Published in final edited form as:
*Med Decis Making*. 2010 ; 30(2): 258–266. doi:10.1177/0272989X09337791.

# Nearest neighbor and logistic regression analyses of clinical and heart rate characteristics in the early diagnosis of neonatal sepsis

**Yuping Xiao, M.S.**[1], **M. Pamela Griffin, M.D.**[2], **Douglas E. Lake, Ph.D.**[1], and **J Randall Moorman, M.D.**[1,3]

[1]Department of Internal Medicine, University of Virginia Health System, Charlottesville, VA 22908

[2]Department of Pediatrics, University of Virginia Health System, Charlottesville, VA 22908

[3]Department of Cardiovascular Research Center, University of Virginia Health System, Charlottesville, VA 22908

## Abstract

**Objectives—**To test the hypothesis that nearest-neighbor analysis adds to logistic regression in the early diagnosis of late-onset neonatal sepsis.

**Design—**We tested methods to make the early diagnosis of neonatal sepsis using continuous physiological monitoring of heart rate characteristics, and intermittent measurements of laboratory values. First, we tested the hypothesis that nearest-neighbor analysis makes reasonable predictions about neonatal sepsis with performance comparable to an existing logistic regression model. We systematically developed the most parsimonious model by excluding the least efficacious clinical data. Second, we tested the hypothesis that a combined nearest-neighbor and logistic regression model gives an outcome prediction that is more plausible than either model alone. Training and test data sets of heart rate characteristics and laboratory test results over a 4-year period were used to create and test predictive models.

**Measurements—**Nearest-neighbor, regression and combination models were evaluated for discrimination using ROC areas and for fit using Wald statistic.

**Results—**Both nearest-neighbor and regression models using heart rate characteristics and available laboratory test results were significantly associated with imminent sepsis, and each kind of model added independent information to the other. The best predictive strategy employed both kinds of models.

**Conclusion—**We propose nearest-neighbor analysis in addition to regression in the early diagnosis of sub-acute, potentially catastrophic illnesses like neonatal sepsis, and we recommend it as an approach to the general problem of predicting a clinical event from a multivariable data set.

## INTRODUCTION

The early diagnosis of sepsis in the neonatal intensive care unit should be an excellent application for data mining approaches. First, we wished to tested the hypothesis that nearest-neighbor analyses make predictions on neonatal sepsis that are comparable to the existing

Corresponding author and reprints: Douglas E. Lake, Ph.D., Box 801395, UVAHS, Charlottesville, VA 22908, dlake@virginia.edu.

Presented in part at the Pediatric Academic Societies annual meeting, May, 2005.

**CONFLICT OF INTEREST STATEMENT**

Medical Predictive Science Corporation of Charlottesville, VA has a license to market technology related to heart rate characteristics monitoring of newborn infants, and supplied partial funding for this study. Drs. Griffin and Moorman have an equity share in this company.

regression models that relate HRC index and laboratory tests to neonatal sepsis. Second, we wished to test the hypothesis that the combined model of nearest-neighbor and regression, not necessarily using the same clinical data, is more accurate than either model by itself.

## BACKGROUND

Neonatal sepsis is a major cause of morbidity and mortality in premature infants hospitalized in neonatal intensive care units (NICUs) (1). It is difficult to diagnose in its earliest and most treatable stages because clinical signs and laboratory test abnormalities are subtle and non-specific (2;3), and thus it commonly presents in advanced stages as systemic inflammatory response syndrome (4;5). We regard neonatal sepsis as an example of many sub-acute illnesses with sub-clinical phases during which treatment should be highly effective in preventing potential catastrophe.

To aid in the early diagnosis of neonatal sepsis, we developed heart rate characteristics (HRC) monitoring (6–9) based on the observation that reduced variability and transient decelerations of heart rate occur in the hours to days prior to the clinical presentation (10). These abnormalities can be quantified using novel time-series measures (11–13) incorporated into a predictive model based on logistic regression. The resulting “HRC index” has highly significant association with sepsis in internal and external validation studies.

We know, though, that physicians rely primarily on experience and not regression equations to diagnose illness, and employ pattern recognition to distill complex presenting features into a short list of possible diseases. This invaluable exercise is not usually formalized, and most diagnostic test results arrive independently and without contextual interpretation. Thus the common clinical discourse of “whenever I see *x* and *y* I think of *z*” is not codified for universal use. Formal approaches to statistical pattern recognition have been developed in other fields (14), and “data mining” is a recently coined term for these strategies, which often exploit computer processing of large databases. Moreover, in general, to achieve optimal prediction of the outcome we need to know the probability distribution of all the variables. As this is impossible in many practical problems, decision rules that do not need knowledge of distributions are favored. Nearest-neighbor analysis, described below, is one such strategy. It takes advantage of a large number of known samples in the training set to compensate for the lack of distributions, and it is reliable for a mixture of continuous and discrete variables (15). Specific to clinical medicine is the idea that a single abnormal finding could supersede many normal findings in diagnosing or predicting illness, and all clinicians recall patients in whom dire diagnoses were made based on a single crucial abnormal finding in a sea of normal ones.

The current study was motivated by the recent finding that laboratory test results add independent information to HRC monitoring in diagnosis of neonatal sepsis (16). Our analysis has been based on regression, which recognizes only monotonic relationships between lab values and risk. This reasoning fails for tests like the WBC or body temperature, which might be abnormally high *or* low in illness. Thus for the new study, we have used a pattern-recognition technique called “nearest-neighbor” analysis (14). The principle is simple. For a patient with a set of findings, one finds the most similar infants in their experience and lists their diagnoses and outcomes. Nearest-neighbor analysis has been widely used in pattern recognition studies of many kinds, but has been relatively underused in clinical medicine. Haddad and co-workers used this approach to detect coronary artery disease using patterns of perfusion scintigraphy in 100 patients in whom the presence or absence of disease was established by angiography (17). Qu and Gotman developed a patient-specific seizure detection algorithm based on EEG waveforms in the presence and absence of seizures (18). Most recently, Lutz and co-workers were able to forecast effects of psychotherapy based on reference responses of 203 clients (19).

A particular strength of nearest-neighbor analysis is independence from assumptions about "normal" levels of test results, or about relationships among test results. The results arise entirely from experience, mimicking at least part of a physician's thought process. Since sepsis elicits a complex systemic inflammatory response syndrome with dysfunction of multiple organs, it seems sensible to consider as many simultaneous processes as possible. On the other hand, however, some laboratory tests are taken much less frequently than the others, making the database smaller the more processes we take into consideration at the same time. Last but not least, while many variables contribute to a model, some may be more closely related to the model outcome than the others, and before the roles of all variables are understood fully, including as many variables as possible in one model may turn out to be impractical and potentially problematic. Thus we need a model, or a combination of multiple models that can include many important variables and also deal with the real world problem of intermittent sampling of laboratory measures.

## RESEARCH QUESTION

Does nearest neighbor analysis add to existing logistic regression methods for early diagnosis of neonatal sepsis?

## METHODS

### Study population

We studied all admissions to the University of Virginia NICU from July 1999 to July 2003 that were 7 or more days of age. The clinical research protocol was approved by the Human Investigations Committee of the University of Virginia. Laboratory test results were available from an electronic archive. The analysis was limited to the time that HRC data were available, or 92% of the total time. Infants were followed prospectively to identify cases of sepsis, but health care personnel were not aware of the result of the HRC monitoring. We defined sepsis to be present when a physician suspected the diagnosis, obtained a blood culture that grew bacteria not ordinarily considered to be a contaminant, and initiated antibiotic therapy of 5 or more days duration. This definition is consistent with the diagnosis of "proven sepsis" used by the Center for Disease Control (20). Sepsis was diagnosed in the usual course of care, that is, no cultures were done for study purposes.

### Data sets

Both the nearest-neighbor and logistic regression analyses call for a training data set and a test data set. Each point represents a summary of the past 12 hours, and points were calculated every 6 hours. The HRC index was available at each point, and laboratory test results were intermittently available. Thus each data point in the test set could be evaluated for HRC values, and many could be evaluated for one or more individual laboratory test results when they were available. In our analysis, the duration of a test result was 12 hours. For this work, HRC and laboratory test data obtained from 1999–2003 in the University of Virginia NICU were split randomly and nearly evenly into a training set of patients and a test set of patients. A total of 676 patients (with more than 70,000 records) were included in this study, and among them 326 patients were in the training set. All were obtained after 7 days of age, and we neglected data during the 7 days after a positive blood culture. The sepsis event was defined to occur over a 24 hour period beginning 6 hours before the positive blood culture and ending 18 hours after. We justify this selection based on our prior work using regression modeling that shows this is the epoch in which most of the diagnostic test results are available (8;16). BW and days of age were included in all the models, as we reasoned that clinicians were always aware of these parameters. Cases of sepsis were individually reviewed for data accuracy. Health care personnel were blinded to the results of the HRC monitoring.

## Nearest-neighbor models

Conceptually, each point from the test set was placed among the points of the training set in a multidimensional space, and its nearest neighbors identified. For each of the 36,000 points in the training set, the distance from each point in the 35,000 point test set was calculated. The distances were calculated for HRC values, days of age and birth weight for all the data, and the process was repeated with additional contributions from each laboratory test individually and in several combinations. Self-records were excluded.

A very important aspect of implementing accurate nearest-neighbor models is the selection of the distance metric used to measure similarity. The Mahalanobis distance is a robust choice for measuring similarity (15). The Mahalanobis distance between two records $\mathbf{x}_i$, $\mathbf{x}_j$ is calculated by

$$d_{i,j}=\|\mathbf{A}(\mathbf{x}_i-\mathbf{x}_j)\|,$$

where $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T$ is a vector of available clinical data features of a record, and $\mathbf{A}$ is an $n \times n$ matrix that transforms each feature vector so that $x_{ik}$ ($k = 1,2,\cdots,n$) have the same scale and are uncorrelated. The transformation matrix $\mathbf{A}$ can be estimated from the training data set by

$$\mathbf{A}=\mathbf{V}\mathbf{S}^{-\frac{1}{2}},$$

where $\mathbf{S}$ is a diagonal matrix consisting of eigenvalues of the covariance matrix of the data set, and $\mathbf{V}$ is the orthogonal matrix of corresponding eigenvectors. The important results of this transformation are that variables with large numerical values (like WBC) do not dwarf those with small values (like I:T ratio), and variables that are highly correlated (like $pCO_2$, pH and $HCO_3$) are uncoupled in the analysis. Mahalanobis distance can be viewed as a simple Euclidean distance following that linear transformation.

For a new test record $\mathbf{x_i}$, distances from all records in the training data set to it are sorted in ascending order. The new record is assigned a value between 0 and 1 postulated to represent the chance of illness in the next 24 hours. If there are $k_s$ records that occurred within 24 hours of sepsis among the $k_i$ nearest neighbors in the training set, then the probability of the new record is also within 24 hours of illness is given by

$$p_i=\frac{k_s}{k_i}.$$

For this work, $k_s$ was selected to be 10 to ensure a reasonable degree of statistical accuracy of this probability. Results were similar for values between 5 and 10. There were 220 records near sepsis from 88 episodes of sepsis in the training set. Were they randomly distributed, the size of the neighborhood would be 10/220, or less than 5% of the data set. Thus the strategy allows effective localization of infants with similar clinical features within the large data set.

Each combination of HRC and laboratory values can be thought of as a predictive model and evaluated using standard metrics such as ROC area and Wald statistic.

## Logistic regression models

We used previously described techniques and adjusted for repeated measures (6–8;16;21).

### Combining nearest-neighbor and logistic regression models

Individual nearest-neighbor models with HRC and laboratory values were combined so that if one or more laboratory values were available, the maximum of the predictions was made the final prediction of the combined model. If none laboratory values were available, the final prediction adopted the outcome of the single model with only HRC and other basic variables (such as birth weight and days of age).

Models combining probability estimates from the nearest-neighbor and logistic regression analyses were based on bivariable logistic regression. The statistical significance of added information was calculated using the Wald chi-square test.

### HRC index and laboratory results

The HRC index has been previously described (6;8). Briefly, it is an internally and externally validated multivariable-regression based measure that is proportional to the risk of acute illness in infants in the NICU. In this analysis, we used individual measurements of the standard deviation, sample entropy (12;13;22) and sample asymmetry (R1 and R2 measures) (11), the major elements of the HRC index. Laboratory test results were obtained from an electronic archive.

### Study design

The first goal was to test the hypothesis that nearest-neighbor analyses alone make effective predictions on neonatal sepsis. All models included days of age and birth weight, as this information is always available to the clinician. We systematically tested nearest-neighbor models that added one laboratory test at a time, and all combinations of 2 laboratory results. We then added 4-dimensional HRC data (S.D., two sample asymmetry statistics, and sample entropy – the components of the regression-based HRC index) and repeated the procedure. ROC and Wald statistic were used to measure model performance.

The second goal was to test the hypothesis that nearest-neighbor and logistic regression models add independent information to each other, and that combined models improve prediction of neonatal sepsis. The approach was to prepare multiple nearest-neighbor models and regression models, evaluate the predictive performances of each model and compare them.

## RESULTS

Table 1 shows the demographic features and rates of sepsis in the training and overall data sets. Figure 1 shows the method for calculating the proportion of sick neighbors. As an example, a three-dimensional space is shown, with each axis representing a measurement modality – here, birth weight, day of age, and white blood cell count. The complete analysis included higher dimensions. The filled points are measurements from past infants who were within 24 hours of the diagnosis of sepsis. Two populations can be identified with different WBC values, as expected from clinical observation that septic infants might have abnormally high *or* low values.

Table 2 shows modeling results. The major finding is that many nearest-neighbor and logistic regression models were significantly associated with upcoming sepsis, with ROC areas as high as 0.71. The nearest-neighbor results validate the general approach of using the similarity of present data to past cases to estimate the risk of imminent illness. Most interestingly, a *combined model* consisting of the maximum (or the worst) prediction of several performed better, with ROC area 0.85, validating the clinicians' reasoning that a single clearly abnormal result trumps a number of other normal results.

To determine an *optimal model*, we tested all 256 possible combinations of HRC index and laboratory tests. The best predictive performance was returned for the combination of HRC index, WBC, I:T ratio and $HCO_3$, with ROC area 0.86 This is the model that was further evaluated in Figures 4 and 5.

The column of Table 2 entitled "Both" shows the results of combined models using both nearest-neighbor analysis and logistic regression. The important result is that the two kinds of models often added independent information to each other, as shown in the right-most columns. This validates the approach of using multiple predictive models, each of them incorporating multiple variables such as laboratory test results and HRC.

Figure 2 shows fit (Wald statistic) as a function of discrimination (ROC area) for bivariable regression models relating sepsis to probability estimates from logistic regression and nearest-neighbor models using the same predictor variables. Strategic combination of multiple models led to improved fit and discrimination.

Figure 3 shows the proportion of time that each model was applicable - that is, the appropriate lab tests were available - as well as the proportion of time that each model contributed the highest prediction value in the final model. The important result is that all models contributed to the final prediction probability. Since the variables were selected for their relevance to neonatal sepsis, the result is unlikely to be due to outliers from irrelevant data.

As noted above, the optimal model utilized HRC, WBC, I:T ratio and $HCO_3$, and Figure 4 shows an evaluation of its performance. The smooth line shows the output of the predictive model using coefficients determined from the training set. The circles are the observed results from the test set, and the boxes describe the 95% confidence limits determined by bootstrap. There is good agreement, with a sharp increase in predicted and observed probability in the top 10%. We defined this as a high-risk group (16), and we similarly defined low-risk (lowest 70%) and intermediate-risk (70th–$90^{th}$ percentile) groups.

Figure 5 shows the time-dependence of the change in risk stratified by these groupings. Initially, there is a very large distinction between the low- and high-risk groups. For example, there is a 20-fold increase in the relative risk of sepsis in the high-risk group compared to the low-risk group, from 0.28-fold to 5.5-fold increase in the average risk. Seven days after a measurement, there is still a more than 3-fold increase in the high-risk group compared to the low-risk group, from 0.70-fold to 2.3-fold increase in the average risk.

## DISCUSSION

We studied the use of predictive models based on nearest-neighbor analysis and on logistic regression in the early diagnosis of late-onset neonatal sepsis. Our major finding was that combining nearest-neighbor and logistic regression models, each based on multiple variables, lead to improved prediction. We incorporated the reasoning of clinicians that lessons learned from past cases were valuable.

Neonatal sepsis seems to be a particularly apt clinical problem to use nearest-neighbor analysis in creating predictive models. These infants are continuously monitored and have frequent lab testing but, since no single test has extremely high predictive performance, physicians are almost always uncertain of the diagnosis until signs of severe illness are present. The analytical strategy presented here should be useful in quantifying the experience of other physicians with other patients in an observational data base. The prediction result, of course, requires further contextual interpretation by the physician.

Nearest-neighbor analysis might find general utility in other clinical situations where the stakes are high and the data plentiful but unsorted. The combination of nearest-neighbor and logistic regression is especially appealing in multivariate applications where there are both linear and nonlinear associations. Logistic regression may have superior performance in handling the linear processes and nearest-neighbor may be more effective in treating the nonlinear components. The challenge to the algorithm-developer is optimal handling of continuous and intermittent data with unequal magnitude and variation, and unknown correlations. To adapt to the general situations that laboratory measures were taken at irregular times when there might be a need, we developed a highly flexible model that combines individual models associated with those laboratory tests but makes the final estimate based on what is available. Since nearest-neighbor rules are based on the assumption of independently distributed variables (15), we used Mahalanobis distance measure to deal with the problem of different kinds of data with unequal magnitude and unknown correlations.

## Limitations of the study

HRC data were not available for online inspection by physicians, and results are likely to be different when they are. In our hospital, real-time HRC monitoring has been in use since September, 2003, and has resulted in diagnosis and treatment of bacterial sepsis with no or only symptoms (9). CRP and newer tests for systemic inflammation are not in routine use at our hospital, but are likely to add useful diagnostic information (23). Finally, diagnosis of the presence or absence of neonatal sepsis is often uncertain. Blood cultures have notoriously poor diagnostic accuracy, especially in this setting where only very small blood samples can be spared (24). Moreover, neurodevelopmental abnormalities are the same in infants with clinical sepsis regardless of the blood culture result (25). As a result, the most recent guidelines for diagnosis of bloodstream infection in neonates lean heavily on clinical and laboratory findings other than blood cultures (26), with a preference for multivariable analysis (27).

The major limitation of any nearest-neighbor analysis is the database itself. If populated with incorrect or irrelevant data, there is obviously a deterioration of predictive performance. A more specific limitation is the nature of the outcome itself, the clinical and laboratory diagnosis of neonatal sepsis. Since blood cultures are a tarnished gold standard, there is irreducible uncertainty in the precise diagnosis of an infant with obvious clinical signs of illness but negative blood cultures. There is increasing awareness that many such infants have systemic inflammation and are vulnerable to identical neurodevelopmental impairment as infants with the same clinical illness but positive blood cultures (25). Further, the diagnosis of neonatal sepsis is relatively rare - an episode per 6–12 infant-months, or <1% of the time (1). As a result, most neighbors are “well” no matter how abnormal the HRC and clinical data. Our training set of 36,000 6-hour records held only 220 occurring within 24 hours of sepsis. To guard against including inappropriately large search spaces, we found the 10 nearest “sick” neighbors and calculated their proportional frequency. Were the abnormal records scattered randomly, our search space would include less than 5% of the total dataset.

In the nearest-neighbor analysis we used Mahalanobis distance metric instead of the more straight-forward Euclidean distance because it corrects for problems of analyzing correlated data with different scales. This advantage outweighs any theoretical limitation, but we note that the Mahalanobis distance was designed for multivariate normal data.

We have only used logistic regression and nearest-neighbor analysis to explore how data mining might aid physicians in the early diagnosis of neonatal sepsis. There are many other kinds of analysis that might be added or substituted, including neural and Bayesian networks, and decision tree analysis. There are many other possible variables to measure, including continuous ones like $O_2$ saturation monitoring and intermittent ones such CRP and other new lab markers of systemic inflammation. There are other disease processes for which this

approach would be useful in addition to systemic inflammatory response syndrome. Our targets are sub-acute, potentially catastrophic illnesses or complications. Examples include acute exacerbations of chronic asthma or chronic obstructive pulmonary disease, pneumonia, urinary tract infection and sepsis after brain or spinal cord injury, cancer recurrence, exacerbation of inflammatory bowel disease, severe hypoglycemia in insulin-requiring diabetes mellitus, complications of nursing home care such as bedsores, relapse of congestive heart failure, new or recurrent infection in HIV disease, community outbreaks of influenza or other infection, or effects of bioterrorism agents. For each setting, an observational data base that houses continuous and intermittent measures can be developed, and predictive models derived.

### Future work

We foresee challenges in implementation of nearest-neighbor analysis. First, since the major target events are infrequent compared to the testing, we can expect many false-positive results. We find this acceptable since the finding of a recent increase in risk of the target illnesses need result in only no-invasive or minimally invasive testing such as imaging or blood sampling. This seems warranted by the potentially catastrophic outcome of late diagnosis and late treatment. The goal is an early detection system for increased risk, not a substitute for a physician. Second, as with all predictive models, there is danger of overfitting. Here, we have guarded against this by developing and validating predictive algorithms in separate populations, and by appropriate adjustments when repeated measures are employed. Third, not all clinical problems are suitable for informatics monitoring approaches. Truly sudden catastrophes with either no prodrome or ones that are too short to allow intervention are unlikely to be amenable to predictive algorithms. It is possible, for example, that arterial thrombosis leading to acute myocardial infarction or stroke has no prodrome. Finally, the issue of liability must be prospectively addressed. Consider a situation in which a monitored patient's risk for, say, sepsis in the setting of spinal cord injury increases more than three-fold in the middle of the night. Is one negligent to ignore this until office hours?

## CONCLUSION

Nearest neighbor analysis adds to existing logistic regression methods for early diagnosis of neonatal sepsis. Nearest-neighbor analysis is a novel approach to predicting imminent illness, and both logistic regression and nearest-neighbor analysis models based on heart rate characteristics and laboratory test results contribute independent information. The best predictions employed both kinds of models, and were driven by the single most abnormal finding. This approach to predicting illness may prove useful in other clinical situations.

## Acknowledgments

We thank WE King for suggesting nearest-neighbor analysis to us.

Supported by NIGMS-64640; American Heart Association, Mid-Atlantic Affiliate; Children's Medical Center Research Fund, University of Virginia; Virginia's Center for Innovative Technology; and Medical Predictive Science Corporation, Charlottesville, VA.

## REFERENCES

1. Stoll BJ, Hansen N, Fanaroff AA, Wright LL, Carlo WA, Ehrenkranz RA, et al. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics 2002;110(2 Pt 1):285–291. [PubMed: 12165580]
2. Fanaroff AA, Korones SB, Wright LL, Verter J, Poland RL, Bauer CR, et al. Incidence, presenting features, risk factors and significance of late onset septicemia in very low birth weight infants. The National Institute of Child Health and Human Development Neonatal Research Network. Pediatr Infect Dis J 1998;17(7):593–598. [PubMed: 9686724]

3. Escobar GJ. The neonatal "sepsis work-up": personal reflections on the development of an evidence-based approach toward newborn infections in a managed care organization. Pediatrics 1999;103(1 Suppl E):360–373. [PubMed: 9917478]
4. Bone RC. Important new findings in sepsis. JAMA 1997;278(3):249. [PubMed: 9218676]
5. Brilli RJ, Goldstein B. Pediatric sepsis definitions: Past, present, and future. Pediatr Crit Care Med 2005;6(3):S6–S8. [PubMed: 15857561]
6. Griffin MP, O'Shea TM, Bissonette EA, Harrell FE Jr. Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. Pediatr Res 2003;53:920–926. [PubMed: 12646726]
7. Griffin MP, O'Shea TM, Bissonette EA, Harrell FE Jr. Lake DE, Moorman JR. Abnormal heart rate characteristics are associated with neonatal mortality. Pediatr Res 2004;55:782–788. [PubMed: 14739356]
8. Griffin MP, Lake DE, Bissonette EA, Harrell FE Jr. O'Shea TM, Moorman JR. Heart rate characteristics: novel physiomarkers to predict neonatal infection and death. Pediatrics 2005;116:1070–1074. [PubMed: 16263991]
9. Moorman JR, Lake DE, Griffin MP. Heart rate characteristics monitoring in neonatal sepsis. IEEE Transactions in Biomedical Engineering. 2005 in press.
10. Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. Pediatrics 2001;107:97–104. [PubMed: 11134441]
11. Kovatchev BP, Farhy LS, Cao H, Griffin MP, Lake DE, Moorman JR. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. Pediatr Res 2003;54(6):892–898. [PubMed: 12930915]
12. Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. Am J Physiol 2002;283:R789–R797.
13. Richman JS, Moorman JR. Physiological time series analysis using approximate entropy and sample entropy. Am J Physiol 2000;278:H2039–H2049.
14. Fukunaga, K. Introduction to statistical pattern recognition. San Diego: Academic Press; 1990.
15. Devijver, PA.; Kittler, J. Pattern recognition: a statistical approach. London: Prentice-Hall; 1982.
16. Griffin MP, Lake DE, Moorman JR. Heart rate characteristics and laboratory tests in neonatal sepsis. Pediatrics 2005;115:937–941. [PubMed: 15805367]
17. Haddad M, Adlassnig KP, Porenta G. Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. Artif Intell Med 1997;9 (1):61–78. [PubMed: 9021059]
18. Qu H, Gotman J. A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: possible use as a warning device. IEEE Trans Biomed Eng 1997;44(2):115–122. [PubMed: 9214791]
19. Costa M, Moody GB, Henry I, Goldberger AL. PhysioNet: an NIH research resource for complex signals. J Electrocardiol 2003;36 Suppl:139–144. 139–144. [PubMed: 14716615]
20. Garner JS, Jarvis WR, Emori TG, Horan TC, Hughes JM. CDC definitions for nosocomial infections, 1988. Am J Infect Control 1988;16(3):128–140. [PubMed: 2841893]
21. Harrell, FE, Jr.. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. Berlin: Springer; 2001.
22. Richman JS, Lake DE, Moorman JR. Sample entropy. Methods Enzymol 2004;384:172–184. [PubMed: 15081687]
23. Malik A, Hui CP, Pennie RA, Kirpalani H. Beyond the complete blood cell count and C-reactive protein: a systematic review of modern diagnostic tests for neonatal sepsis. Arch Pediatr Adolesc Med 2003;157(6):511–516. [PubMed: 12796229]
24. Jawaheer G, Neal TJ, Shaw NJ. Blood culture volume and detection of coagulase negative staphylococcal septicaemia in neonates. Arch Dis Child Fetal Neonatal Ed 1997;76(1):F57–F58. [PubMed: 9059190]
25. Stoll BJ, Hansen NI, Adams-Chapman I, Fanaroff AA, Hintz SR, Vohr B, et al. Neurodevelopmental and growth impairment among extremely low-birth-weight infants with neonatal infection. JAMA 2004;292(19):2357–2365. [PubMed: 15547163]

26. Haque KN. Definitions of bloodstream infection in the newborn. Pediatr Crit Care Med 2005;6 (3):S45–S49. [PubMed: 15857558]

27. Escobar GJ. What have we learned from observational studies on neonatal sepsis? Pediatr Crit Care Med 2005;6(3):S138–S145. [PubMed: 15857547]

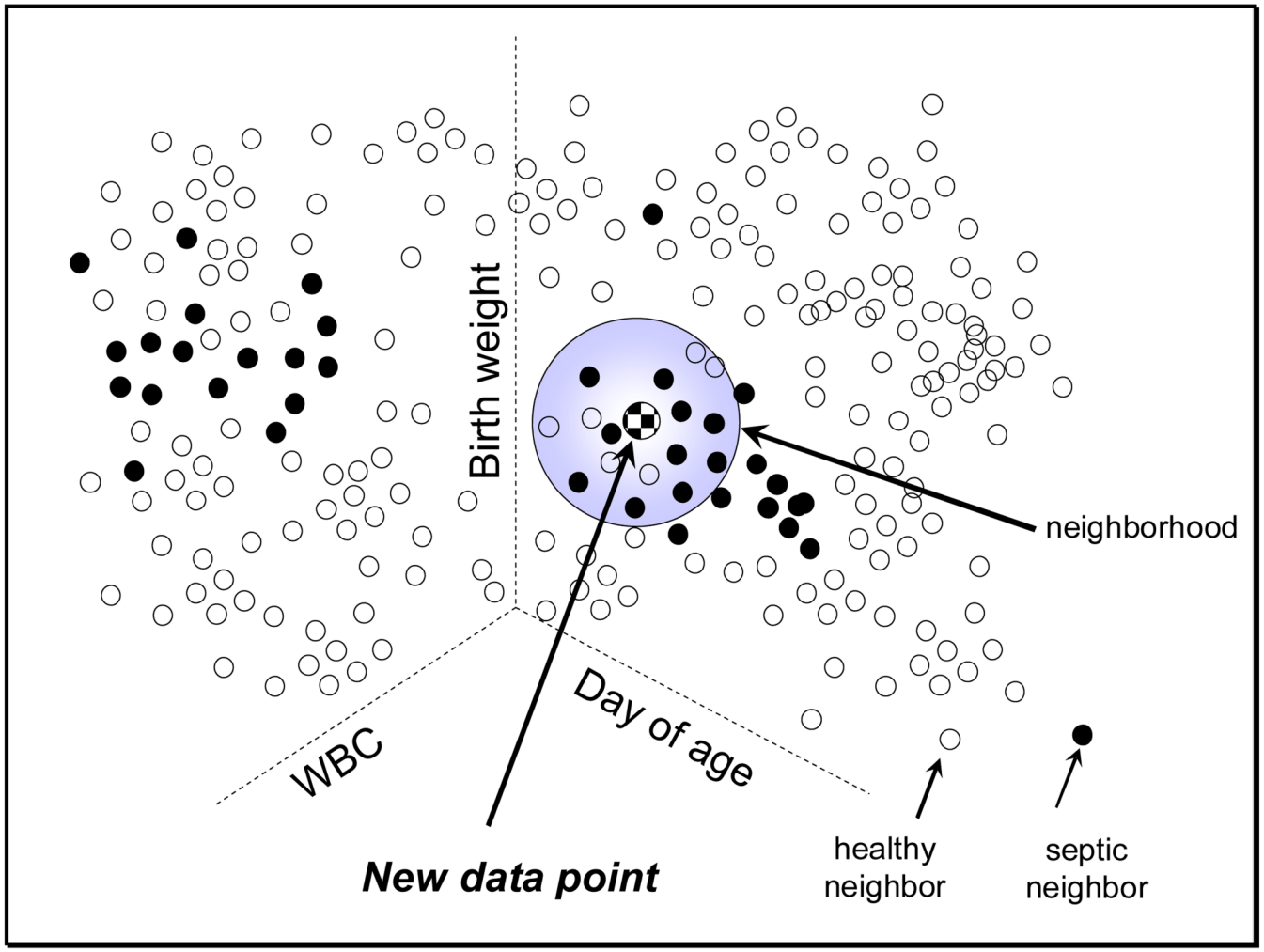


**Figure 1.**
Method of nearest-neighbor analysis. The plot is a stylized representation of a three-dimensional space in which each point is a WBC value measured on a known day of age in an infant of known birth weight. Filled points occurred in infants within 24 hours of a positive blood culture obtained for signs of sepsis. There are two clusters of filled points, indicating sepsis occurring in infants with low or high WBC values. The test point is at the center of the shaded sphere. Conceptually, the sphere is enlarged until it contains 10 filled points, and the proportion of filled to total points is calculated. This probability measure is the nearest-neighbor analysis result of the likelihood of imminent sepsis.

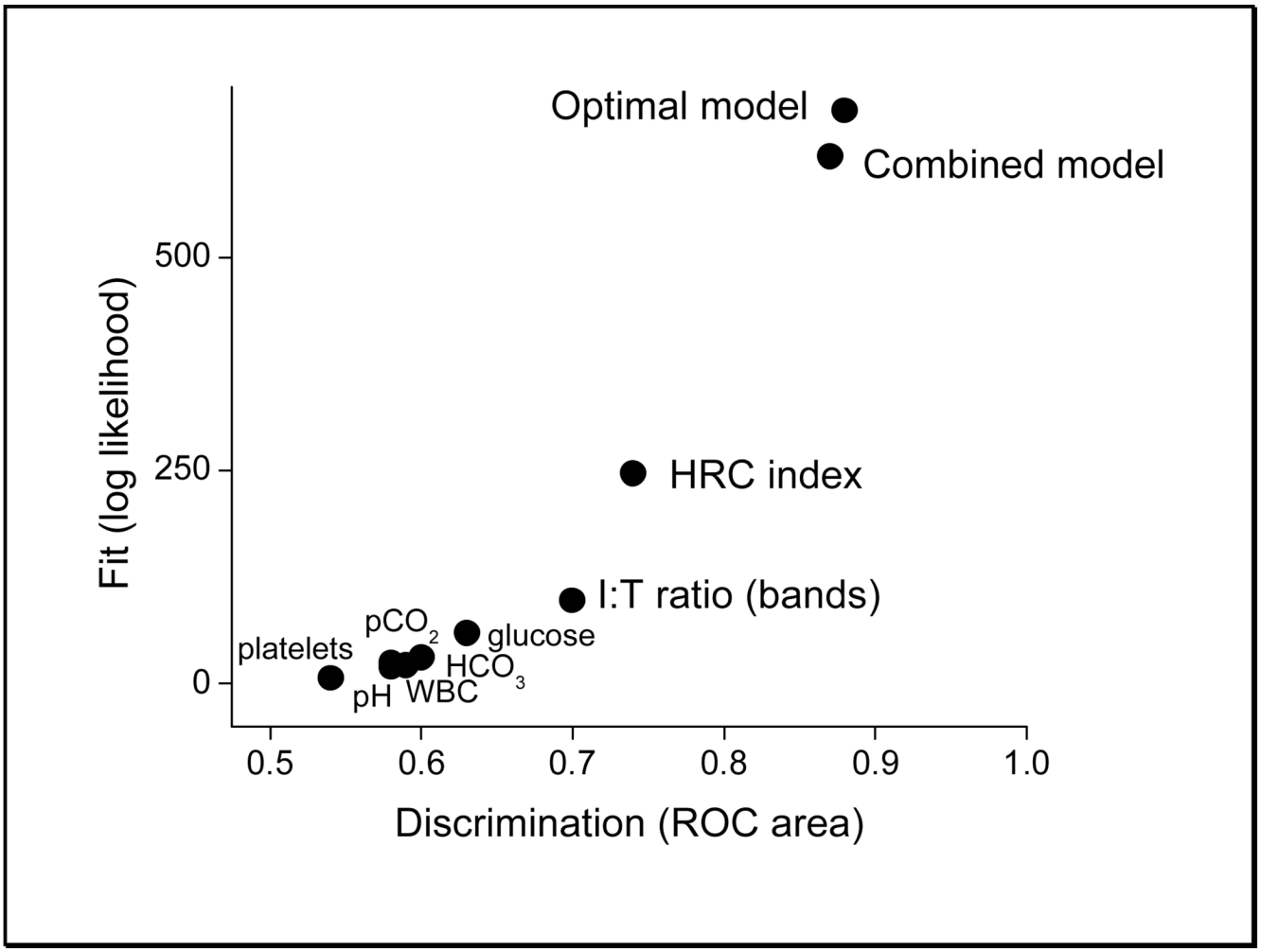


**Figure 2.**

Combination of multiple nearest-neighbor models lead to improved prediction of illness. The fit of the predictive models, calculated as the Wald Chi-square statistic of the data given the model, is plotted as a function of the area under the receiver-operating characteristic curve. Of the models using a single variable (in addition to day of age and birth weight), HRC index had the best performance and I:T ratio the next best. Either strategy of model combination had better performance.

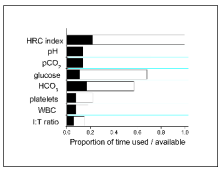


**Figure 3.**
Availability and utilization of predictive variables. The filled section of each bar is the proportion of the total time for which the model incorporating the specified variable led to the highest predicted probability of illness, and the open section is the proportion of time for which the variable was available. By the study design, HRC was available all of the time.

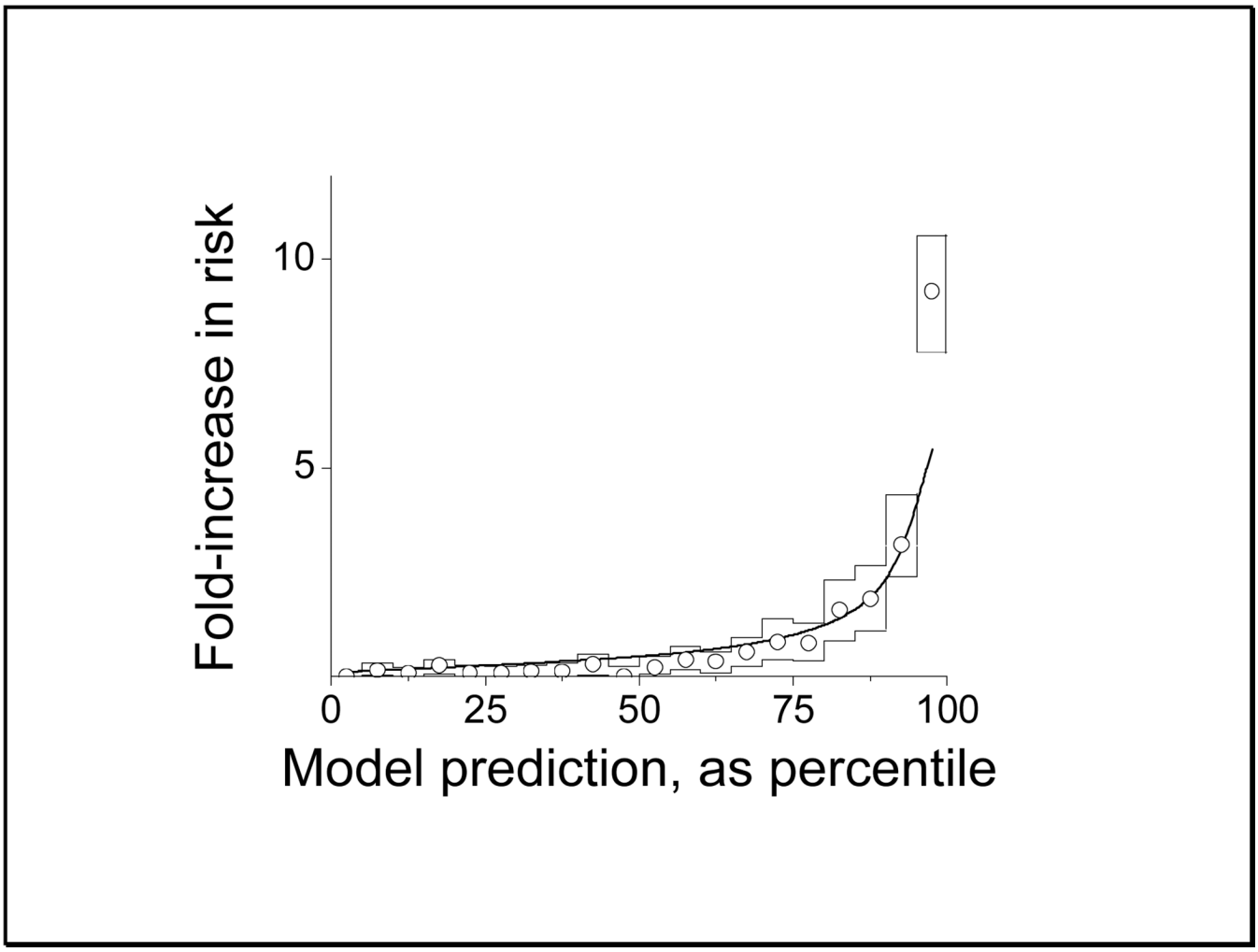


**Figure 4.**
Internal validation of the nearest-neighbor model of selected lab tests. The plot shows predicted (smooth line) and observed (circles) rates of sepsis based on percentile of the predicted probability. The boxes show the 95% confidence limits of the observed rates.

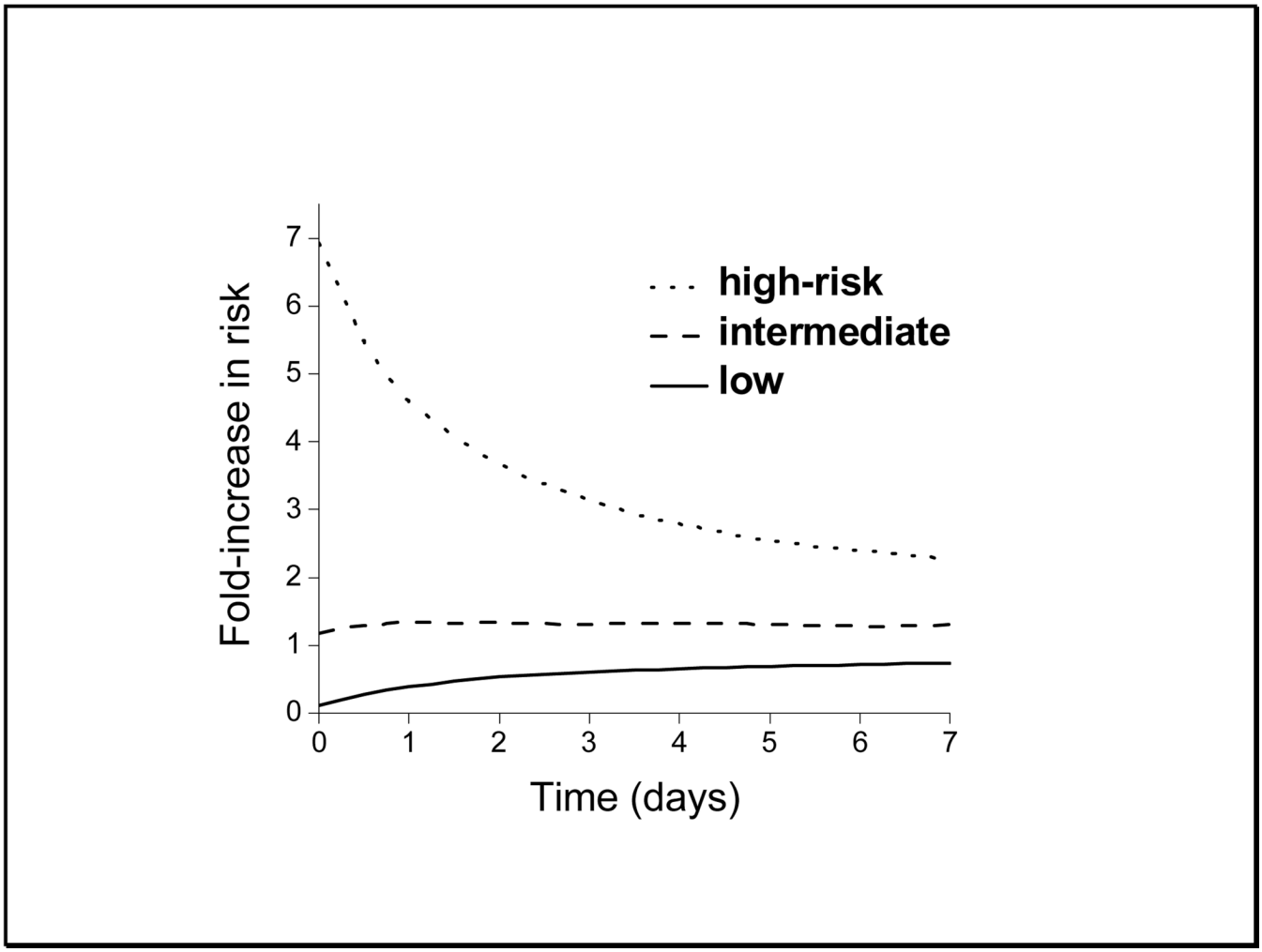


**Figure 5.**
Relative risks of sepsis in low-, intermediate- and high-risk groups based on the model of selected lab tests.

## Table 1

Patient population and laboratory test results

| | Training set | All infants |
|---|---|---|
| N | 327 | 676 |
| #HRC | 36,309 | 71,254 |
| BW (g) | 1454 (973,2620) | 1581 (974,2700) |
| <1500 g | 166 (51%) | 317 (47%) |
| GA (weeks) | 31 (27, 36) | 31 (27, 36) |
| Episodes of sepsis | 88 (67 infants) | 163 (120) |
| Laboratory test results | | |
| WBC | 6723 | 13,014 |
| I:T ratio | 5616 | 10,887 |
| glucose | 24,592 | 48,312 |
| platelet count | 8275 | 16,144 |
| pH, $pCO_2$ | 25,865 | 50,464 |
| $HCO_3$ | 20,828 | 40,630 |

#HRC = number of 6-hour HRC records; BW = birth weight and GA is gestational age, and are given as median ($25^{th}$, $75^{th}$ percentiles); WBC = white blood cell count; I:T ratio = ratio of immature to total white blood cell forms.

**Table 2**

Modeling results.

| | N-n ROC | N-n Wald stat | | LR ROC | N-n Wald stat | | Both: ROC | Both: Wald stat | N-n add? | LR add? |
|---|---|---|---|---|---|---|---|---|---|---|
| HRC | 0.71 | 134 | | 0.74 | 105 | | 0.74 | 137 | * | * |
| WBC | 0.57 | 12 | | 0.53 | 3.1 | | 0.58 | 15 | * | |
| I:T ratio | 0.67 | 58 | | 0.69 | 60 | | 0.70 | 74 | * | * |
| platelet count | 0.61 | 33 | | 0.62 | 55 | | 0.63 | 58 | | * |
| glucose | 0.51 | 0.4 | | 0.54 | 3.5 | | 0.54 | 3.8 | | |
| $pCO_2$ | 0.56 | 5.5 | | 0.58 | 5.8 | | 0.58 | 6.9 | | * |
| pH | 0.57 | 6.3 | | 0.58 | 6.2 | | 0.59 | 9.6 | | * |
| $HCO_3$ | 0.53 | 0.9 | | 0.58 | 9.0 | | 0.60 | 15 | * | * |
| *Combined model* | 0.85 | 317 | | 0.87 | 311 | | 0.87 | 472 | * | * |
| *Optimal model* | 0.86 | 358 | | 0.87 | 319 | | 0.88 | 480 | * | * |

N-n = nearest-neighbor analysis; LR = logistic regression; both = bivariable regression model using results of nearest-neighbor analysis and multivariable logistic regression models; ROC = receiver-operating characteristic area; Wald stat = Wald statistic; HRC = heart rate characteristics; WBC = white blood cell count; I:T ratio = ratio of immature to total white blood cell forms; N-n add * = nearest-neighbor model added significant information to logistic regression model ($p<0.05$); LR add * = logistic regression model added significant information to nearest-neighbor model ($p<0.05$).