Journal of Pediatric Surgery 54 (2019) 799–804

Contents lists available at ScienceDirect

Journal of Pediatric Surgery

journal homepage: www.elsevier.com/locate/jpedsurg

Statistical evaluation of diagnostic tests: A primer for pediatric surgeons☆,☆☆

Steven J. Staffa *, David Zurakowski

Department of Surgery, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 2 March 2018
Received in revised form 2 May 2018
Accepted 4 June 2018

Key words:
Biostatistics
Diagnostic tests
ROC analysis
Sensitivity
Specificity

ABSTRACT

Background/Purpose: Diagnostic tests are of paramount importance for informing decision making in the surgical setting. Certain statistical methods are necessary to properly analyze data for diagnostic or prognostic tests involving biomarkers and risk factor data. Our goal is to provide a useful primer for the surgical researcher when performing a diagnostic research study in order to best analyze their data.
Methods: We present the key concepts and statistics for diagnostic tests and receiver operating characteristic (ROC) curve analysis, and we illustrate each with hypothetical surgery research examples. We use hypothetical data regarding CT imaging and WBC count in their diagnostic ability in predicting acute appendicitis, an extremely common surgical condition, while reviewing the statistical concepts of sensitivity, specificity, positive and negative predictive value, positive and negative likelihood ratio, relative risk, odds ratio, and ROC curves. Then we will consider a hypothetical a risk factor analysis on 30-day readmission to illustrate how multiple predictors can be combined.
Conclusions: The statistical concepts presented are useful to the pediatric surgeon researcher in assessing the ability of diagnostic tests, which will translate into decision making and patient management implications in the clinical setting.
Type of Study: Review Article
Level of Evidence: N/A

Contents

Diagnostic tests are of paramount importance for decision making in surgical setting. Most pediatric surgeons are familiar with the basic concepts of specificity and sensitivity and use them to determine the usefulness of specific diagnostic tests in clinical practice. On the other hand,

☆ Declarations of interest: None.
☆☆ Financial Disclosures: None.
* Corresponding author at: Department of Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115. Tel.: +1 617 355 8254.
E-mail address: Steven.Staffa@childrens.harvard.edu (S.J. Staffa).

https://doi.org/10.1016/j.jpedsurg.2018.06.010
0022-3468/© 2018 Elsevier Inc. All rights reserved.

the more clinically applicable concepts of positive and negative predictive values are underutilized. In addition, many diagnostic tests used in pediatric surgery are assigned a grade rather than a simple positive or negative result. To measure the sensitivity and specificity of a graded test, one must first select a threshold grade above which the result will be considered positive and below which it will be considered negative. Such thresholds need not be arbitrary, but rather the optimum cutoff point can be determined by choosing a cutoff that optimizes the combination of sensitivity and specificity. The ability of a graded test to discriminate between two possible results can be evaluated using a receiver operating characteristic (ROC) curve.

The objective of this paper is to review the statistics associated with the evaluation of diagnostic tests and Receiver Operating Characteristic (ROC) Curve analysis in a univariate and a multivariable setting. We will demonstrate these concepts with hypothetical data regarding Computed Tomography (CT) imaging and white blood cell (WBC) counts in patients with suspected acute appendicitis, an extremely common condition requiring surgical intervention. The gold standard for the determination of acute appendicitis is the surgical finding. We will then expand upon the principles of diagnostic tests to a multivariable framework using multivariable logistic regression and multivariable ROC analysis. A glossary of statistical terminology of the statistics for diagnostic tests can be found in Appendix A.

1. Dichotomous diagnostic tests

The simplest diagnostic test is one where the results are used to classify patients into two groups. Such tests with only two possible outcomes are known as dichotomous tests. As an example, we may which to study the diagnostic accuracy of CT imaging in detecting acute appendicitis. Table 1 shows the CT imaging and surgical finding data for this hypothetical example. Note that the cells are labeled with the traditional “a”, “b”, “c”, and “d” in order to generalize the formulas in this article. The question that arises in the clinical setting is “How good is CT at distinguishing patients with and without acute appendicitis?” In other words, “To what degree can I rely on the interpretation of CT imaging in making judgements about whether or not a pediatric patient truly has acute appendicitis?”

1.1. Sensitivity and specificity

One method of measuring the value of CT imaging in detecting acute appendicitis is to calculate the proportion of patients with and without acute appendicitis that were correctly classified CT imaging. These proportions are known as the sensitivity and specificity of a test, respectively.

Sensitivity is calculated as the proportion of patients with acute appendicitis that were correctly classified by CT imaging. In other words, this is the probability that a patient with the disease or condition tested positive for the condition [1]. In this example, of the 277 patients with acute appendicitis, 261 tested positive upon CT imaging. The sensitivity of CT imaging in the detection of acute appendicitis is therefore 94% (sensitivity = a/(a + c) = 261/277 = 0.94). In other words, 94% of patients with acute appendicitis were correctly classified using CT imaging prior to surgery.

Table 1
Hypothetical example

CT Imaging	Surgical Finding		Total
	Acute Appendicitis (+)	No Acute Appendicitis (−)	
Test Positive (+)	261 (a)	5 (b)	266
Test Negative (−)	16 (c)	88 (d)	104
Total	277	93	370

Specificity is calculated as the proportion of patients without acute appendicitis that were correctly classified by CT imaging. This is the probability that a patient who is disease-free had a test result negative finding [1]. Of the 93 patients truly without acute appendicitis, 88 were correctly classified. The specificity of CT imaging in the evaluation of the presence or absence of acute appendicitis is therefore 95% (specificity = d/(b + d) = 88/93 = 0.95). This means that 95% of patients without acute appendicitis were correctly classified as disease-free by CT imaging.

It is desirable to provide an estimate of the precision of the computed sensitivity and specificity. The 95% confidence interval (CI) is the most common tool for reporting the precision around an estimate. Although other methods have been described [2], the modified Wald method is an easy to use method and valid option for constructing 95% CIs for proportions such as sensitivity and specificity [3]. The 95% confidence interval is constructed as follows:

$$95\%\mathrm{CI} = p \pm 1.96\sqrt{\frac{2p(1-p)}{\mathrm{n}+4}}$$

In this formula, p denotes the observed proportion, and n denotes the marginal total (in other word the denominator) for the sensitivity or the specificity. With a larger sample size, the sensitivity and specificity observed will be more precise, and the corresponding 95% confidence intervals will be narrower. Since a proportion must lie between 0 and 1, note that the upper confidence limits must be trimmed to 1 if it is calculated to be larger than 1, and note that the lower confidence limit must be set to 0 if it is calculated to be less than 0. Appendix B is an Excel function that computes the 95% CI for a proportion using the modified Wald method. The surgeon can use this Excel program to easily insert their data and obtain the estimated 95% CI.

Using this approach, we can compute the 95% CI for our sensitivity of CT in the detection of acute appendicitis to be (0.90, 0.98). Similarly, the 95% CI for the specificity of CT is (0.88, 1.0). Looking at the 95% confidence intervals for sensitivity and specificity, we see that our estimate of sensitivity is more precise (95% CI: 0.90, 0.98) as compared to our estimate of specificity (95% CI: 0.88, 1.0), with less uncertainty, as demonstrated by a narrower confidence interval.

1.2. Positive and negative predictive values

Sensitivity and specificity tell us the rate of true positive and true negative results, but they do not measure how well CT imaging predicts acute appendicitis. Therefore we wish to obtain an estimate of the probability that the CT imaging will give the correct result. Such estimates are provided by the positive and negative predictive values.

Positive predictive value (PPV) is the probability that a patient who is found to have acute appendicitis upon CT imaging actually has acute appendicitis. Negative predictive value (NPV) is the probability that a patient who is found to not have acute appendicitis upon CT imaging truly does not have acute appendicitis [4]. Calculations of the PPV and NPV require an estimate of the prevalence acute appendicitis in the clinical population [5]. This reflects the fact that, the rarer that the condition is, the more confident we can be that a negative finding on CT imaging correctly indicates absence of acute appendicitis and the less confident that a positive finding on CT indicates presence of acute appendicitis. If the prevalence of the condition is low, the PPV will be low even if both the sensitivity and specificity are high. In other words, there is a greater potential for false positive tests when evaluating rare conditions, even when the rate of true positives and true negatives is high.

Unlike sensitivity and specificity, PPV and NPV are complex functions rather than simple proportions [6]. Bayes theorem is the most valid approach for determining predictive values [7]. These two formulas are stated as:

$$\mathrm{PPV} = \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1-\text{specificity})(1-\text{prevalence})}$$

$$\mathrm{NPV} = \frac{(\text{specificity})(1-\text{prevalence})}{(\text{specificity})(1-\text{prevalence}) + (1-\text{sensitivity})(\text{prevalence})}$$

If we assume a prevalence of 85% for acute appendicitis in this population of patients suspected to have acute appendicitis undergoing a CT scan, then we can calculate the following PPV and NPV for CT imaging as a diagnostic test of the condition.

$$\mathrm{PPV} = \frac{(0.94)(0.85)}{(0.94)(0.85) + (1-0.95)(1-0.85)} = 0.99$$

$$\mathrm{NPV} = \frac{(0.95)(1-0.85)}{(0.95)(1-0.85) + (1-0.94)(0.85)} = 0.74$$

Among patients who tested positive on CT imaging, we expect 99% to actually have acute appendicitis, and among patients who tested negative on CT imaging, we expect 74% to be acute appendicitis-free. Our population has a high prevalence of acute appendicitis, and therefore our PPV is very large. If acute appendicitis was rare in our population, then the PPV would not be as large and the NPV would be larger.

1.3. Likelihood ratios

Likelihood ratios (LRs) indicate the value of a diagnostic test for increasing certainty of a particular diagnosis. The positive likelihood ratio (LR+) is the likelihood of obtaining a positive test result in a patient with a particular condition divided by the probability of obtaining a positive test result in patients without the particular condition. Similarly, the negative likelihood ratio (LR−) is the likelihood of obtaining a negative test result in a patient without the particular condition [6]. In clinical situations, likelihood ratios may have advantages over other statistics of a diagnostic test because they can be interpreted in the context of an increased or decreased likelihood of having a certain surgical condition based on the test result across different levels of test results [1]. Please see below after odds ratios are introduced for further information. The positive likelihood ratio and the negative likelihood ratio can be computed by the following simple formulas:

$$\mathrm{LR}+ = \frac{\text{sensitivity}}{(1-\text{specificity})}$$

$$\mathrm{LR}- = \frac{(1-\text{sensitivity})}{\text{specificity}}$$

In our clinical hypothetical example, the LR+ for CT imaging in diagnosing acute appendicitis is (0.94) / (1 − 0.95) = 18.8, and the LR− for CT imaging is (1 − 0.94) / (0.95) = 0.06. Note that the LR is not bound by 0 and 1, but rather is any positive real value. The higher the LR+ and the farther away from 1.0, the better the test result for ruling in a particular condition or disease, whereas the smaller the LR− and again the farther away from 1.0, the better the test is for ruling out a particular condition [8]. The LRs for CT imaging show that it has strong ability to rule in and rule out acute appendicitis.

1.4. Odds ratios

The odds ratio (OR) is familiar statistic to most surgeons, and is a useful measure of association when an outcome is binary. The odds ratio is the ratio of two odds, that is, it is the odds of an outcome in one group divided by the odds of the outcome in another group. Therefore, an odds ratio of 1.0 indicates no association between the predictor or group variable and the outcome. An odds ratio greater than 1 indicates a higher odds of the outcome in the first group (in the numerator), and an odds ratio less than 1 indicates a higher odds of the outcome in the second group (in the denominator) [9].

From a 2 × 2 table such as Table 1, the odds ratio can be estimated as the ratio of the diagonal products (ad/bc). That is, the odds ratio for CT imaging in predicting acute appendicitis is as follows:

$$\mathrm{OR} = \frac{ad}{bc} = \frac{261 \times 88}{5 \times 16} = 287$$

Thus, the odds of true acute appendicitis upon surgery are 287 times greater if a patient tests positive on CT imaging as compared to if a patient tests negative on CT imaging.

The 95% confidence interval for the natural logarithm of the OR can be computed using the following formula based on a Taylor series expansion and using Table 1:

$$95\%\mathrm{CI} = \ln(\mathrm{OR}) \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

From here, we can obtain the 95% confidence interval for the OR by exponentiating the lower and upper confidence limits for the natural logarithm of the OR (i.e. raising "e" to the power of each). With this formula, we calculate the 95% confidence interval for the natural logarithm of the OR to be $\ln(287) \pm 1.96\sqrt{\frac{1}{261} + \frac{1}{5} + \frac{1}{16} + \frac{1}{88}} =$ (4.6, 6.7). Thus the 95% confidence interval for the OR of 287 is $(e^{4.6}, e^{6.7}) = (99, 812)$.

Furthermore, the likelihood ratio can be used along with the odds to obtain posttest odds, and subsequently the posttest probability using the following formulas [1].

$$\text{Posttest odds} = \text{Pretest odds} \times \text{Likelihood ratio}$$

$$\text{Posttest probability} = \frac{\text{Posttest odds}}{1-\text{Posttest odds}}$$

1.5. Relative risks (risk ratios)

The relative risk, as computed by the risk ratio (RR), is measured as the conditional probability of the outcome among the patients in one exposure group divided by the conditional probability of the outcome among the patients in the other exposure group. The relative risk is an alternative to the odds ratio, although it has a different interpretation. The risk ratio may be more useful with prospective data, and when an outcome is rare in the study population, the odds ratio can be used estimate the risk ratio [10]. In our example, the outcome of acute appendicitis occurs in 277 out of 370 patients (75%). Therefore the outcome is not rare and the risk ratio cannot be estimated by the odds ratio. In Table 1, the risk ratio can be estimated by the following formula:

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{261/266}{16/104} = \frac{0.98}{0.15} = 6.5$$

The numerator of the risk ratio is the risk of a surgical finding of acute appendicitis conditional on a positive test result on CT imaging (98%), and the denominator is the risk of acute appendicitis given a negative test result (15%). The relative risk of 6.5 suggests that the risk of a surgical finding of acute appendicitis in the group of patients with a positive test result on CT imaging is 6.5 times higher than the risk of acute appendicitis in the group of patients with a negative test result.

The 95% confidence interval for the risk ratio may provide a sense of the precision of the estimate of relative risk, and can be obtained by exponentiation the lower and upper confidence limits for the natural logarithm of the risk ratio [11]. Using Table 1, the formula to obtain the confidence interval for the natural logarithm of the risk ratio is the following:

$$95\%\ \mathrm{CI} = \ln(\mathrm{RR}) \pm 1.96\sqrt{\frac{b/a}{(a+b)}+\frac{d/c}{(c+d)}}$$

With this formula, we calculate the 95% confidence interval for the natural logarithm of the risk ratio to be $\ln(6.5) \pm 0.451 = (1.4, 2.3)$. Thus the 95% confidence interval for the RR of 6.5 is $(e^{1.4}, e^{2.3}) = (4.1, 10.2)$.

It should be noted that the evaluation of diagnostic tests requires a sufficient sample size to include patients who fall into the categories of true positives, true negatives, false positives and false negatives. In our example of acute appendicitis, if we had a very small number of patients with this outcome, we would not be able to accurately estimate sensitivity or other statistics such as PPV and NPV. Furthermore, in calculating odds ratios and risk ratios it is necessary that the cells of the 2 × 2 table not contain any zeros.

2. Graded tests

In pediatric surgery, many diagnostic tests do not give a clear-cut positive or negative result. Many tests yield a grade or a score and the surgeon must then decide upon a cutoff value to classify the result as indicating either the presence or absence of a condition. As a continuation of our first hypothetical example, suppose all 370 patients had preoperative lab testing done revealing their white blood cell (WBC) count (10^3 cells per mm^3). A plot of all individual data points is shown in Fig. 1. Each patient can be categorized into 5 ordinal categories of WBC count. The hypothetical distribution of WBC counts is shown in Table 2. We wish to assess the diagnostic ability of this continuous biomarker (that we will treat at ordinal), WBC count, in determining acute appendicitis. In this example we are considering a continuous predictor variable (WBC count). It is not essential that WBC count follows a normal distribution because we are using it to create an ordinal variable. However, if a continuous variable is highly skewed by the Shapiro–Wilk test or the Kolmogorov–Smirnov test, a log-transformation may be useful to normalize the variable prior to analysis.

Table 2
White blood cell biomarker data

Category	WBC Count (10^3 cells per mm^3)	Surgical Finding	
		Acute Appendicitis (+)	No Acute Appendicitis (−)
1	0–4	9	29
2	4–8	17	37
3	8–12	21	18
4	12–16	76	7
5	16–20	110	2
6	> 20	44	0
Total		277	93

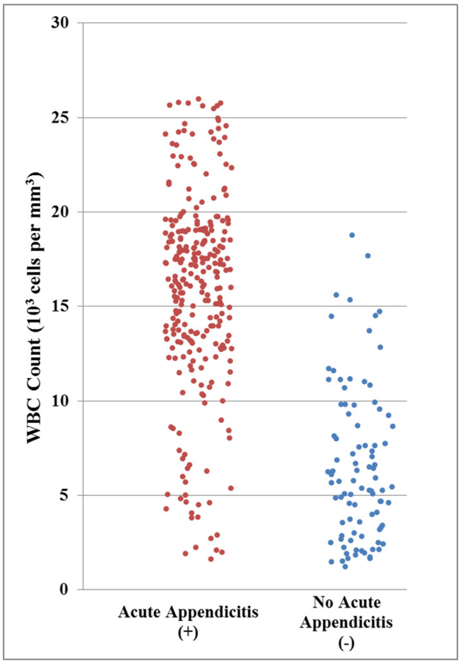

Fig. 1. Plot of all data points for WBC count (10^3 cells per mm^3) by presence (red) or absence (blue) of surgical finding of acute appendicitis. Note that the WBC counts among patients found to have acute appendicitis are generally higher than those without acute appendicitis. In order to determine the diagnostic ability of WBC count in predicting acute appendicitis ROC analysis is required.

If the surgeon considered all WBC category 2 or higher (WBC count > 4×10^3 per mm^3) as a test positive for acute appendicitis, the sensitivity would be 269/277 = 97% and the specificity would be 29/93 = 31%. If the cutoff value were raised to category 3 (WBC count > 8×10^3 per mm^3), the sensitivity and specificity would be 91% and 71%, respectively. Likewise, a cutoff of category 4 (WBC count > 12×10^3 per mm^3) would result in a sensitivity of 83% and a specificity of 90%. Finally, a cutoff of category 5 to determine a test positive (WBC count > 16×10^3 per mm^3) would result in a sensitivity and specificity of 56% and 98%, respectively. These results as well as the PPV, NPV, LR+ and LR− for each operating point are summarized in Table 3. Note that as the threshold changes, sensitivity and specificity respond in opposite directions from each other. While taking into consideration the relative consequences of false negative and false positive results, the surgeon must choose a cutoff that produces an acceptable combination of sensitivity and specificity. A test with good diagnostic performance is one that has a cutoff value at which both sensitivity and specificity are reasonably high. The way that the surgeon can go about determining this optimal cutoff is by a Receiver Operating Characteristic (ROC) analysis.

2.1. Receiver operating characteristic curves

The ability of a graded test to discriminate between results (patients with vs. without acute appendectomy) can be measured by a receiver operating characteristic (ROC) curve. An ROC graph shows the relationship between sensitivity (y-axis) and 100 − specificity (x-axis) plotted at each possible cutoff (called "operating points") [12]. In other words, the ROC curve describes the test's performance as the relationship between the true-positive rate and the false-positive rate. If a test discriminates well, its ROC curve rapidly approaches a true-positive rate or sensitivity of 100%. On the other hand, a test that discriminates poorly has a diagonal ROC curve. Diagnostic performance is evaluated by the

Table 3
Diagnostic test statistics for each operating point

Cutoff Category for Testing Positive	WBC Count Cutoff (10^3 cells per mm^3)	Sensitivity	Specificity	PPV[a]	NPV[a]	LR+	LR−
1	≥ 0	1.00	0.00	0.85	.	1.0	.
2	≥ 4	0.97	0.31	0.89	0.63	1.4	0.10
3	≥ 8	0.91	0.71	0.95	0.57	3.1	0.13
4	≥ 12	0.83	0.90	0.98	0.48	8.6	0.19
5	≥ 16	0.56	0.98	0.99	0.28	25.9	0.45
6	≥ 20	0.16	1.00	1.00	0.17	.	0.84

WBC = White Blood Cell; PPV = Positive Predictive Value; NPV = Negative Predictive Value; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio.

[a] PPV and NPV assume a prevalence of acute appendicitis of 85% in this population of patients with suspected acute appendicitis.

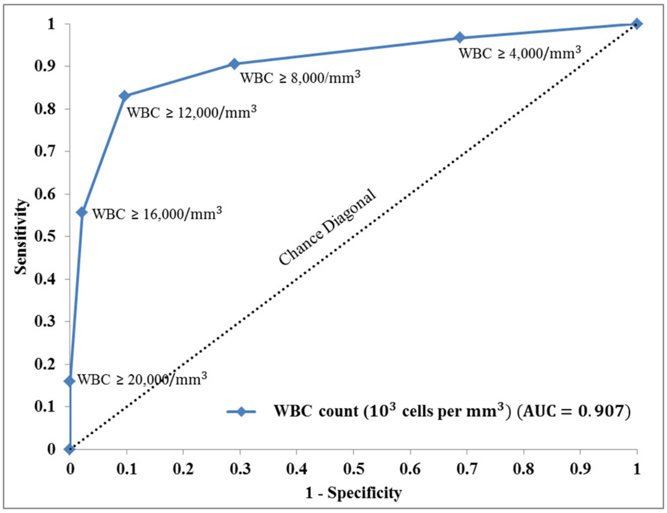

Fig. 2. ROC Curve for WBC count (10^3 cells per mm^3) in discriminating acute appendicitis. The ROC curve displays the trade-off between sensitivity and specificity for each possible threshold of WBC count. The AUC of 0.907 indicates very good discriminatory ability, and the optimal cutoff of WBC count ≥ 12,000 cells per mm^3 has a sensitivity of 83% and a specificity of 90%. The chance diagonal represents no discriminatory ability (equivalent to the flip of a coin).

Table 4
Individual risk factors of 30-day readmission

Neonatal Status	30-Day Readmission (+)	No 30-Day Readmission (−)	Total
Neonate (+)	81	31	112
Older than Neonate (−)	24	118	142
Total	105	149	254
History of Previous Surgery	30-Day Readmission (+)	No 30-Day Readmission (−)	Total
Previous Surgery (+)	66	19	85
No Previous Surgery (−)	39	130	169
Total	105	149	254
Oxygen Support	30-Day Readmission (+)	No 30-Day Readmission (−)	Total
Oxygen Support (+)	55	7	62
No Oxygen Support (−)	50	142	192
Total	105	149	254

area under the ROC curve (AUC). In the case of perfect discrimination, the AUC will equal 1.0, while an area of 0.5 indicates discrimination equivalent to a coin toss or random guessing. An AUC of 0.600 is usually deemed fair, 0.700 is deemed good, 0.800 is deemed very good, and 0.900 is deemed excellent. Area under the ROC curve can be determined by commercially available software packages and is estimated using trapezoidal approximations. The hypothetical data on patients' WBC counts generate a steep ROC curve (Fig. 2). The area under the curve is 0.907, indicating that WBC count is very good at discriminating between patients with and without acute appendicitis. Note that each point on the curve corresponds to a different cutoff for considering a WBC count as test positive (operating points).

3. Multivariable ROC analysis

Thus far we have considered the discriminatory ability of one variable in predicting a binary outcome. We illustrated the case where this diagnostic test (predictor) was binary such at CT imaging test results, and we also considered how to proceed in the situation of a continuous biomarker using ROC analysis. However, there may be more than one variable in the data that can discriminate an outcome. In this case, multivariable logistic regression is invaluable in order to multiplex predictors together to perform an ROC analysis. Multivariable modeling is useful for the assessment of the independent contributions of multiple predictors or covariates, as well as the evaluation of the associations between a predictor variable of interest and the binary outcome, while controlling for potential confounding.

For this analysis we will consider a second hypothetical example of a clinical risk model [13]. Suppose we wish to multiplex or combine risk factors of 30-day readmission following surgery. The risk factors that we will consider are neonatal status (age ≤ 28 days) at time of surgery, history of previous surgery, and oxygen support at time of surgery. Our hypothetical data set has no missing values for all variables and includes 254 patients, 105 of which were readmitted within 30-days.

A 2 × 2 table can be created for each of these risk factors with 30-day readmission following surgery. These tables are shown in Table 4. Each factor on its own has good sensitivity and specificity for predicting 30-day readmission. The sensitivity and specificity of neonatal status for predicting 30-day readmission are 81/105 = 78% and 118/149 = 79%, respectively. For history of previous surgery, the sensitivity is 66/105 = 63% and the specificity is 130/149 = 87%. Finally, for oxygen support at time of surgery, the sensitivity is 55/105 = 52% and the specificity is 142/149 = 95% in predicting cases with and without readmission. While each factor is useful on its own, we wish to use all three factors in predicting 30-day readmission following surgery. To do so, we must use multivariable logistic regression [14].

We can fit a multivariable logistic regression model with 30-day readmission as the dichotomous outcome, and neonatal status, history of previous surgery, and oxygen support at the time of surgery as predictors. This model will produce a fitted probability of readmission following surgery for each patient. Patients with identical values for their risk factors (i.e. patients with the same "covariate pattern") will have the same predicted probability of 30-day readmission. Since we have 3 risk factors and all are dichotomous, there are a total of 8 covariate patterns. A flow diagram for the probability of 30-day readmission is a useful way to display the risk for a given patient (Fig. 3).

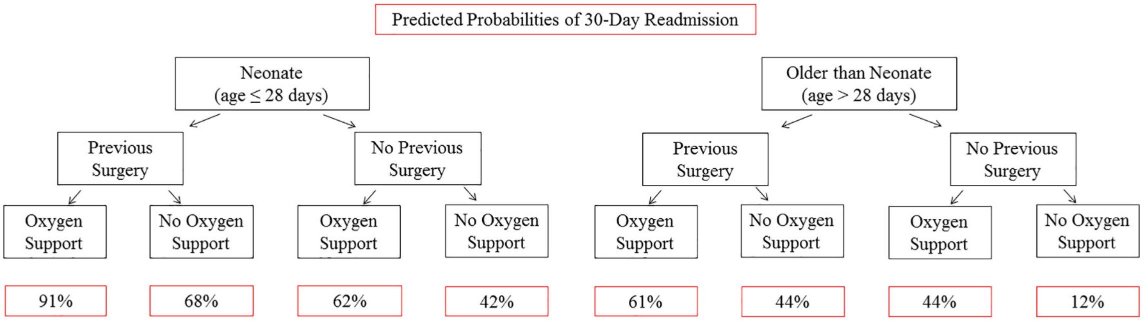

Fig. 3. Flow diagram of risk of 30-day readmission based on the risk factors of neonatal status (age ≤ 28 days), history of previous surgery and oxygen support at time of surgery. The predicted risks of 30-day readmission are can be obtained using multivariable logistic regression modeling. The predicted risks are displayed in decreasing order by neonatal age category and within each stratum.

4. Conclusions

The fundamental concepts of diagnostics tests and ROC analysis are widely applicable to pediatric surgical research [15–18]. The statistical concepts of sensitivity, specificity, positive and negative predictive values, positive and negative likelihood ratios, odds ratios, ROC curves, and AUC are applicable to many types of research studies, including diagnostic testing studies, biomarker studies, and risk factor analyses. Although this is a large statistical area with numerous details and concepts, our article can be used as a practical guide for the pediatric surgeon when evaluating diagnostics tests.

Appendix A. Glossary of statistical terminology

Accuracy: the ability of a diagnostic test to correctly classify patients with and without the disease state or outcome of interest.

Area under the curve (AUC): the area under the ROC curve which indicates the discriminatory ability of an independent or test variable on a binary disease state or outcome of interest. An AUC of 0.600 is usually fair, 0.700 is good, 0.800 is very good, and 0.900 is deemed excellent.

Bayes theorem: a mathematical formula that uses the prevalence of a condition in the analysis population to correctly calculate the positive and negative predictive values.

False positive: a patient who has a positive test result but does not have the disease state or outcome of interest as determined by the gold standard.

False negative: a patient who has a negative test results but truly has the disease state or outcome of interest as determined by the gold standard.

Gold standard: the most widely accepted method of determining disease status.

Logistic regression: a statistical modeling technique for a binary outcome variable and one or more predictor variables. Can be used to obtain odds ratios, 95% confidence intervals, and P values for each predictor variable.

Negative likelihood ratio (LR−): calculated as equal to (1 − sensitivity) / specificity. Can be used to determine the ability of the test to rule-out a disease state or outcome of interest.

Negative predictive value (NPV): the probability of not having the disease state or outcome of interest conditional on a negative test result. It is correctly calculated using Bayes theorem and the prevalence of the disease state in the analysis population.

Odds ratio (OR): calculated as the odds of the outcome in one group divided the odds of the outcome in a comparison group.

Positive likelihood ratio (LR+): Calculated as equal to sensitivity / (1 − specificity). Can be used to determine the ability of the test to rule-in a disease state or outcome of interest.

Positive predictive value (PPV): the probability of having the disease state or outcome of interest conditional on a positive test result. It is correctly calculated using Bayes theorem and the prevalence of the disease state in the analysis population.

Prevalence: the proportion of individuals with a disease state or outcome of interest among all in the population being considered.

Receiver operating characteristic (ROC) curve: plots the sensitivity and 1 − specificity for all possible cutoff points (operating points) for a predictor to classify patients as test positive and a binary outcome.

Relative risk (RR): calculated as the risk (or conditional probability) of the outcome in one group divided the risk (or conditional probability) of the outcome in a comparison group.

Sensitivity: characterizes a diagnostic test's ability to correctly classify patients with the disease or outcome of interest.

Specificity: characterizes a diagnostic test's ability to correctly classify patients without the disease or outcome of interest.

True negative: a patient who has a negative test result and does not have the disease state or outcome of interest as determined by the gold standard.

True positive: a patient who has a positive test result and has the disease state or outcome of interest as determined by the gold standard.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpedsurg.2018.06.010.

References

[1] Fletcher RH, Fletcher SW. Risk: looking backward. Clinical epidemiology: the essentials. 4th ed. New York: Lippincott Williams & Wilkins; 2005; 91–104.

[2] Blyth CR. Still HA. Binomial confidence intervals. J Am Stat Assoc 1983;78:108–16.

[3] Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. 2nd ed. New York: John Wiley & Sons; 2011; 166–74.

[4] Umberger RA, Hatfield LA, Speck PM. Understanding negative predictive value of diagnostic tests used in clinical practice. Dimens Crit Care Nurs 2017;36:22–9.

[5] Parikh R, Mathai A, Parikh S, et al. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 2008;56:45–50.

[6] Zurakowski D, Johnson VM, Lee EY. Biostatistics in clinical decision making for cardiothoracic radiologists. J Thorac Imaging 2013;28:368–75.

[7] Schulman P. Bayes' theorem—a review. Cardiol Clin 1984;2:319–28.

[8] Sackett DL, Haynes RB, Guyatt GH. The interpretation of diagnostic data. Clinical epidemiology. 2nd ed. New York: Lippincott Williams & Wilkins; 1991; 69–152.

[9] Bland JM, Altman DG. The odds ratio. BMJ 2000;320:1468.

[10] Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 1998;280:1690–1.

[11] Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York, NY: John Wiley & Sons Inc; 2000.

[12] Soreide K, Korner H, Soreide JA. Diagnostic accuracy and receiver-operating characteristics curve analysis in surgical research and decision making. Ann Surg 2011; 253:27–34.

[13] Grunkemeier GL, Jin R. Receiver operating characteristic curve analysis of clinical risk models. Ann Thorac Surg 2001;72:323–6.

[14] Vittinghoff E, Glidden DV, Shiboski SC, et al. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. 2nd ed. New York: Springer; 2012.

[15] Arbuthnot M, Mooney DP. The sensitivity and negative predictive value of a pediatric cervical spine clearance algorithm that minimizes computerized tomography. J Pediatr Surg 2017;52:130–5.

[16] Bonadio W, Peloquin P, Brazg J, et al. Appendicitis in preschool aged children: regression analysis of factors associated with perforation outcome. J Pediatr Surg 2015;50:1569–73.

[17] Alshehri A, Lo A, Baird R. An analysis of early nonmortality outcome prediction in esophageal atresia. J Pediatr Surg 2012;47:881–4.

[18] Fawley JA, Peterson EL, Christensen MA, et al. Can omphalocele ratio predict postnatal outcomes? J Pediatr Surg 2016;51:62–6.