Predicting and Understanding Bronchopulmonary Dysplasia in Premature Infants

Laney Kuenzel, under the mentorship of Suchi Saria and Professor Daphne Koller

CS 229 Final Report
December 10, 2010

I. Introduction

Bronchopulmonary dysplasia (BPD) is a lung disorder that affects infants, primarily those born prematurely. Defined as the requirement for oxygen therapy for at least 28 postnatal days, BPD occurs in nearly a third of infants with birth weight under 1000 grams [1]. BPD is associated with far-reaching negative consequences such as further respiratory problems, cerebral palsy, and cognitive impairment [2].

Unfortunately, BPD is one of the most poorly understood complications of prematurity. In particular, there is no consensus on the pathogenesis of the disease. Among the commonly hypothesized causes of BPD are ventilator-induced injury, lung immaturity, lung inflammation due to infection, and genetic predisposition [2, 3].

Previous work on predicting BPD has focused primarily on correlating eventual BPD diagnosis with laboratory measurements, medicine administrations, and mechanical ventilator settings. Such studies have consistently shown BPD to be significantly associated with certain abnormal blood gas levels (e.g., low blood pH) as well as aggressive ventilation [4–6].

Despite the abundance of studies seeking to identify factors associated with BPD diagnosis, a large gap still exists in the body of literature on BPD prediction: very little is known about what characterizes the physiological signals (such as heart rate and respiratory rate) of infants eventually diagnosed with BPD. We believe that this area is under-explored largely because it has only recently become possible to obtain fine-grained physiological time series data for hospitalized infants. We were fortunate enough to have access to this type of data, enabling us to conduct novel research on the relationship between physiological signals and BPD. This endeavor was particularly exciting due to its potential to generate important new medical knowledge.

In our search for physiological signatures for BPD, we focused specifically on three signals: heart rate, respiratory rate, and oxygen saturation. We had two main reasons for choosing these particular signals.

First, it makes sense from a biological perspective that lung problems would manifest in these three signals. The lungs serve to introduce oxygen from inhaled air into the bloodstream and to release carbon dioxide from the blood as exhaled air. Therefore, we would expect infants with poor lung function to exhibit low oxygen levels (motivating our use of the oxygen saturation signal) and high carbon dioxide levels. Furthermore, in infants with lung problems, we would expect to observe altered patterns in the breathing rate and heart rate (motivating our use of the respiratory rate and heart rate signals) as the body attempts to respond to the blood gas imbalance by adjusting the amount of air entering and exiting as well as the speed with which the blood is circulating.

Second, these three signals are recorded noninvasively for every infant in a neonatal intensive care unit (NICU). Consequently, a predictive model for BPD based only on features of these three signals could easily be adopted by any NICU. This is not the case for a model with features that rely on more invasive, expensive, or nonstandard measurements.

For these reasons, we formulated our overall goal as understanding whether and how BPD manifests in an infant's heart rate, respiratory rate, and oxygen saturation signals. The remainder of this paper describes our efforts toward this goal.

II. Data

We had access to minute-interval time series data collected from monitoring devices attached to premature infants during their entire stay in Stanford Hospital's NICU. For these infants, we also had data on all of the clinical events, such as laboratory tests and ventilator setting changes, that occurred during their hospitalization.

Of the infants admitted to Stanford's NICU between March 2008 and March 2009, we considered those satisfying the following criteria: gestational age ≤ 34 weeks, birth weight ≤ 2000 grams, length of life ≥ 28 days (allowing for BPD diagnosis), and availability of ≥ 5000 minutes of monitor data. Of these infants, the thirty with a positive BPD diagnosis were included in the study. As negative examples, an additional thirty-seven infants were chosen at random from those diagnosed with respiratory distress syndrome, an indicator of breathing problems at the time of birth, but not BPD.

In this set of infants, low gestational age and low birth weight were found to be highly predictive of BPD, with areas under the ROC curve (AUCs) of 0.93 and 0.91, respectively. For this reason, we also created an age/weight-matched set of twenty infants (ten with BPD and ten controls), for which gestational age and birth weight had much lower AUCs of 0.62 and 0.59, respectively.

III. Initial Exploration of Features

For each infant, we had a very large quantity of data, including several long physiological time series and information on tens of thousands of diverse clinical events. As a first step, we wanted to determine which of the many available pieces of data would be most useful to us for predicting BPD. We drew from the BPD literature, our discussions with Stanford NICU clinicians, and our observations from visualizing the data to identify 160 potentially interesting features of the first 5000 minutes of data available for each infant.

Around forty of these features were functions of the physiological signals, such as mean, range, standard deviation, and amount of time below or above threshold values. Sixty more features were related to laboratory measurements (e.g., platelet count and blood oxygen level) identified as predictive

TABLE I
CONSERVATIVE MEAN AUCS FOR TOP FEATURES

Feature	Full Set	Match Set
Number of dextrose administrations	0.84	0.82
Number of ISTAT blood gas measurements	0.84	0.78
Number of ventilator setting increases	0.82	0.82
Number of ventilator FiO_2 setting increases	0.80	0.79
Number of ventilator rate setting increases	0.77	0.77
Number of 10% dextrose administrations	0.75	0.78
Maximum airway resistance measurement	0.84	0.74
Mean blood pH measurement	0.82	0.74
Minimum blood pH measurement	0.80	0.84
Maximum blood carbon dioxide measurement	0.79	0.84
Range of blood carbon dioxide measurements	0.77	0.78
Range of blood pH measurements	0.74	0.87
Minimum blood oxygen measurement	0.74	0.77
Median RR range over 5-minute windows	0.77	0.77
Mean abs. diff. between consecutive RR values	0.77	0.74

of BPD in previous studies. For each type of measurement, we included as features the first, mean, minimum, and maximum values, the range of values, and the number of measurements taken. Given that existing work has shown aggressive ventilation to be correlated with BPD, we also included about thirty features related to ventilator type and settings. Rounding out the set of features we considered were the numbers of administrations of over twenty types of medicine.

To determine the predictive value of each feature, we computed the conservative mean AUC (with $k = 10$ folds) as proposed by Khosla *et al.* in order to penalize features sensitive to variations in sampling [7]. The fifteen top-performing features across both the full set and the age/weight-matched set are displayed in Table I along with their conservative mean AUCs. We observed that these top features fell into three categories:

(i) Six of the features measured the frequency of various clinical interventions (namely medicine administrations, blood draws, and ventilator adjustments) ordered by the NICU doctors. The strong performance of these intervention frequencies as features suggests that Stanford clinicians have a good sense of which infants are at highest risk for BPD or other complications and therefore require the most treatment and surveillance.

(ii) Seven of the features were functions of measurements other than the physiological signals. More specifically, six were related to blood gas values and the other one was a function of the airway resistance measurement, which is taken only for infants on a particular type of ventilator. Interestingly, the frequency of blood draws was generally more predictive of BPD than the actual measurement values resulting from those draws. Assuming that the NICU doctors order more blood draws for the infants that they deem sicker, this result suggests that the doctors' assessment of an infant's health status is richer in predictive information than the blood gas levels for that infant.

(iii) The remaining two features were functions of the respiratory rate. We were surprised to find that of the over forty physiological features we considered, only these two were among the fifteen most informative.

As described in Section I, our aim in this project was to shed light on the relationship between physiological signals and BPD. We found in this initial exploration that most of the highly predictive features for BPD were related not directly to physiology but instead to intervention frequencies and laboratory measurements. Based on this result, we decided to focus our further efforts on finding physiological proxies for the informative interventions and measurements.

IV. MOTIVATION FOR "PROXY" APPROACH

Essentially, we decided to examine the relationship between our three physiological signals and the interventions and measurements corresponding to the features in categories (i) and (ii) above. More specifically, we hoped to identify physiological signatures characterizing the time when a given intervention is ordered, in the case of category (i), or the time when an infant has a certain measurement value, in the case of category (ii). Our rationale was that because the intervention frequencies and measurement values were so predictive of BPD, good physiological proxies for them would likely be predictive of BPD as well.

In addition to their potential value in BPD prediction, physiological proxies for the informative features would be useful in several other ways. In the case of category (i), we could use physiological signatures for a given intervention to create a tool reporting whether an infant is exhibiting physiology which typically precedes that intervention. Such a tool would help NICU doctors decide whether and when a certain intervention is necessary.

Physiological proxies for the blood gas and resistance measurements in category (ii) would be valuable for two reasons beyond BPD prediction. First, we could incorporate such proxies into a tool that would let doctors noninvasively obtain a rough estimate of blood gas levels or resistance for an infant. This type of tool would reduce the number of invasive blood draws performed, benefiting both the health of the infants and the hospital's budget. Second, independent of any prediction application, an understanding of how abnormal measurements manifest in the physiological signals of premature infants would constitute valuable medical knowledge.

Why did we believe that physiological proxies for our predictive features would exist at all? In terms of interventions, Stanford NICU clinicians informed us that they decide what interventions to order based partly on an infant's physiological signals. For example, doctors consider low oxygen saturation and high respiratory rate as signs that the ventilator is not working effectively and requires setting increases. In terms of measurements, we expected that abnormal blood gas values would be reflected in the physiological signals since the body sets the heart and respiratory rates based on blood levels of oxygen and carbon dioxide.

V. Experimental Setup

A. Data

In our first set of experiments, we considered five interventions with predictive frequencies: dextrose administrations, ISTAT blood gas measurements (note that "ISTAT" refers to a type of handheld blood gas meter), and increases in three types of ventilator settings. We examined the 60-minute intervals of our physiological signals preceding these interventions. The intervals were taken from the first 5000 minutes of monitor data available for each of the infants in our set. As negative examples, we wanted to find intervals during which the doctor considered the infant's state and could have ordered the intervention but did not. For this purpose, we chose 60-minute intervals preceding times when the doctor entered electronic comments about the infant but did not order the intervention in question during the hour before or after comment entry. This procedure resulted in sets of 1862, 966, 930, 806, and 146 intervals for the five interventions.

In our second set of experiments, we took the 60-minute intervals prior to measurements of blood pH, blood oxygen, blood carbon dioxide, and airway resistance that occurred during the first 5000 minutes of monitor data for the infants in our set. We had 827, 827, 771, and 189 intervals, respectively, corresponding to these four measurements.

B. Feature Extraction

For each interval, we computed 448 features based on the heart rate, respiratory rate, and oxygen saturation signals. We used five different approaches to obtaining features:

(i) We computed simple functions of the signals, such as mean, range, and variability.

(ii) We calculated time-lagged correlation between each pair of signals. We believed that correlation might be informative based on our visualization of the data. Moreover, research on other complications of prematurity shows that sick infants often have impaired autoregulation [8], leading to synchronization of physiological signals.

(iii) We used the discrete Fourier transform (DFT) to obtain features capturing the frequency contents of each signal.

(iv) We applied the Time Series Topic Model (TSTM) developed by Saria *et al.* [9]. The TSTM segments the physiological signal into regions ("words") generated by the same autoregressive process, indicating similar short-term dynamics. The TSTM also learns higher-level "topics" corresponding to different distributions over words. As features, we used both word and topic frequencies obtained from the TSTM.

(v) We learned a two-layer belief network using the sparse Restricted Boltzmann Machine (RBM) algorithm proposed by Lee *et al.* [10]. A single example consisted of the sixty values each of heart rate, respiratory rate, and oxygen saturation that occurred during one 60-minute interval. We pre-processed the data by applying PCA whitening and then learned a sparse RBM model with 400 hidden units. With the resulting hidden unit probabilities, we trained a second sparse RBM layer with 400 hidden units. As features for a given interval, we used the inferred values of the second-layer hidden units resulting from feeding the interval's signals forward through the trained model.

TABLE II
AUCs for PLR Classifiers to Predict Interventions

	Features		
Intervention Type	Non-RBM	RBM	All
Dextrose administration	**0.67**	0.65	0.64
ISTAT blood gas measurement	**0.55**	0.54	0.54
Increase in any ventilator setting	**0.77**	0.72	**0.77**
Increase in ventilator FiO$_2$ setting	**0.75**	0.70	**0.75**
Increase in ventilator rate setting	0.60	**0.65**	0.62

We normalized each feature to have mean zero and unit standard deviation so that the weights learned by classifiers would be meaningful in comparison to one another.

C. Predicting Interventions

For each of the five interventions, we trained a penalized logistic regression (PLR) classifier using the algorithm proposed by Zhu and Hastie [11]. On a training set consisting of 70% of the intervals, we performed five-fold cross-validation to select the optimal regularization parameter. More specifically, for each of a range of possible parameter values and for each fold, we computed the AUC of the ROC curve for the classifier's outputted probabilities. Using the parameter which maximized the conservative mean AUC over the folds, we trained a PLR classifier on the entire training set and then used this classifier to make predictions on our test set.

We repeated this procedure three times: once with all 448 of our features, once with only the 400 RBM features (i.e., those of type (v) in Section V-B), and once with only the 48 non-RBM features. We separated the features in this way because we were especially interested to compare the belief networks with our other feature extraction methods in terms of how effectively they captured the information in the physiological signals. The resulting AUCs are reported in Table II and will be discussed in Section VI.

D. Predicting Measurement Values

We first tried linear regression to predict measurement values but found that the amount of error was unacceptably high for each measurement type. We instead decided to attempt the simpler task of predicting whether the values fell above or below a given threshold. We observed that NICU doctors often mentally represent an infant's blood gas values with ternary ("high", "normal", or "low") or binary ("abnormal" or "normal") values. Thus, we reasoned that a model making binary predictions of measurement value would be useful to NICU doctors.

As thresholds separating "low" from "high" for blood pH, oxygen, and carbon dioxide, we used values indicated as clinically meaningful by Stanford physicians. For airway resistance, since such a value was not available, we took the ROC curve generated for the "maximum resistance" feature

TABLE III
AUCS FOR PLR CLASSIFIERS TO PREDICT MEASUREMENTS

	Features		
Measurement Type	Non-RBM	RBM	All
Blood pH	0.86	0.76	**0.87**
Blood oxygen	**0.74**	0.58	0.68
Blood carbon dioxide	0.69	0.68	**0.72**
Airway resistance	**0.72**	0.56	0.65

from Section III and selected the threshold with the best sensitivity and specificity (i.e., the one leading to the point on the ROC curve closest to the upper left corner (0,1)).

We used the same procedure described in Section V-C to train PLR classifiers. Table III shows the resulting AUC values.

VI. RESULTS AND DISCUSSION

We found that in each case but one, the non-RBM classifier outperformed its RBM counterpart. Furthermore, when we learned classifiers using the non-RBM and RBM features together, we obtained AUCs which were not substantially higher than those for the non-RBM classifiers. For the remainder of this discussion, then, we focus on the non-RBM classifiers.

A. Predicting Interventions

We found that we were unable to accurately predict whether a given interval preceded a dextrose administration, blood gas measurement, or ventilator rate increase. To better understand the problems with our models, we ran the classifiers for these three interventions on their respective training sets to compute training AUCs.

For dextrose administrations and blood gas measurements, the training AUCs were low (0.64 and 0.68, respectively), indicating that the classifiers failed to separate even the training examples well. We believe that the most likely explanation for this poor performance is that NICU doctors rely primarily on factors other than the physiological signals in deciding when to order these two interventions. Indeed, one Stanford clinician informed us that he mainly orders ISTAT blood draws after adjusting the ventilator, and thus his choice of whether to order the intervention at a given time does not depend strongly on the infant's physiology. In terms of dextrose administrations, we observed that they generally occurred at regularly spaced intervals which varied in length for different infants, suggesting that doctors set a dextrose administration schedule early on rather than deciding based on the past hour's physiological signals. Given that the frequency of dextrose administrations was predictive of BPD, it would be fruitful to investigate how NICU doctors set this dextrose schedule.

On the other hand, the ventilator rate increase classifier had a training AUC (0.82) much higher than its testing AUC. Thus, it appears that the classifier was overfitting, especially given the relatively small number of examples (146) for this intervention.

We were fairly successful in predicting whether an interval preceded a ventilator setting increase or a ventilator FiO_2 (fraction of inspired oxygen) increase. We note that the two classification tasks were similar, since the majority of setting increases were FiO_2 increases. One of our main goals in creating these classifiers was to identify physiological signatures useful for BPD prediction. To do so, we examined the weights that the two classifiers assigned to the features. (Recall that we had normalized feature values so that the weights would be comparable.) We found that all of the highly weighted features for both classifiers were functions of the oxygen saturation (OS) signal. This result was not surprising given that NICU clinicians told us that they often adjust the ventilator in response to desaturation events (i.e., periods of low OS). Interestingly, the top features for both classifiers included not only simple functions of the OS signal like mean and variability but also several features obtained from the signal's discrete Fourier transform (DFT). This observation suggests that the frequency content of the OS signal captures useful information about desaturation events and therefore has potential for BPD prediction.

B. Predicting Measurement Values

The classifier for blood pH performed very well. The other three classifiers were also reasonably successful in differentiating high values from low. As discussed in Section IV, these classifiers could be extremely useful in a NICU, giving doctors noninvasive real-time estimates of an infant's blood gas values.

To gain insight into BPD prediction, we examined the feature weights that our classifiers learned. One general observation was that of our nine total classifiers, eight assigned very high weight to the mean OS feature. This finding raises the question of why the mean OS feature was not among the most predictive of BPD in our initial feature exploration. We hypothesized that the answer was our segmentation of the signal into intervals and that, more specifically, the distribution of mean OS values over intervals captures more useful information than the signal's overall mean OS. Indeed, we found that the minimum of mean OS values over 60-minute intervals was more predictive of BPD than either the mean or the minimum OS value over the first 5000 minutes. This result demonstrates the benefit of focusing on short intervals of the physiological signals.

Besides mean OS, the measurement classifiers learned high weights for other simple OS features such as variability. Additionally, DFT-based features of the signals appeared frequently among the top features for all four classifiers, again suggesting that the frequency contents of the physiological signals have promise for predicting BPD.

We made two observations which were particularly interesting from a medical perspective. First, we found that the most highly weighted feature for the resistance classifier was a function of the heart rate (HR) signal's DFT. Based on the sign of the weight, periods of high airway resistance are characterized by HR signals with more high-frequency components. This result sheds light on the effect of lung dynamics on heart rate and, more broadly, the interrelation between the respiratory and circulatory systems of premature infants. Our second interesting observation was that both the blood pH and blood oxygen classifiers learned high weight for the correlation between the HR and OS signals. Beyond suggesting a feature for BPD prediction, this link between high HR/OS alignment and low blood pH and blood oxygen levels provides insight into an unhealthy infant's ability (or lack thereof) to autoregulate his body systems.

VII. Creating a BPD Classifier

By examining the weights learned by our various classifiers, we identified sixteen features as potentially useful for BPD prediction: four simple OS features, ten DFT features, HR/OS correlation, and mean RR. With the addition of the two physiology-based features found to be predictive of BPD in our initial exploration, we had a set of eighteen features.

Rather than classifying infants as BPD or control, we decided to create a model that would classify 60-minute intervals as BPD or not. We had several reasons for making this choice. First, we saw an example in Section VI-B in which focusing on 60-minute intervals led to a more predictive feature for BPD than working with the entire 5000-minute signal at once. Second, our previous experiments were performed using 60-minute intervals, so some of our features (namely those based on the DFT) applied specifically to intervals of that length. Finally, we believed that an interval classifier would be more valuable to NICU doctors, as it would let them obtain predictions every hour and thereby gauge changes in an infant's BPD risk over time and with different treatments.

Ideally, we would train and test our BPD interval classifier on data taken from a new set of infants, since we already used data from our current set to create the intervention and measurement classifiers. Unfortunately, as obtaining data for new infants requires efforts by multiple people to label and pre-process the data into usable form, we were unable to create a new data set within the time frame of this project.

Instead, we learned a BPD classifier using the same data set from our previous experiments. We started by segmenting each infant's first 5000 minutes of data into 60-minute intervals. Then, with the eighteen features described above, we trained a PLR classifier on 500 randomly-chosen intervals, using the Bayesian information criterion to select the optimal value of the regularization parameter. We ran the resulting classifier on all of the remaining intervals. To obtain an overall BPD risk score for a given infant, we computed the proportion of that infant's intervals which were classified as BPD (excluding from the computation any intervals which were included in the training set). These scores achieved AUCs of 0.89 and 0.80 on the full set and age/weight-matched set, respectively.

VIII. Conclusion and Future Work

The final product of our project was a physiology-based BPD interval classifier which achieved strong results on our data set. We would like to train and test this classifier on a new and larger set of infants to verify whether accurate BPD prediction is indeed possible with our chosen features.

In addition to developing a BPD classifier, we created nine intervention and measurement classifiers, several of which performed well and could be very useful in a NICU. We also identified informative features related to signal frequency content and HR/OS correlation, a valuable finding not just for predicting BPD but also for understanding it biologically.

There are a number of directions in which we would like to extend this work. For one, given our observation that segmenting a signal into intervals can yield more predictive features, we would like to repeat our experiments using different interval lengths to determine which one is optimal. We are also interested in further exploring the relationship between cross-signal correlation and BPD by developing more sophisticated measures of correlation tailored to our specific application. In terms of learning algorithms, we would like to check whether SVMs can outperform our PLR classifiers.

Finally, we are interested in characterizing the differences between BPD and control infants in terms of response to interventions, particularly ventilator adjustments. Our hypothesis is that interventions cause more dramatic changes in the physiological signals of a BPD infant than a control, since BPD infants are less able to autoregulate effectively. In a preliminary experiment, we examined intervals extending from 30 minutes before to 30 minutes after an intervention. Using as features the differences in simple functions of the signals before and after the intervention, we attempted to classify the intervals as BPD or control. Our classifiers performed poorly, so we would like to try to improve them by identifying features that better capture response characteristics.

IX. Acknowledgments

I would like to thank Suchi Saria for her guidance throughout the course of the project, Dr. Anand Rajani and Dr. Valerie Chock for answering our questions about the NICU, and Jiquan Ngiam for letting us use and helping us understand his sparse RBM code.

References

[1] Walsh, M. *et al.* Summary proceedings from the bronchopulmonary dysplasia group. *Pediatrics* **117**(3), S52-56 (2006).

[2] Kinsela, J., Greenough, A. & Abman, S. Bronchopulmonary dysplasia. *Lancet* **367**(9520), 1421-31 (2006).

[3] Jobe, A. & Ikegami, M. Mechanisms initiating lung injury in the preterm. *Early Hum Dev* **53**(1), 81-94 (1998).

[4] Van Marter, L. *et al.* Do clinical markers of barotrauma and oxygen toxicity explain interhospital variation in rates of chronic lung disease? *Pediatrics* **105**(6), 1194-1201 (2000).

[5] Garland, J. *et al.* Hypocarbia before surfactant therapy appears to increase bronchopulmonary dysplasia risk in infants with respiratory distress syndrome. *Arch Pediatr Adolesc Med* **149**(6), 617-22 (1995).

[6] Yoder, B., Anwar, M. & Clark, R. Early prediction of neonatal chronic lung disease: a comparison of three scoring methods. *Pediatr Pulm* **27**(6), 388-94 (1999).

[7] Khosla, A. *et al.* An integrated machine learning approach to stroke prediction. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2010).

[8] Kaiser, J. The association of high-magnitude cerebral passivity and intraventricular hemorrhage in premature infants. *Pediatrics* **124**(1), 384-86 (2009).

[9] Saria, S. *et al.* Discovering shared and individual latent structure in multiple time series. Under review (2010).

[10] Lee, H., Ekanadham, C. & Ng, A. Sparse deep belief net model for visual area V2. *NIPS* (2007).

[11] Zhu, J. & Hastie, J. Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**(3), 427-43 (2004).