www.nature.com/pr

Pediatric RESEARCH

CLINICAL RESEARCH ARTICLE

A critical evaluation of current definitions of necrotizing enterocolitis

Shiloh R. Lueschow[1], Timothy J. Boly[2], Elizabeth Jasper[3,4,5], Ravi M. Patel[6] and Steven J. McElroy[1,2]

BACKGROUND: Necrotizing enterocolitis (NEC) is a devastating intestinal disease of premature infants, with significant mortality and long-term morbidity among survivors. Multiple NEC definitions exist, but no formal head-to-head evaluation has been performed. We hypothesized that contemporary definitions would perform better in evaluation metrics than Bell's and range features would be more frequently identified as important than yes/no features.

METHODS: Two hundred and nineteen patients from the University of Iowa hospital with NEC, intestinal perforation, or NEC concern were identified from a 10-year retrospective cohort. NEC presence was confirmed by a blinded investigator. Evaluation metrics were calculated using statistics and six supervised machine learning classifiers for current NEC definitions. Feature importance evaluation was performed on each decision tree classifier.

RESULTS: Newer definitions outperformed Bell's staging using both standard statistics and most machine learning classifiers. The decision tree classifier had the highest overall machine learning scores, which resulted in Non-Bell definitions having high sensitivity (0.826, INC) and specificity (0.969, ST), while Modified Bell (IIA+) had reasonable sensitivity (0.783), but poor specificity (0.531). Feature importance evaluation identified nine criteria as important for diagnosis.

CONCLUSIONS: This preliminary study suggests that Non-Bell NEC definitions may be better at diagnosing NEC and calls for further examination of definitions and important criteria.

Pediatric Research (2022) 91:590–597; https://doi.org/10.1038/s41390-021-01570-y

IMPACT:

- This article is the first formal head-to-head evaluation of current available definitions of NEC.
- Non-Bell NEC definitions may be more effective in identifying NEC based on findings from traditional measures of diagnostic performance and machine learning techniques.
- Nine features were identified as important for diagnosis from the definitions evaluated within the decision tree when performing supervised classification machine learning.
- This article serves as a preliminary study to formally evaluate the definitions of NEC utilized and should be expounded upon with a larger and more diverse patient cohort.

INTRODUCTION

Necrotizing enterocolitis (NEC) is an inflammatory bowel disease that primarily afflicts premature infants due to prematurity of the intestine and prematurity of the immune system leading to an inability to immunomodulate.[1–5] NEC is associated with extreme levels of intestinal inflammation and ranges in pathology from patchy to total intestinal necrosis (NEC totalis).[1–5] NEC is the leading cause of gastrointestinal morbidity and mortality in preterm infants with an estimated 30–50% mortality rate depending on disease severity.[1–5] Therefore, NEC represents a significant ailment in preterm neonates. Although NEC was formally described in 1965 by Mizrahi et al., the etiology has yet to be fully established despite decades of research. Treatment strategies for NEC remain limited, non-targeted, and have potential drawbacks and risks, including short bowel syndrome and intestinal failure.[1,4,6] In the past few decades, little improvement has been made in treatment strategies available for NEC, potentially because clinicians have difficulty diagnosing NEC until more severe disease stages have been reached. Over the years, many studies have also been targeted at identifying and evaluating biomarkers for NEC, but there have been mixed results in the sensitivity and specificity of the potential biomarkers discovered.[7,8] Additionally, some biomarkers identified have been criticized for their subjective nature, which leads to questions of consistency across institutions.[7,8]

[1]Department of Microbiology & Immunology, University of Iowa, Iowa City, IA, USA; [2]Stead Family Department of Pediatrics, University of Iowa, Iowa City, IA, USA; [3]Division of Quantitative Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center, Nashville, TN, USA; [4]Center for Precision Medicine, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [5]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA and [6]Department of Pediatrics, Emory University and Children's Healthcare of Atlanta, Atlanta, GA, USA
Correspondence: Steven J. McElroy (steven-mcelroy@uiowa.edu)

Received: 17 February 2021 Revised: 19 April 2021 Accepted: 26 April 2021
Published online: 21 May 2021

a

	Bell staging			Modified Bell staging						Stanford	United Kingdom	2 of 3	International Neonatal Consortium	Vermont Oxford Netwerk	Centers for Disease Control
	I	II	III	IA	IB	IIA	IIB	IIIA	IIIB						
Abbreviation										ST	UK	2of3	INC	VON	CDC
Definition origin	[9]			[10]						[15]	[12]	[14]	[16]	[13]	[17]
Publication year	1978			1986						2014	2017	2018	2019	2019	2020
Number of features	12	17	20	10	11	13	19	26	27	10	12	12	11	8	7
Risk grouping	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	3/4	1/4	2/4	2/4	0/4	0/4
Exclusion criteria	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	1/4	all	3/4	1/4	0/4
Systemic signs	4/11	4/11	5/11	4/11	4/11	4/11	6/11	10/11	10/11	3/11	0/11	1/11	2/11	0/11	0/11
Intestinal signs	7/16	8/16	9/16	4/16	5/16	7/16	9/16	12/16	12/16	2/16	6/16	2/16	2/16	4/16	4/16
Radiologic findings	1/8	5/8	6/8	2/8	2/8	2/8	4/8	4/8	5/8	2/8	4/8	3/8	2/8	3/8	3/8

b

Category of features	Alterations from original definition criteria
Risk grouping	None of the features were altered from original defintion in this category
Exclusion criteria	SIP and congenital anomaly eliminated, Fed <80ml/kg/day and GA ≥36 GA considered as a standard yes/no criteria
Systemic signs	None of the features were altered from original defintion in this category
Intestinal signs	Marked hemorrhage and right low quadrant mass not used
Radiologic findings	Ileus and dilation/distension separated into 2 features, small bowel separation not used

Fig. 1 Description of features of current definitions of NEC and methods of pruning for machine learning classifiers. a The origin of the NEC definition as well as a comparison of the number of various features utilized by the definitions based on the figure from the 2020 publication by Patel et al..[7] **b** A table depicting the various alterations from the original NEC definition criteria utilized after pruning, preprocessing, and accounting for missing information for the machine learning classifiers.

Many definitions for NEC have been developed using different criteria to aid in diagnosis.[4,9,10] The first clinical staging system was proposed by Bell et al. in 1978,[9] which was later modified by Walsh and Kliegman in 1986.[10] Bell staging and Modified Bell staging continue to be the most commonly used clinical definitions of NEC to date. Modified Bell's criteria have been criticized as they are not specific to the disease process of NEC until the disease has progressed to its more severe forms.[4,9–11] In response to the limitations of Bell's criteria, six contemporary definitions of NEC have been proposed, including The original modification of Bell's criteria (ModBell),[10] United Kingdom (UK),[12] Vermont Oxford Network (VON),[13] Two of three (2of3),[14] Stanford (ST),[15] International Neonatal Consortium (INC),[16] and Centers for Disease Control (CDC)[17] (Fig. 1a). Although two recent studies have been published determining the capacity of definitions to distinguish between NEC and spontaneous intestinal perforation (SIP) using machine learning techniques, to date, no head-to-head evaluation has been performed to determine the functionality of the individual definitions of NEC to correctly make a diagnosis between NEC infants and those without NEC.[18,19] To attempt to close this knowledge gap, we utilized a single-center, retrospective cohort and performed traditional statistical measures of diagnostic performance and supervised classification machine learning on the current definitions of NEC to evaluate the sensitivity, specificity, and accuracy of each definition, and supervised machine learning was used to determine the area under the receiver operating characteristic curve (AUROC). Finally, the most important features for each NEC definition were identified through the respective decision tree classifier.

MATERIALS AND METHODS

The dataset and statistical evaluation

We conducted a retrospective cohort study of infants born at the Stead Family Children's Hospital at the University of Iowa from February 1, 2008, to September 1, 2019 with an International Classification of Diseases (ICD)-9 based diagnosis of "concern for NEC" and/or "concern for intestinal perforation" following institutional review board approval (University of Iowa IRB #201410743). No infants were excluded and 219 infants were eligible for analysis. One infant had two separate inclusion events, and these were collected separately for a total of 220 patient events used for analysis. Demographic data as well as data pertaining to each of the NEC definitions were obtained from medical records on the day of illness onset (Table 1). The clinical diagnosis of NEC, to which NEC definitions were evaluated, was determined by a single blinded investigator based on the medical record and supporting clinical information. The decision was based on: physical findings, such as bloody stools and abdominal distension with discoloration; laboratory evidence of inflammation, including leukocytosis, elevated C-reactive protein, or thrombocytopenia; radiographic evidence concerning for pneumatosis intestinalis or portal venous gas; and when available, evidence of intestinal inflammation and necrosis present on specimens from surgery or autopsy. Infants were considered to have definitive NEC if there were >2 physical exam findings, definite laboratory evidence of inflammation, and radiographic changes. Pathology was used to support the diagnosis when available. All other infants were placed in the non-NEC cohort. Those who were diagnosed with SIP were also assigned to the non-NEC group as the primary goal of our study

Table 1. Demographic data of the 219 patients included in this study.

Characteristic	Clinical diagnosis of NEC ($n = 102$) Number with SIP ($n = 0$)	No diagnosis of NEC ($n = 117$) Number with SIP ($n = 30$)
Gender		
Male, *n* (%)	57 (56)	61 (52)
Female, *n* (%)	45 (44)	56 (48)
Race		
White, *n* (%)	76 (74.5)	87 (74.2)
Black, *n* (%)	17 (16.6)	19 (16.2)
Other, *n* (%)	9 (8.8)	11 (9.4)
Gestational age		
<28 weeks, *n* (%)	42 (41.2)	29 (24.8)
28 0/7–31 6/7 weeks, *n* (%)	28 (27.5)	17 (14.5)
32 0/7–36 6/7 weeks, *n* (%)	23 (22.5)	37 (31.6)
>37 weeks, *n* (%)	9 (8.8)	34 (29.1)
Birth weight		
<1000 g, *n* (%)	44 (43.1)	36 (30.8)
1000–1500 g, *n* (%)	22 (21.6)	7 (5.9)
1501–2500 g, *n* (%)	26 (25.5)	40 (34.2)
>2500 g, *n* (%)	10 (9.8)	34 (29.1)
Range (average) postnatal age at the onset of symptoms, days	0–84 (17.3)	0–120 (24.6)

Demographics data includes gender, race, gestational age, birth weight, and average age at the onset of symptoms.

was to determine the accuracy of NEC diagnosis. SIP was diagnosed in those with abdominal distension or discoloration, with radiographic evidence of intraperitoneal free air, and no or minimal inflammatory response. When available, pathology was used to support this diagnosis. It is important to note that, while all infants in this study had an ICD-9 code of "Concern for NEC", many of these infants did not ultimately have NEC, which explains our need for the NEC and non-NEC cohorts despite all having the same ICD-9 code.

Presence or absence of definition criteria were then determined for each case. NEC definitions that were evaluated included Modified Bell (IA-IIIB), ModBell(IIA+), UK, VON, CDC, 2of3, ST, and INC (Fig. 1a). Due to the retrospective nature of this study, no infant had complete documentation (e.g., absence of findings) of each of the criteria/evaluations necessary for each individual definition to be examined on its own, thus we utilized a best fit strategy based on the criteria available. When a criterion was not documented, it was considered not present. The best fit definition was determined per patient based on the definition containing the most inclusion criteria without also containing exclusion criteria. If multiple definitions had equal total inclusion criteria met, the infant was categorized into all appropriate definitions. They were not categorized into a definition if any exclusion criteria were present. Due to the large number of inclusion criteria for the ModBell(IIA+) classification, most infants were initially categorized under this classification. To better assess the more recent definitions, a second categorization was performed excluding ModBell(IIA+). The sensitivity, specificity, and accuracy for each clinical definition was then determined.

Preprocessing for machine learning

From the original dataset of 67 features, which included clinical systemic findings, abdominal exam findings, and radiographic characteristics, pruning was done involving several filtering steps where criteria were eliminated (Fig. 1b). Criteria were removed from the dataset if they were too diverse to be meaningful or if there was not an easily justifiable way to assign a meaningful category number to a patient. An example of this was major congenital anomaly, where out of the 220 patients, only 57 had a major congenital anomaly. Although gastroschisis was the most commonly identified anomaly (13/57), there were approximately 42 different anomalies recognized with some only being in one patient and some patients having multiple different anomalies. Features with excessive missing information were also removed, which was defined as having data available for less than one-quarter of the patient population. In this study, only one of the exclusion criteria was missing that much information, which was age at the onset of SIP (Fig. 1b). Features where all answers were "no"/zero for all patients were also removed, which included marked hemorrhage, right lower quadrant mass, and small bowel separation (Fig. 1b). Two features were added to the analysis, which included splitting "Ileus, intestinal dilation or distension" into two separate features, "Ileus" and "intestinal dilation or distension." Also, "ultrasound used" was added to the all features definition, which was not an original feature for a specific definition.

A total of 43 different features were utilized by at least one definition, but only 41 features made it through the pruning step to go into the preprocessing step. Thirty-three out of the 41 features were "yes"/"no" responses, which were converted to 1 and 0, respectively. Information not documented was assumed to be "no" and converted to zero. For the purposes of the machine learning classifiers, the exclusion criteria depicted in Fig. 1b were not considered as exclusion criteria. SIP and major congenital anomaly were not utilized as criteria because as mentioned before they did not pass though the pruning step. The exclusion criteria "Fed <80 ml/kg/day" and gestational age "(GA) ≥ 36 weeks" were both considered as regular yes/no features instead of exclusion criteria. One patient was missing information for the majority of features and was excluded from machine learning analyses. Finally, for features like GA or volume of feeding at NEC onset where the outcome was a range of information, missing data were imputed with the mean for that feature. After preprocessing, the criteria needed for each of the NEC definitions was individually compiled to use in the machine learning models. After condensing the Bell staging (II, III) and Modified Bell staging (IIA, IIB, IIIA, IIIB) into one definition, ModBell(IIA+), seven NEC definitions had available criteria. Because SIP was removed in the pruning step, the VON and CDC definitions were identical and were compressed into one definition represented as VON in figures.

Analysis of NEC definition criteria

Scikit-learn was used within Jupyter Notebook (version 6.0.0) to set up each of the machine learning classifiers.[20–22] Within Scikit-learn, six different supervised machine learning classifiers were used: K nearest neighbors (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Simple Neural Network (SNN), Random Forest (RF), and Decision Tree (DT).[20–22] For each definition, the data were individually split into a training (75%) and test set (25%).[21,22] External validation was not performed due to the lack of sufficient patient data. Each model run was trained to take in all the various criteria for each definition and classify each patient based on the presence of NEC (one) or absence (zero). Each classifier was then evaluated based on the test set performance for sensitivity, specificity, accuracy, and the AUROC.[21,22] ModBell(IIA+) encompassed Bell stage II and III, and ModBell IIA, IIB, IIIA, and IIIB, so to produce a single score for this definition, the average score was taken for each metric from these individual definitions to be the representative score. To optimize each of the classifiers, different parameters were adjusted (Fig. 2a, b). For KNN, the number of neighbors was altered, while for NB the type utilized including

A

	KNN	SNN	NB	RF	SVM	DT
ST	1	lbfgs solver	Multinomial	50 estimators	Gamma = 2	Max depth = 2
UK	3	2 layers	Multinomial	500 estimators 5 max features	Gamma = 2	Max depth = 3
2of3	3	lbfgs solver 3000 maxit	Multinomial	500 estimators 5 max features	Gamma = 5	Max depth = 4 Min leaf = 5
INC	1	lbfgs solver 2 layers 4000 maxit	Multinomial	100 estimators 5 max features	Gamma = 2	Max depth = 2 Min leaf = 5
VON	3	2 layers	Multinomial	100 estimators	Gamma = scale	Max depth = 2 Min leaf = 5

B

	KNN	SNN	NB	RF	SVM	DT
Bell 2	3	lbfgs solver	Multinomial	50 estimators	Gamma = 2	Max depth = 2
Bell 3	2	1000 maxit	Multinomial	500 estimators 4 max features	Gamma = 2	Max depth = 2
ModBell 2A	3	1400 maxit	Gaussian	500 estimators 5 max features	Gamma = scale	Max depth = 2
ModBell 2B	1	lbfgs solver 1000 maxit	Multinomial	500 estimators	Gamma = 3	Max depth = 2
ModBell 3A	3	2 layers	Multinomial	100 estimators	Gamma = scale	Max depth = 2
ModBell 3B	3	2 layers	Multinomial	500 esimators	Gamma = scale	Max depth = 2 Min leaf = 3

Fig. 2 Optimal parameters used for machine learning classifiers. a The optimal parameters for each Non-Bell NEC definition using the various machine learning classifiers analyzed in this study. **b** The optimal parameters for each Bell and Modified Bell NEC definition that constituted the ModBell(IIA+) definition using the various machine learning classifiers analyzed in this study.

Gaussian, Multinomial, and Complement was changed (Fig. 2a, b).[20–22] The tweaked parameters for SVM were gamma set either to scale or two through five and linear kernel with C (regularization parameter) set to 0.025 (Fig. 2a, b).[20–22] For SNN, the parameters altered were limiting the maximum number of iterations (maxit); changing the solver to Limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) to accommodate the smaller dataset; along with limiting the number of layers to 2 rather than 3 (Fig. 2a, b).[20–23] The parameters altered for RF were the number of estimators (50, 100, and 500 were tested) and limiting the maximum number of features parameter (Fig. 2a, b).[20–22] Finally, for DT the parameters altered were the maximum depth and minimum number of samples per leaf (Fig. 2a, b).[20–22]

Feature importance evaluation

For the decision tree classifier, feature importance evaluation within Scikit-learn was performed for 14 definitions of NEC, including Bell I, Bell II, Bell III, ModBell IA, ModBell IIA, ModBell IIIA, ModBell IB, ModBell IIB, ModBell IIIB, UK, VON, 2of3, ST, and INC.[21,22] Additionally, an "All features" definition was produced and evaluated based on the entire dataset of 41 features utilized by at least one of the definitions. For each definition's decision tree classifier, feature importance evaluation generated a feature importance score for each feature. All features with an importance score ≥0.1 were considered important and were further evaluated. Finally, after determination of the most important features, a new dataset was established using the most important features and evaluated with a decision tree classifier using the metrics described above.

RESULTS

Sensitivity and specificity evaluation

The ModBell(IIA+) and UK definitions had the highest sensitivity in standard statistics (0.706 and 0.745, respectively), while the other newer definitions of NEC including ST, CDC, VON, 2of3, and INC had lower sensitivity (0.412, 0.314, 0.314, 0.294, and 0.225, respectively) (Fig. 3a). Sensitivity was more consistent and higher using machine learning classifiers with the exception of SVM, where 4 of the 6 definitions had extremely poor scores ranging from 0 to 0.130 (Fig. 3a). The overall average sensitivity score using machine learning classifiers (excluding the SVM classifier) among the definitions was 0.708. Within the remaining classifiers, sensitivity scores ranged from a low in ST definition at 0.522 using the DT classifier to a high in UK and INC definitions at 0.826 using the NB and DT classifiers, respectively (Fig. 3a).

The ModBell(IIA+), CDC, UK, and ST definitions had low specificity in standard statistics (0.402, 0.479, 0.504, and 0.53, respectively), while the 2of3 and INC definitions had high specificity (0.88 and 0.897, respectively, Fig. 3b). As seen with sensitivity, increased specificity was noted using machine learning classifiers (Fig. 3b). SVM specificity scores were eliminated from average specificity score calculation to maintain consistency with sensitivity evaluation and to prevent skewing from extreme

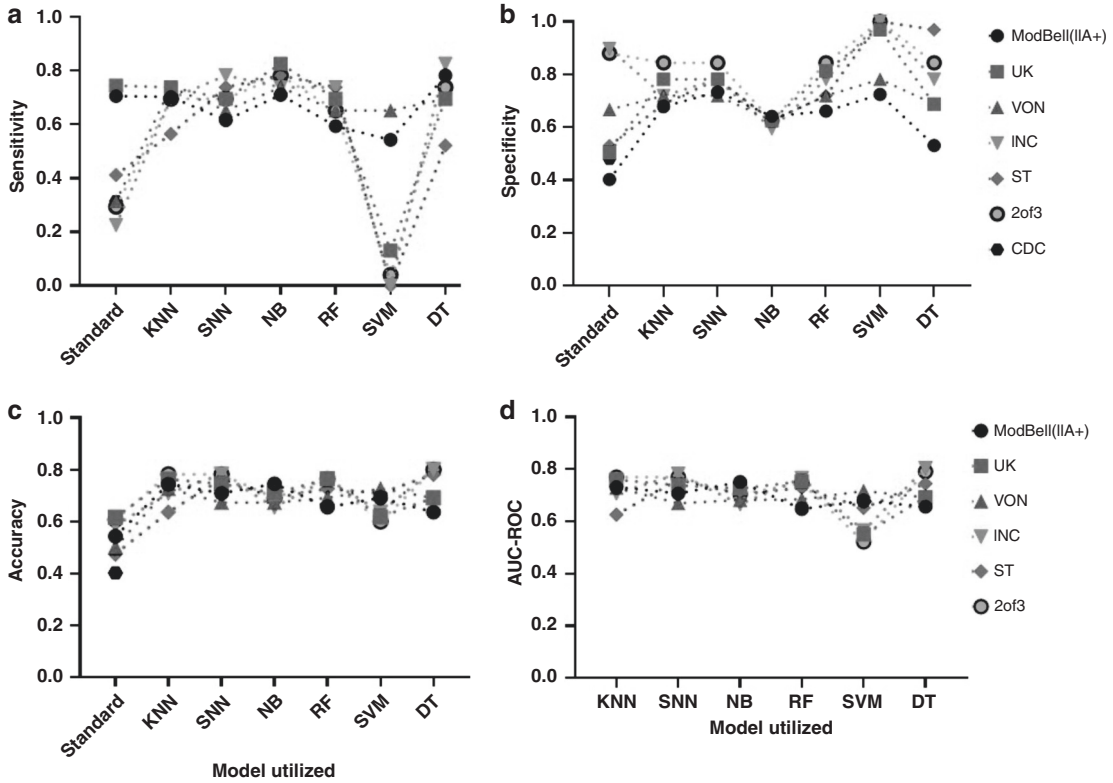


Fig. 3 Graphical depiction of sensitivity, specificity, accuracy, and AUROC of NEC definitions using standard statistical methods and machine learning classifiers. a Sensitivity, **b** specificity, **c** accuracy, and **d** area under the receiver operator curve (AUROC) of various definitions of NEC including Severe Bell (ModBell(IIA+)), United Kingdom (UK), Vermont Oxford Network (VON), 2of3, Stanford (ST), and International Neonatal Consortium (INC) based on standard statistics (standard), *K* nearest neighbors (KNN), simple neural network (SNN), Naive Bayes (NB), random forest (RF), support vector machine (SVM), and decision tree (DT) modeling. Of note, the Centers for Disease Control (CDC) definition is only located in the standard statistics analysis as discussed in the "Methods" and the AUROC for standard statistics was not calculated.

scores. The average specificity score for the definitions across the remaining classifiers was 0.728, which was similar, yet higher than the average for sensitivity. The scores for specificity ranged from a low specificity in the ModBell(IIA+) definition with 0.531 in the DT classifier to a specificity score of 1.0 for 2of3, INC, and ST definitions using the SVM classifier (Fig. 3b). Of note, the 2of3 definition had consistently better specificity scores across both standard statistics and all machine learning classifiers in comparison to most other definitions ranging from 0.625 in NB to 1.0 in SVM with an average score across all classifiers and statistics of 0.84 (Fig. 3b).

When evaluating sensitivity and specificity together, both the INC and 2of3 definitions had higher performance using machine learning classifiers than most other definitions, particularly in the DT classifier (Fig. 3a, b). Additionally, the ModBell(IIA+) definition had the overall lowest sensitivity scores in three of the six machine learning classifiers, including RF, NB, and SNN, and also had the overall lowest specificity scores in four of the six machine learning classifiers, including KNN, RF, SVM, and DT (Fig. 3a, b).

Accuracy of the NEC definitions and AUROC evaluation of machine learning classifiers

The UK definition had the highest overall accuracy score (0.616) utilizing standard statistical methods followed closely by the 2of3 definition (0.607) (Fig. 3c). In comparison, the lowest performing definitions were CDC (0.402) and ST (0.475) (Fig. 3c). Similar to the sensitivity and specificity scores, higher overall accuracy scores were achieved by using machine learning classifiers with an average score across classifiers of 0.715. For the machine learning classifiers, the accuracy scores ranged from a low score by the UK definition of 0.618 in the SVM classifier to a high score of 0.800 from both the INC and 2of3 definitions in the DT classifier (Fig. 3c).

AUROC scores were only calculated for the machine learning classifiers. The average AUROC score across all definitions and classifiers was 0.705, which was the lowest average of the four metrics examined in this study (sensitivity, specificity, accuracy, and AUROC; Fig. 3d). The AUROC scores ranged from a low with the UK definition in the SVM classifier of 0.550 to a high with the INC definition of 0.804 in the DT classifier (Fig. 3d).

Consistent with the sensitivity and specificity scores, the INC and 2of3 definitions had consistently higher scores compared to almost all the other classifiers in both accuracy and AUROC (Fig. 3c, d). Also, in both accuracy and AUROC, ModBell(IIA+) had the lowest overall scores for both the RF and DT classifiers.

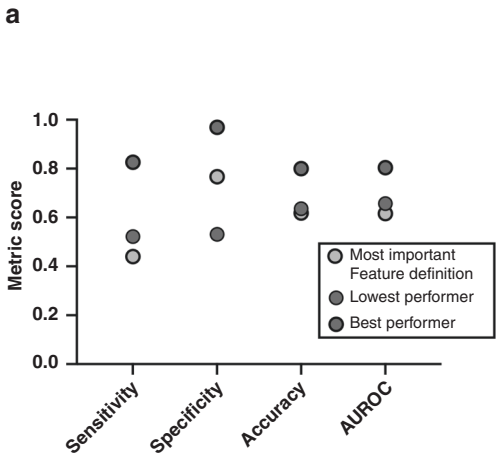
Feature importance evaluation

Using the decision tree classifier, the most important features for each of the 14 current NEC definitions were identified. These features as well as their feature importance scores are shown in Table 2. The most frequently used of the most important features was pneumatosis, which was identified as important for 11 of the 14 definitions (Table 2). Additionally, pneumatosis contained the highest feature importance scores (0.801 for VON and 0.851 for INC; Table 2). When all 41 available features were combined together as a separate definition designated as "All features" and analyzed for feature importance, 3 features were identified as important: volume of feeding at NEC onset, disseminated intravascular coagulation (defined as prothrombin time ≥15 s or fibrinogen ≤100), and occult rectal bleeding (Table 2). Interestingly, these three features had not previously been identified as

Table 2. Most important features identified from the decision tree classifier for each NEC definition.

Feature identified as important	NEC definition identified as important for
Apnea	Bell Stage 1 (0.368), Bell Stage 1A and 1B (0.414)
Lethargy	Bell Stage 1 (0.335), Bell Stage 2, 3, 2A, 2B, 3A, and 3B (0.245), Bell stage 2A and 2B (0.357)
Guaiac-positive GI bleed	Bell Stage 1 (0.196), Bell Stage 1A and 1B (0.208)
Pneumatosis	Bell Stage 2 and 3 (0.725), Bell Stage 2A, 2B, 3A, and 3B (0.713), UK (0.618), VON (0.801), 2of3 (0.458) ST (0.542), INC (0.851)
Gestational age	UK (0.179), 2of3 (0.155)
Postnatal age at NEC onset	2of3 (0.270), ST (0.355), INC (0.149)
Volume of feeding at NEC onset	All features (0.654)
Disseminated intravascular coagulation (DIC)	All features (0.128)
Occult rectal bleeding	All features (0.136)

Parentheses denote the importance score for each of the features within the definition they were identified as important. Of note, the "All Features" definition includes all the features that were utilized by the various definitions.



b

Feature identified as important	Feature importance score
Apnea	0
Lethargy	0.107
Guaiac-positive GI bleed	0
Pneumatosis	0.082
Gestational age	0.142
Postnatal age at NEC onset	0.049
Volume of feeding at NEC onset	0.406
DIC	0.078
Occult rectal bleeding	0.136

Fig. 4 Graphical depiction of sensitivity, specificity, accuracy, and AUROC of the most important feature definition and its relative importance scores. a Sensitivity, specificity, test set accuracy, and AUROC evaluation of the most important features definition in the decision tree classifier. **b** Features identified as important from the most important features.

important in the individual 14 definitions analyzed with the DT classifier.

The nine most important features (Table 2) were combined into a new definition designated as "most important features." This new definition was unable to train a better decision tree model compared to the other definitions' DT models despite the fact that it contained only features that had been identified as important for the other models to choose between NEC and non-NEC. As seen in Fig. 4a, the most important feature definition had sensitivity (0.440), specificity (0.767), accuracy (0.618), and AUROC (0.616) scores that were all lower than the results from the best performing NEC definitions and were also lower than the lowest performing NEC definitions aside from specificity. The feature importance scores for each of the nine features in the most important features definition are listed in Fig. 4b. Range features, including GA, volume of feeding at NEC onset, and postnatal age at NEC onset, represented almost half of the feature importance for the most important feature definition (Fig. 4b). Although pneumatosis was identified as the most important feature and had the highest observed feature importance score for the NEC definitions (Table 2), in the most important feature definition, it ranked only fifth most important out of the nine features (Fig. 4b).

DISCUSSION

NEC remains a leading cause of morbidity and mortality in preterm infants.[1,4,6] However, because the pathophysiology of NEC remains incompletely understood and NEC is likely a spectrum of disease, creating a uniform definition has been challenging.[7] Many definitions for NEC have been developed[4,9,10] (Fig. 1a), but to date, no head-to-head evaluation has been performed to evaluate the capacity of the individual definitions of NEC to correctly make a diagnosis. To attempt to close this knowledge gap, we performed standard statistical measures and supervised machine learning techniques on a single-center, retrospective cohort to evaluate the sensitivity, specificity, and accuracy of each definition. Overall, based on standard statistics, the ModBell(IIA+) and UK definitions had low specificity (0.402 and 0.504, respectively) but high sensitivity (0.706 and 0.745, respectively) for classifying a clinician-based diagnosis of NEC, while more contemporary definitions (VON, INC, and 2of3) had better specificity (0.667, 0.897, 0.880 respectively) but lower sensitivity (0.314, 0.225, 0.294). Machine learning provided overall better scores compared to standard statistics. The highest machine learning scores were obtained using the DT classifier in all the four metrics analyzed, particularly for the 2of3 and INC definitions. These definitions performed consistently better than other definitions in almost all the classifiers other than the NB and SVM. When evaluating feature

importance for the decision trees, nine features were determined to be most important. However, when making a classifier using only these nine features, the results suggested that sensitivity was sacrificed for higher specificity and the test set accuracy and AUROC scores were worse than most pre-defined definitions, especially 2of3 and INC.

An important finding of this study was the high performance of the Non-Bell NEC definitions, including UK, VON, 2of3, ST, and INC. These definitions were better at distinguishing between NEC and non-NEC than ModBell(IIA+) in most machine learning models (Fig. 3). This suggests that further evaluation needs to be done on the newer definitions to elucidate what feature(s) allow improved diagnostic performance compared to classical Bell staging. The improved performance of the Non-Bell definitions was interesting as the Bell and Modified Bell definitions constituting the ModBell (IIA+) definition used between 13 and 26 features to make a diagnosis, while Non-Bell definitions only used between 7 and 11 features.[21] The higher performance with fewer features suggests that more limited measures for a NEC definition may be more clinically useful, may be more informative of disease presence, and may have greater performance in identifying infants with a clinician's diagnosis of NEC.

Of the six machine learning classifiers utilized in this study, DT had the highest overall scores in sensitivity, specificity, accuracy, and AUROC compared to the five other classifiers examined in this study. This meant that the DT classifier was best able to fit the data and may suggest that scaling is an issue in our dataset. One of the strengths of the decision tree classifier is that it is unaffected by improper scaling.[21] Three of our features, including GA (range in weeks = 22.1–40.5), postnatal age at NEC onset (range in days = 0–120), and volume of feeding at NEC onset (range in milliliters = 0–179), had input values that covered a range of data. These three features were not scaled or normalized during the preprocessing step, which resulted in significantly higher input values than the zero/one that was utilized for the yes/no features. Based on the DT classifier's superior performance, future datasets would likely benefit from scaling or normalizing some of these range features, so that the classifier does not give more consideration to one feature over the other.

Finally, evaluation of feature importance in the decision tree classifier highlighted features that were important to distinguish clinical NEC diagnosis versus not.[21,22] Although a subset of nine features were identified, when combined, they were unable to train a better decision tree than combinations of criteria from pre-established definitions.[5,21] This may suggest that interactions between features are important and feature engineering may help machine learning models perform better on the data. Ultimately, this further emphasizes the need for additional evaluation of the features available for NEC diagnosis.[5,21]

Our study had several limitations. While NEC diagnosis was defined by physical findings; laboratory evidence of inflammation; radiographic findings; and when available, pathology findings, the absolute presence of NEC was still determined by a single investigator. Our study was also a retrospective study and thus not all charted diagnoses and evaluations were available for each patient. This resulted in the need to find "best fit" definitions when doing the statistical analysis instead of individually assessing each patient with each NEC definition as a whole. Further, this resulted in having to do the statistical analysis in two separate sets with one focusing on the Bell definitions and one focusing on all the contemporary definitions to maintain adequate feature numbers. The limited evaluation for some patients also impacted the machine learning datasets, since machine learning cannot handle missing data points and must have representation for all patients for all criteria utilized. To address this, features with excessive missing information (over 3/4ths of the data) or criteria that were over or under diverse to be meaningful were eliminated from the machine learning analysis. Further, any missing data for features that were kept had to be filled in as was described above in the "Methods." These accommodations may partially explain the difference in results between the machine learning models and the standard statistical approach.

Furthermore, all patients came from one hospital and there was not a true "healthy" cohort as all patients either were diagnosed with NEC or had NEC concern; therefore, the generalizability to other hospitals or infants without some form of gastrointestinal pathology needs to be analyzed further. Infants diagnosed with NEC were determined based on information at the time of NEC onset and not throughout the course; therefore, the utility of these different definitions to predict onset of NEC or throughout the dynamic course of NEC needs additional consideration. The cohort for our study included all infants with an ICD-based diagnosis of "NEC" and/or "intestinal perforation." Some infants who were determined to not have NEC were determined to have SIP (30/117). The patients with SIP were not excluded from the study due to the already minimal patient numbers available. This represents a limitation for machine learning because the training and test sets were randomly split, and depending on the number of SIP infants in the training set for "No NEC," it may have slightly skewed the model.

Finally, although some of the Non-Bell NEC definitions have exclusion criteria, due to the nature of machine learning we were unable to apply these features as true exclusions. Instead, exclusion criteria of definitions such as enteral feeding <80 ml/kg/day and GA ≥36 weeks were applied as traditional features in our machine learning models and this again may partly explain the machine learning model's better performance for Non-Bell definitions using these criteria compared to standard statistics. Feature importance analysis showed that not much weight was given to these two criteria, but future analysis may benefit from further examining the role of these exclusion criteria in diagnosis.

In conclusion, standard statistics and six different supervised machine learning classifiers were used to evaluate current definitions of NEC. Results suggested that in statistics the ModBell(IIA+) and UK definitions had high sensitivity but low specificity and VON, INC, and 2of3 definitions had low sensitivity but high specificity. The Non-Bell definitions tended to perform better in all four evaluation metrics in most supervised machine learning classifiers. The superior performance of the Non-Bell definitions, which utilized fewer features than Bell staging, suggests that further evaluation of these newer definitions is critical in the quest to find an optimal definition. Additionally, for many definitions there was a trade-off between obtaining high sensitivity but low specificity or obtaining low sensitivity but high specificity. The potential optimal diagnostic criteria may contain a consensus from one or more definitions with high sensitivity in combination with one or more definitions with high specificity. Of the six classifiers utilized in this study, the DT classifier performed the best with scores in all metrics of around 0.8 for the 2of3 and INC definitions, although evaluation scores may be improved for the other classifiers and definitions in future analyses by scaling the data, adding more informative features, feature engineering, and further optimization of parameters. Future work may also include graphical visualization and clustering analysis to help in evaluation of how best to fit the data and aid in identifying important features. Additionally, evaluation of features that are not currently applied in the definitions but may be relevant for diagnosis have the potential to enhance the definitions and lead to better ability to diagnose. As an example, it has been recognized in the field that formula feeding leaves an infant at greater risk for NEC than breast feeding, but type of feeding is not currently utilized as a feature in any definition for NEC.[1,4,21] Also, it has been well established that microbiome plays a significant role in NEC with dysbiosis, often presenting as a bloom in Enterobacteriaceae, commonly observed shortly before NEC onset.[1,4,5,24] Adding a microbiome component to the evaluation

for NEC may also provide for a more informed definition. Most importantly, external validation and evaluation of performance of definitions in a larger and more diverse cohort of infants will be crucial to determine the utility of the various NEC definitions for classification of the disease. Ultimately, the goal is to more effectively diagnose NEC to provide a more accurate prognosis and guide additional diagnostic evaluation and treatment for patients at risk of this devastating disease.

ACKNOWLEDGEMENTS

We would first like to acknowledge Dr. Thomas Casavant from the University of Iowa Center for Bioinformatics and Computational Biology (CBCB) for critical advice regarding the machine learning classifiers. We would also like to acknowledge Dr. Kelli Ryckman from the University of Iowa Epidemiology Department for aiding in acquiring the data utilized in this study. S.R.L., T.J.B., and S.J.M. were supported by the Stead Family Department of Pediatrics, University of Iowa.

AUTHOR CONTRIBUTIONS

S.R.L., T.J.B., and S.J.M. all contributed significantly to conception, design, analysis, and interpretation of the data as well as drafting the article and revising critically. E.J. contributed significantly to acquisition of data and critical revisions. R.M.P. contributed significantly to critical revisions.

ADDITIONAL INFORMATION

Competing interests: R.M.P. and S.J.M. are scientific advisors to the NEC Society.

Consent statement: Patient consent was not required.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Claud, E. C. & Walker, W. A. Hypothesis: inappropriate colonization of the premature intestine can cause neonatal necrotizing enterocolitis. *FASEB J.* **15**, 1398–1403 (2001).
2. Gordon, P. V., Christensen, R., Weitkamp, J.-H. & Maheshwari, A. Mapping the new world of necrotizing enterocolitis (NEC): review and opinion. *EJ Neonatol. Res.* **2**, 145–172 (2012).
3. Mizrahi, A. et al. Necrotizing enterocolitis in premature infants. *J. Pediatr.* **66**, 697–706 (1965).
4. Neu, J. & Walker, W. A. Necrotizing enterocolitis. *N. Engl. J. Med.* **364**, 255–264 (2011).
5. Tanner, S. M. et al. Pathogenesis of necrotizing enterocolitis. *Am. J. Pathol.* **185**, 4–16 (2015).
6. Wejryd, E. et al. Low diversity of human milk oligosaccharides is associated with necrotising enterocolitis in extremely low birth weight infants. *Nutrients* **10**, 1556 (2018).
7. Patel, R. M. et al. Defining necrotizing enterocolitis: current difficulties and future opportunities. *Pediatr. Res.* **88**, 10–15 (2020).
8. D'Angelo, G. et al. Current status of laboratory and imaging diagnosis of neonatal necrotizing enterocolitis. *Ital. J. Pediatr.* **44**, 84 (2018).
9. Bell, M. J. et al. Neonatal necrotizing enterocolitis: therapeutic decisions based upon clinical staging. *Ann. Surg.* **187**, 1–7 (1978).
10. Walsh, M. C. & Kliegman, R. M. Necrotizing enterocolitis: treatment based on staging criteria. *Pediatr. Clin. N. Am.* **33**, 179–201 (1986).
11. Gordon, P. V., Swanson, J. R., Attridge, J. T. & Clark, R. Emerging trends in acquired neonatal intestinal disease: is it time to abandon Bell's criteria? *J. Perinatol.* **27**, 661–671 (2007).
12. Battersby, C. et al. Development of a gestational age–specific case definition for neonatal necrotizing enterocolitis. *JAMA Pediatr.* **171**, 256 (2017).
13. Vermont Oxford Network. Vermont Oxford Network manual of operations: part 2 data definitions and infant data forms. https://vtoxford.zendesk.com/hc/en-us/articles/360013115393-2019-Manual-of-Operations-Part-2-Release-23-2-PDF (2019).
14. Gephart, S. M. et al. Changing the paradigm of defining, detecting, and diagnosing NEC: perspectives on Bell's stages and biomarkers for NEC. *Semin. Pediatr. Surg.* **27**, 3–10 (2018).
15. Ji, J. et al. A data-driven algorithm integrating clinical and laboratory features for the diagnosis and prognosis of necrotizing enterocolitis. *PLoS ONE* **9**, e89860 (2014).
16. Caplan, M. S. et al. Necrotizing enterocolitis: using regulatory science and drug development to improve outcomes. *J. Pediatr.* **212**, 208.e1–215.e1 (2019).
17. Centers for Disease Control. CDC/NHSN Surveillance definitions for specific types of infections. https://www.cdc.gov/nhsn/pdfs/pscmanual/17pscnosinfdef_current.pdf (2021).
18. Lure, A. C. et al. Using machine learning analysis to assist in differentiating between necrotizing enterocolitis and spontaneous intestinal perforation: a novel predictive analytic tool. *J. Pediatr. Surg.* https://doi.org/10.1016/j.jpedsurg.2020.11.008 (2020).
19. Irles, C. et al. Estimation of neonatal intestinal perforation associated with necrotizing enterocolitis by machine learning reveals new key factors. *Int. J. Environ. Res. Public Health* **15**, 2509 (2018).
20. Kluyver, T. et al. In *Positioning and Power in Academic Publishing: Players, Agents, and Agendas* (eds Loizides, F. & Scmidt, B.) 87–90 (IOS Press, 2016).
21. Müller, A. C. & Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists* (O'Reilly Media Inc., 2017).
22. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
23. Byrd, R. H., Lu, P. & Nocedal, J. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16**, 1190–1208 (1995).
24. Elgin, T. G., Kern, S. L. & McElroy, S. J. Development of the neonatal intestinal microbiome and its association with necrotizing enterocolitis. *Clin. Ther.* **38**, 706–715 (2016)., corrected publication 2021