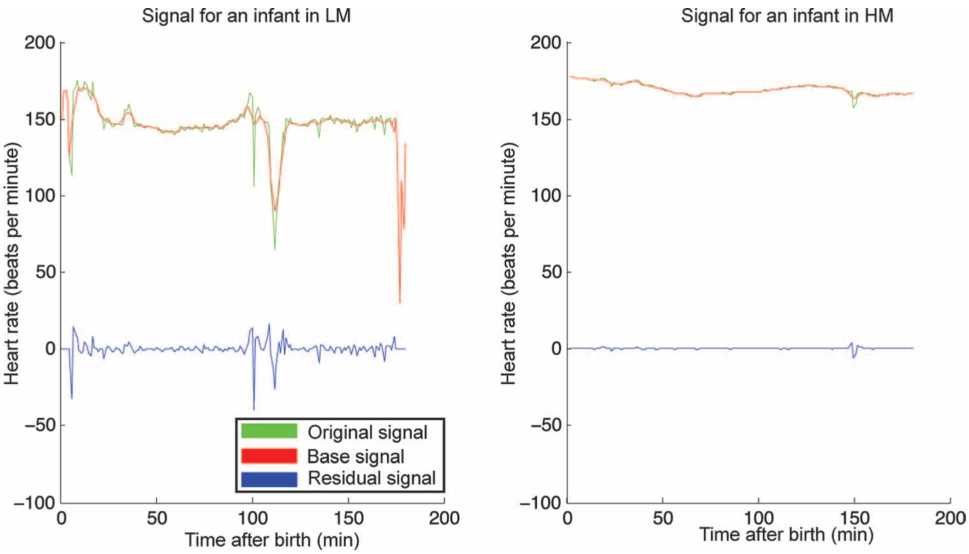

Fig. 4. Processing signal subcomponents. Differing heart rate variability in two neonates matched for gestational age (29 weeks) and weight (1.15 ± 0.5 kg). Original and base signals are used to compute the residual signal. Differences in variability can be appreciated between the neonate predicted to have HM (right) versus LM (left) by PhysiScore.

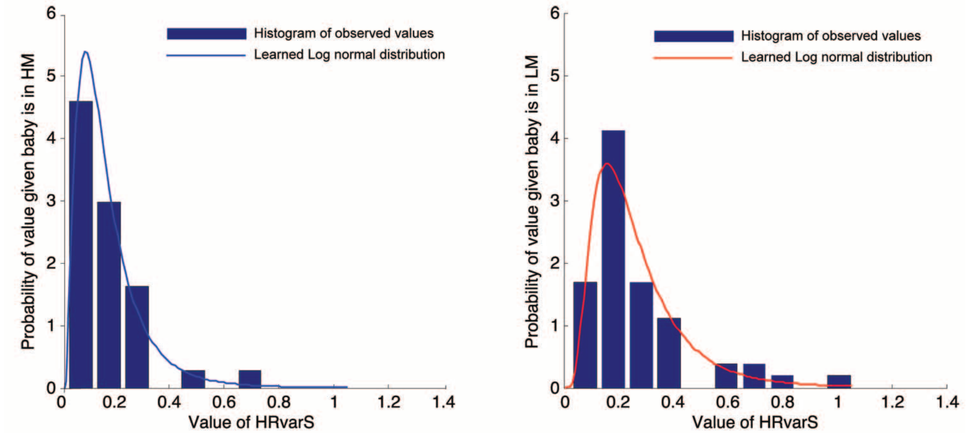

Fig. 5. Distribution of residual heart rate variability (HRvarS) in all infants. Learned parametric distributions overlaid on the data distributions for HRvarS displayed for the HM versus LM categorization.

Advanced computational techniques in modern medical settings

The use of computer-based techniques to integrate and interpret patterns in patient data to automate morbidity prediction has the potential to improve medical care. The current U.S. governmental mandate to improve electronic health record use and gain economic benefit from using digital data (*18*) facilitates the use of computer-based tools. Flexible Bayesian modeling with almost no tunable parameters allows our approach to be easily applied to a range of different prediction tasks, allowing use of the highly informative but underused data obtained daily for thousands of acutely ill patients.

MATERIALS AND METHODS

Ethics statement

All work was performed under protocol 8312 approved by Stanford's Panel on

Human Subjects. Waiver of Individual Authorization was approved under 45 CFR 164.512(i)(2)(ii)(A),(B),(C) on the basis that the data collection was part of routine care, no intervention or interaction with the patients occurred, and the data were processed anonymously.

General study strategy

After enrollment, we used a subset of patients ($n = 12$) to develop physiologic data processing methods. We combined state-of-the-art techniques from machine learning to build our framework that (i) processed these physiological parameters using nonlinear models, (ii) used regularization to do automatic feature selection, and (iii) combined relevant weights using multivariate logistic regression to produce the predictive PhysiScore (physiological features plus birth weight and gestational age). This framework has essentially no tunable parameters. Thus, unlike traditional frameworks that require separate feature selection and modeling steps followed by model testing using data, our framework combined these steps to allow direct testing of the predictive ability of this score on all 138 subjects by the leave-one-out method (*15*) to prospectively identify infants at high risk of severe complications.

Study population

Inborn infants admitted to the NICU of Lucile Packard Children's Hospital from March 2008 to March 2009 were eligible for enrollment. A total of 145 preterm infants met inclusion criteria: gestational age ≤34 completed weeks, birth weight ≤2000 g, and availability of cardiorespiratory (CR) monitor data within the first 3 hours of birth. Seven infants found to have major malformations were subsequently excluded.

Thirty-five neonates had HM complications. Of these, 32 had long-term morbidities (moderate or severe BPD, ROP stage 2 or greater, grade 3 or 4 IVH, and/or NEC). Four neonates died after the first 24 hours of life. There were 103 preterm neonates with only common problems of prematurity (RDS and/or PDA) and so were considered LM. Five infants with a <2-day history of mechanical ventilation for RDS, but no other early complications, were transferred before ROP evaluation and marked as LM.

Outcome annotation

Electronic medical records, imaging studies, and laboratory values were reviewed by pediatric nurses and verified by a physician. All significant illnesses during the hospitalization were recorded. Morbidities were identified with previously described criteria: BPD (*19*), ROP (*20*), NEC (*21*), and IVH (*22*). For IVH and ROP, the highest unilateral grade or stage was recorded, respectively. Acute hemodynamic instability was also noted: hypotension (defined as a mean arterial blood pressure less than gestational age or poor perfusion) requiring ≥3 days of pressor support or adrenal insufficiency requiring hydrocortisone.

Physiologic signal processing

Time series heart rate, respiratory rate, and oxygen saturation data are collected from all CR monitors. Heart and respiratory rate signals are processed to compute a base and residual signal. The base signal represents a smoothed, long-term trend; it is computed with a moving average window of 10 min. The residual signal is obtained by taking the difference between the original signal and the base signal; it characterizes short-term variability most likely linked to sympathetic function (Fig. 4). The variance features were motivated by analysis using the model in (*23*) on our preliminary set of 12 patients. For heart and respiratory rates, we compute the base signal mean, base signal variance, and residual signal variance. For the oxygen saturation, we compute the mean and the ratio of the time the oxygen saturation is below 85%.

Statistical methods

Sensitivity, specificity, AUC, and significance values (*12*) were computed for all comparisons. All statistical analyses were performed with software developed for this project (available for academic use upon request.) We used the leave-one-out method for all evaluations. With this method, predictive accuracy was evaluated for each patient separately. For each patient, we learned the model parameters with the data from all other patients as the training set and evaluated predictive accuracy on the held-out patient. This technique was repeated for each subject, so that each subject's clinical data were prospectively obtained. This method of performance evaluation is computationally intensive but is a well-established statistical method for measuring performance when the sample set size is limited (*15*).

Nonlinear models of risk factors

To implement Eq. 1, we must determine how to integrate continuous-valued risk factors, including the physiological measurements, into our risk model. Several approaches exist in the literature. One common approach is to define a "normal" range for a measurement and use a binary indicator whenever the measurement is outside that range. Although this approach can most easily be implemented in a clinical setting, it provides only coarse-grained distinctions derived from extreme values. Another approach is to predetermine a particular representation of the continuous-valued measurement, usually either the feature itself, or a quadratic or logarithmic transformation, as selected by an expert (*24*, *25*).

We used a different approach based on a Bayesian modeling paradigm (*26*). This approach captures the nonlinear relationships between the risk factor and the outcome and takes into account that the overall behavior of a factor can vary greatly between sickness categories. For each risk factor v_i, we separately learned a parametric model of the distribution of observed values in the training set $P(v_i|C)$ for each class of patient C (HM and LM). The parametric model is selected and learned with maximum-likelihood estimation (Fig. 5) from the set of long-tailed probability distributions of exponential, Weibull, lognormal, normal, and gamma. Specifically, for each parametric class, we fit the maximum likelihood parameters and then select the parametric class that provides the best (highest likelihood) fit to the data. The log odds ratio of the risk imposed by each factor was incorporated into the model.

An important advantage of our approach is that explicit missing data assumptions can be incorporated. When standard laboratory results (for example, complete blood count) are not recorded, we assume that they are missing at random and not correlated with outcome. Their contribution if missing is 0 and $\log P(v_i|\mathrm{HM})/P(v_i|\mathrm{LM})$ otherwise. Blood gas measurements, however, are likely obtained only for profoundly ill patients and hence are not missing at random. Thus, for each measurement type i, we define $m_i = 1$ if measurement v_i is missing and $m_i = 0$ otherwise. We now learn the distribution $P(m_i|C)$ (the chance that the measurement i is missing for each patient category C) and $P(v_i|C, m_i = 0)$ (the distribution of the observed measurements) as described above. The factor contribution for measurement i is computed as

$$f(v_i) = \begin{cases} \log P(v_i|\mathrm{HM}, m_i = 0)/P(v_i|\mathrm{LM}, m_i = 0) + \\ \quad \log P(m_i = 0|\mathrm{HM})/P(m_i = 0|\mathrm{LM}) & m_i = 0 \\ \\ \log P(m_i = 1|\mathrm{HM})/(m_i = 1|\mathrm{LM}) & m_i = 1 \end{cases} \quad (2)$$

This formulation allows us to account both for the observed measurement, if present, and for the likelihood that a particular measurement might be taken for patients in different categories.

This approach has additional advantages. Putting all factors in a probabilistic framework provides a comparable representation for different risk factors, allowing them to be placed within a single, integrated model. Using a parametric representation of each continuous measurement alleviates issues arising from data scarcity. Uncovering the dependence between the risk factor and the illness category automatically reduces data requirement by eliminating the need for cross-validation to select the appropriate form. Unlike most previous methods, we used different parametric representations for patients in different categories, better capturing disease-induced changes in patient physiology. Finally, we obtained an interpretable visual summary of the likelihood of low patient morbidity over the range of values for each factor (Fig. 3B).

Learning the PhysiScore parameters

To learn the score parameters b and w, we maximized the log likelihood of the data in the training set with a ridge penalty as

$$\arg\max_{w,b} \sum_{j=1}^{n} \log P(H|v_1^j,v_2^j \ldots v_{18}^j) - \lambda \sum_i w_i^2 \tag{3}$$

The ridge penalty reduces spurious data dependence by enabling automatic factor selection to control model parsimony and prevents overfitting (*27*, *28*). The hyperparameter λ controls the complexity of the selected model and was set to 1.2 in our experiments. This value was selected early in our development by random 70/30 cross-validation splits, based on experimental analysis showing that the results were not sensitive to the choice of this parameter.

REFERENCES AND NOTES

1. R. Behrman, A. Butler, Eds., *Preterm Birth: Causes, Consequences and Prevention* (National Academies Press, Washington, DC, 2007).
2. P. A. Robertson, S. H. Sniderman, R. K. Laros Jr., R. Cowan, D. Heilbron, R. L. Goldenberg, J. D. Iams, R. K. Creasy, Neonatal morbidity according to gestational age and birth weight from five tertiary care centers in the United States, 1983 through 1986. *Am. J. Obstet. Gynecol.* **166**, 1629–1641 (1992).
3. J. E. Tyson, N. A. Parikh, J. Langer, C. Green, R. D. Higgins; National Institute of Child Health and Human Development Neonatal Research Network, Intensive care for extreme prematurity—moving beyond gestational age. *N. Engl. J. Med.* **358**, 1672–1681 (2008).
4. D. K. Richardson, J. E. Gray, M. C. McCormick, K. Workman, D. A. Goldmann, Score for Neonatal Acute Physiology: A physiologic severity index for neonatal intensive care. *Pediatrics* **91**, 617–623 (1993).
5. D. K. Richardson, J. D. Corcoran, G. J. Escobar, S. K. Lee, SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *J. Pediatr.* **138**, 92–100 (2001).
6. The International Neonatal Network, The CRIB (Clinical Risk Index for Babies) score: A tool for assessing initial risk and comparing performance of neonatal intensive care units. *Lancet* **342**, 193–198 (1993).
7. V. Schulte-Frohlinde, Y. Ashkenazy, A. L. Goldberger, P. Ivanov, M. Costa, A. Morley-Davies, H. E. Stanley, L. Glass, Complex patterns of abnormal heartbeats. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **66**, 031901 (2002).
8. H. Tsuji, F. J. Venditti Jr., E. S. Manders, J. C. Evans, M. G. Larson, C. L. Feldman, D. Levy, Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. *Circulation* **90**, 878–883 (1994).
9. M. P. Griffin, D. E. Lake, J. R. Moorman, Heart rate characteristics and laboratory tests in neonatal sepsis. *Pediatrics* **115**, 937–941 (2005).
10. K. P. Williams, F. Galerneau, Intrapartum fetal heart rate patterns in the prediction of neonatal acidemia. *Am. J. Obstet. Gynecol.* **188**, 820–823 (2003).
11. B. M. Casey, D. D. McIntire, K. J. Leveno, The continuing value of the Apgar score for the assessment of newborn infants. *N. Engl. J. Med.* **344**, 467–471 (2001).
12. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).
13. K. P. Williams, F. Galerneau, Intrapartum influences on cesarean delivery in multiple gestation. *Acta Obstet. Gynecol. Scand.* **82**, 241–245 (2003).
14. V. Tuzcu, S. Nas, U. Ulusar, A. Ugur, J. R. Kaiser, Altered heart rhythm dynamics in very low birth weight infants with impending intraventricular hemorrhage. *Pediatrics* **123**, 810–815 (2009).
15. R. M. Rangayyan, *Biomedical Image Analysis. Biomedical Engineering Series* (CRC Press, Boca Raton, FL, 2005).
16. H. J. McCrea, L. R. Ment, The diagnosis, management, and postnatal prevention of intraventricular hemorrhage in the preterm neonate. *Clin. Perinatol.* **35**, 777–792, (2008).
17. P. Vannemreddy, C. Notarianni, K. Yanamandra, D. Napper, J. Bocchini, Is an endothelial nitric oxide synthase gene mutation a risk factor in the origin of intraventricular hemorrhage? *Neurosurg. Focus* **28**, E11 (2010).
18. *The American Recovery and Reinvestment Act of 2009* (Public Law 111-5) official text (Government Institutes/Bernan Press, Lanham, MD, 2009).
19. R. A. Ehrenkranz, M. C. Walsh, B. R. Vohr, A. H. Jobe, L. L. Wright, A. A. Fanaroff, L. A. Wrage, K. Poole; National Institutes of Child Health and Human Development Neonatal Research Network, Validation of the National Institutes of Health consensus definition of bronchopulmonary dysplasia. *Pediatrics* **116**, 1353–1360 (2005).
20. International Committee for the Classification of Retinopathy of Prematurity, The International Classification of Retinopathy of Prematurity revisited. *Arch. Ophthalmol.* **123**, 991–999 (2005).
21. R. M. Kliegman, M. C. Walsh, Neonatal necrotizing enterocolitis: Pathogenesis, classification, and spectrum of illness. *Curr. Probl. Pediatr.* **17**, 213–288 (1987).
22. L. A. Papile, J. Burstein, R. Burstein, H. Koffler, Incidence and evolution of subependymal and intraventricular hemorrhage: A study of infants with birth weights less than 1,500 gm. *J. Pediatr.* **92**, 529–534 (1978).
23. S. Saria, D. Koller, A. Penn, Discovering shared and individual latent structure in multiple time series. arXiv:1008.2028v1 [stat.ML] (2010).
24. G. Whitlock, S. Lewington, P. Sherliker, R. Clarke, J. Emberson, J. Halsey, N. Qizilbash, R. Collins, R. Peto, Body-mass index and cause-specific mortality in 900 000 adults: Collaborative analyses of 57 prospective studies. *Lancet* **373**, 1083–1096 (2009).
25. R. B. Schnabel, L. M. Sullivan, D. Levy, M. J. Pencina, J. M. Massaro, R. B. D'Agostino Sr., C. Newton-Cheh, J. F. Yamamoto, J. W. Magnani, T. M. Tadros, W. B. Kannel, T. J. Wang, P. T. Ellinor, P. A. Wolf, R. S. Vasan, E. J. Benjamin, Development of a risk score for atrial fibrillation (Framingham Heart Study): A community-based cohort study. *Lancet* **373**, 739–745 (2009).
26. S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists* (Elsevier Academic Press, Amsterdam, ed. 3, 2004).
27. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001).
28. J. Zhu, T. Hastie, Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**, 427–443 (2004).

29. **Acknowledgments:** We thank E. Helfenbein (Phillips Healthcare) for help in setting up a monitor data capture system; N. Llerena for data extraction; A. Pasca for figure production; D. Vickrey for insightful discussions; J. Hall, B. Kogut, P. Hartsell, and A. White for data annotation; and P. Prasad and Clinical Technology and Biomedical Engineering at Stanford for technical support. **Funding:** S.S. is a Rambus Fellow in Computer Science. A.K.R. is a Susan and Lynn Orr Fellow in Neonatology. This work was supported by the Stanford University Department of Computer Science pilot project funds. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** S.S., D.K., and A.A.P. conceived the study. All authors contributed to study design. S.S. and D.K. designed the methods and performed the analysis. A.K.R. extracted the medical annotations from the medical records. S.S. implemented the model and generated the figures. All authors contributed to the preparation of the manuscript and approved the final version. **Competing interests:** The authors declare that they have no competing interests.

Submitted 21 May 2010
Accepted 19 August 2010
Published 8 September 2010
10.1126/scitranslmed.3001304

Citation: S. Saria, A. K. Rajani, J. Gould, D. Koller, A. A. Penn, Integration of early physiological responses predicts later illness severity in preterm infants. *Sci. Transl. Med.* **2**, 48ra65 (2010).