Biostatistics in Clinical Decision Making What can We Get from a 2×2 Contingency Table

Shanshan Zhang[1, *]

[1]Columbian College of Arts and Science, the George Washington University, Washington D.C., 20052, United States

Abstract. Biostatistics is an essential part when making clinical decisions. Applications of 2×2 contingency tables playing a key role in conducting analysis involving binary variables. When it comes to analysis based on 2×2 contingency tables, most people are familiar with the concept of sensitivity and specificity for evaluating a new test, but predictive values and receiver operating characteristic (ROC) curves would also provide information. Besides, Odds Ratio (OR), Risk Ratio (RR), and Chi-square test are measures based on 2×2 tables and commonly applied in retrospective and prospective studies. This article will first review the two kinds of application of 2×2 contingency tables, evaluating a new test compared with a reference standard, and exploring the relationship of exposures and outcomes in retrospective or prospective studies. Two clinical examples are presented to demonstrate these basic biostatistical concepts: diagnostic accuracy of 64-slice multidetector computed tomography (64-MDCT) to identify periampullary duodenal diverticula, and a randomized clinical trial (RCT) to examine the effectiveness of Dexmedetomidine for prevention of delirium in elderly patients after non-cardiac surgery. Correctly understanding these concepts will assist clinicians and medical researchers to analyze the data and interpret the results, and therefore make accurate decisions in clinical practice.

1 Introduction

For individuals, clinical decision making includes diagnosis and treatment, mostly depending on the results of laboratory tests and medical imaging. Clinicians make diagnoses in view of biomarkers shown on the test sheets and propose treatment according to the diagnosis and their experiences. The experience of a doctor comes from a large number of patients with the same disease, that is, the analysis based on population data, so-called biostatistics. When conducting analysis involving binary data, application based on 2×2 contingency table plays a key role and is widespread used in clinical decision making. 2×2 contingency tables are commonly used in clinical decision making. For most clinicians, when it occurs to 2×2 contingency tables, the first thing that comes into mind is the concept of sensitivity and specificity and the use to evaluate some new tests. However, the application of the contingency tables is far beyond that. Predictive values and receiver operating characteristic (ROC) curves can provide relevant information as well. Furthermore, retrospective case-control studies and prospective cohort studies use Odds Ratio (OR), Risk Ratio (RR), and Chi-square test to explore the relationship between exposures and outcomes. These measures are all based on the contingency tables.

The first part of this article will summarize the application of 2×2 contingency tables and the relevant statistical concepts in the context of comparing a new test to a reference standard and exploring the relationship between exposures and outcomes. In the latter part, two clinical examples are provided following the concept introduced to help understand the methodology used for clinical decision making.

2 Fundamental application in medical diagnosis

One of the classical applications of the traditional 2×2 contingency table is to evaluate a new clinical test or a novel kind of treatment compared with the reference standard, or some certain method of treatment. Values are typically organized, as shown in Table 1. In this manner, value a and d represent true positive (TP) and true negative (TN), meaning that a patient has a positive or negative result with a new method, and the case is also confirmed by the reference standard. Count b and c are recognized as false positive (FP) and false negative (FN), respectively, which indicates that the result of new test is opposite to that of the standard test.

* Corresponding author: *shanshan789789@gwu.edu

Table 1. Standard layout of a traditional 2×2 contingency table

		Reference Standard		
		+	-	Marginal totals
New test or treatment	+	a	b	a+b
	-	c	d	c+d
Marginal totals		a+c	b+d	n

2.1 Sensitivity and Specificity

The concepts of sensitivity and specificity were first proposed by Jacob Yerushalmy in 1947 [1], identified as a classical method of evaluating a new clinical test or treatment based on a preference standard. Sensitivity, also called TP rate (TPR), is the fraction of TP among patients tested positive with the new method. Specificity, or TN rate (TNR), similarly is the fraction of TN among those who had a negative result using a new test or treatment. The formulas are as follows:

$$Sensitivity = TPR = \frac{a}{a+b} \quad (1)$$

$$Specificity = TNR = \frac{d}{c+d} \quad (2)$$

How to evaluate a certain test or treatment by sensitivity and specificity depends on the application and context of the method. A highly sensitive test would be selected when there is a vital penalty for not detecting the disease. For example, a cardiothoracic radiologist would prefer a test with high sensitivity, since the missed diagnoses by false negative results are quite likely to delay the treatment of disease, like cancer, which would definitely lead to a fairly poor prognosis. Similarly, highly specific tests are particular important when FP results can harm the patients.

Reporting a confidence interval (CI) is needed to evaluate the precision of the estimates [2] and the accuracy of diagnosis and should be reported in all research. The absence of a CI report would cause considerable differences in the interpretation of results by clinicians. The most common CI reported in research papers is 95% CI, which could be computed basically by hand as:

$$95\%\ CI = p \pm 1.96\sqrt{\frac{p(1-p)}{n}} \quad (3)$$

where p is the calculated sensitivity or specificity, and n denotes the total sample size. This is a basic formula, based on which many alternative adjusted formulas can be obtained, such as Asymptotic Wald Interval, Wilson Score interval [3].

2.2 Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

When comparing the validity of a new test with the gold standard (or preference standard), sensitivity answers the question "If I have the disease (positive result for gold standard), what is the likelihood that I will test positive?", and specificity answers the question "If I do not have the disease (negative for gold standard), what is the probability for testing negative with a new test?". However, PPV answers the question "If I test positive, what is the likelihood that I truly have the disease?", and similarly, NPV answers the question "If I get a negative result, what is the probability that I do not really have the disease?". Sensitivity and specificity evaluate the test from the standpoint that the condition of disease or not is known, whereas PPV and NPV predict the likelihood of the disease based on the test results. The formulas are given by:

$$PPV = \frac{Number\ of\ persons\ who\ have\ the\ disease\ test\ positive}{Total\ number\ of\ person\ testing\ positive} = \frac{a}{a+b} \quad (4)$$

$$NPV = \frac{Number\ of\ persons\ who\ have\ the\ disease\ test\ negative}{Total\ number\ of\ person\ testing\ negative} = \frac{d}{c+d} \quad (5)$$

The predictive value of a test depends on three parameters: sensitivity, specificity, and disease prevalence. A higher prevalence in a population would increase the PPV and decrease the NPV. Therefore, when considering the predictive values of a new test for diagnosing, it is necessary to take the influence of the prevalence of disease into account [4].

2.3 ROC Curves

A perfect test has values of 1 for both sensitivity and specificity, which is rare to achieve and the balance of sensitivity and specificity is thus of importance. A trade-off always exists between the sensitivity and specificity, which is determined by cut-off values. Receiver operating characteristic (ROC) curves offer a graphical demonstration of each cut point to evaluate any diagnostic tests where continuous variables are involved [5]. The optimum cut-off value is the point through which the area under the curve (AUC) is maximized. A ROC curve describes the TPR (sensitivity) against FPR (false positive rate), also identified as 1-specificity. The 45-degree line means no discrimination, indicating that the TPR is equal to FPR. AUC is used to evaluate the quality of the test in a practical sense. The measure ranges from 0.5 (no discrimination line) to a theoretical maximum of 1, and over 0.75 can be considered good or useful for a test or a treatment.

3 Evaluation of the relationship between the outcomes and the exposures for a certain disease

Another situation where the contingency table is very common is in a case-control study or a cohort study for a certain disease. Cohort studies are longitudinal studies, where one or more cohorts are followed prospectively or designed retrospectively, and evaluations with respect to a disease or outcome are conducted to determine whether the exposures are associated with it. Case-control studies

historically compare patients who have the disease or outcome (case) and those who do not have (control), and then compare the occurrence of exposures in each group, in order to explore the relationship between the outcome and the exposures [6]. Values are organized, as shown in Table 2.

Table 2. Layout of 2×2 contingency table in the context of a case-control study or a cohort study

	Disease	Non-disease	Marginal totals
Exposure	a	b	m_1= a+b
Non-exposure	c	d	m_2=c+d
Marginal totals	n_1 = a+c	n_2 = b+d	n

3.1 Odds Ratio, Risk Ratio, and Chi-Square Test

Odds ratio (OR) is a common measure for a case-control study. It is the ratio of odds of disease among the exposure and non-exposure group. Risk ratio (RR) is usually used in a prospective cohort study, and it is the ratio of risks instead of the odds. The odds of disease are the ratio of the number of disease and non-disease persons, whereas the risk of disease, which is more intuitive than the definition of odds, is the proportion of the disease for the exposed and unexposed. The formula of OR and RR is given by:

$$OR = \frac{Odds\ of\ disease\ among\ exposed}{Odds\ of\ disease\ among\ unexposed} = \frac{a/b}{c/d} = \frac{ad}{bc} \quad (6)$$

$$RR = \frac{Risk\ of\ disease\ among\ exposed}{Risk\ of\ disease\ among\ unexposed} = \frac{a/(a+b)}{c/(c+d)} \quad (7)$$

The interpretation of OR is how many times higher the odds of disease among exposed than unexposed when OR is larger than 1, indicating a positive association between exposure and the outcome. When OR is smaller than 1, OR means the exposure reduces the odds of disease by 100 (1-OR) %, suggesting a negative association. OR value of 1 means no association between exposure and the outcome. Chi-square test is used to compare the OR and the value 1. The null hypothesis of the chi-square test is OR=1, whereas the alternative hypothesis is OR≠0. The test statistics is the sum of the contribution of test statistics of each cell in Table 2 and could be calculated as the following formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (8)$$

$$E = \frac{\sum R_i \times \sum C_j}{n} \quad (9)$$

$$df = (R-1) \times (C-1) \quad (10)$$

where O is the observed value (values in each cell); E is the expected value; $\sum R_i$ and $\sum C_j$ denote the sum of the observation for the cell's respective row and column; df is the degree of freedom; R and C denote the number of row and column, respectively. P-value is used to determine the significance of the test statistics, and 0.05 is usually used as the threshold. The 95% CI of OR can also be obtained by the following formula:

$$95\%\ CI_{OR} = \frac{ad}{bc} \times e^{\pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \quad (11)$$

$$95\%\ CI_{RR} = \frac{am_2}{bm_1} \times e^{\pm 1.96\sqrt{\frac{1}{a}-\frac{1}{m_1}+\frac{1}{b}-\frac{1}{m_2}}} \quad (12)$$

OR and RR can also be used to describe the comparison of two tests or two treatments with a contingency table shown in the last section. What needs to note is that if either FN or FP equals zero, then OR cannot be calculated, since the denominator is zero.

4 Clinical Examples

4.1 Clinical Example 1: Diagnostic accuracy of 64-slice multidetector computed tomography (64-MDCT) to identify periampullary duodenal diverticula.

Periampullary duodenal diverticula is a disease with small, bulging pouches located inside a radius of 2 cm of the protrusion of the ampulla of Vater into the duodenum. Detecting duodenal diverticula by Endoscopic retrograde cholangiopancreatography (ERCP) is considered as the gold standard for diagnosis. This method, however, is invasive, and scarcely used to early diagnose. Computed tomography (CT) is suggested as a safer surrogate for ERCP for symptomatic patients, and the accuracy is undoubtedly the most concerning problem. Table 3 presents a contingency table based on a retrospective study among patients who underwent both ERCP and MDCT to detect periampullary duodenal diverticula [7].

Table 3. Diagnosis summary of both ERCP and MDCT

		ERCP		
		+	-	Marginal totals of MDCT
MDCT	+	76	0	76
	-	24	20	44
Marginal totals of ERCP		100	20	120

The sensitivity is 76/100=76%; that is, 76% of patients who have periampullary duodenal diverticula will be diagnosed by MDCT. The specificity is 100%, which is a perfect true negative rate, meaning that there are no false positives among these 120 patients. The 95% CI of sensitivity is as follows:

$$95\%\ CI\ (sensitivity) = 0.76 \pm 1.96\sqrt{\frac{0.76(1-0.76)}{120}} = (0.68, 0.84) \quad (13)$$

The ability of the MDCT to correctly identify cases of periampullary duodenal diverticula ranges from 68% to 84%. Since the specificity here is 100%, the 95% CI cannot be obtained as a range. Tests with high specificity, such as this example, have more confidence in distinguishing TNs from FNs than discerning TPs from FPs. 100% specificity has vital meaning for patients:

emotional damage and unnecessary cost by misdiagnosis could be all avoided.

PPV here is 100%, whereas NPV is 20/44=45.5%. Hence, a patient who is diagnosed as periampullary duodenal diverticula by MDCT would have the disease, based on this dataset, whereas patients with a negative MDCT result still have a likelihood of 45.5% truly having the disease. In this case, OR cannot be obtained as a specific value since FN is zero.

4.2 Clinical Example 2: A randomized clinical trial (RCT) to examine the effectiveness of Dexmedetomidine for prevention of delirium in elderly patients after non-cardiac surgery

This example is based on a randomized double-blind, parallel-arm placebo-controlled clinical trial published in Lancet 2016 [8]. The participants were all aged 65 years or older and underwent non-cardiac surgery, one of the severe postoperative complications of delirium. Table 4 presents a contingency table based on this dataset.

Table 4. Data summary of the RCT in example 2

	Delirium	Non-delirium	Marginal totals
Dexmedetomidine	32	318	350
Placebo	79	271	350
Marginal totals	111	589	700

According to Table 4, OR is $\frac{32/318}{79/271} = 0.35$, meaning that the intervention can reduce 65% the odds of delirium. The 95% CI is given by:

$$95\%\ CI\ (OR) = 0.35 \times e^{\pm 1.96\sqrt{\frac{1}{32}+\frac{1}{318}+\frac{1}{79}+\frac{1}{271}}} = (0.22, 0.54) \quad (14)$$

The results are consistent with the article. Since this a retrospective study, RR can also be used, and the interpretation is more intuitive. Using the previous methodology, RR is $\frac{32/350}{79/350} = 0.41$, meaning that the risk of delirium was reduced by 59% with Dexmedetomidine. Similarly, the 95% CI is as follows:

$$95\%\ CI\ (RR) = 0.41 \times e^{\pm 1.96\sqrt{\frac{1}{32}-\frac{1}{350}+\frac{1}{79}-\frac{1}{350}}} = (0.28, 0.60) \quad (15)$$

P-value < 0.001 in Chi-square test, giving a significant OR and RR, indicating that the incidence of postoperative delirium was significantly lower in the intervention group than in the placebo group. This conclusion is also the same as the article. The authors used OR to measure the effectiveness of Dexmedetomidine. However, in a prospective study, the prevalence of disease in the study cohort is known, and RR is obviously more intuitively interpreted than OR.

5 Conclusion

This article demonstrated two mainstream applications of 2×2 contingency tables in biostatistics for medical decision making. The first application is to evaluate a new test compared with a reference standard. Sensitivity, or true positive rate, and specificity, or true negative rate, are two basic concepts to compare the accuracy of new tests with a standard reference. Predictive values allow researchers to evaluate the performance of the test in a population outside the study sample. The ROC curve and AUC are useful when setting up a reasonable cut point of a test. The other application is to explore the relationship between exposures or interventions with outcomes or diseases. OR is a measure widely utilized in retrospective studies, such as a case-control study due to lack of prevalence in population. In contrast, RR is usually used in prospective studies, such as a prospective cohort study, for a more intuitive clinical interpretation. Chi-square test is a statistical method to detect whether an association exists between exposures and outcomes, the p-value of which is widespread used to measure the significance, and 0.05 is always considered as a threshold. CI of the measures should be reported to provide more information and the precision of diagnostic characteristics.

Two clinical examples provided in the latter part of the article are both based on real data published in journals. The first example demonstrated the application of sensitivity, specificity, and predictive values of MDCT compared with gold standard ERCP to diagnose periampullary duodenal diverticula, and the results are consistent with the original article. The second example illustrated the application of OR, RR, and Chi-square test to explore the association between Dexmedetomidine intervention and the occurrence of delirium. The example used the measure of RR, except OR and Chi-square test conducted in the original article, to argue the relationship more intuitively, and get the same conclusion.

References

1. Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports (1896-1970), 1432-1449.
2. Harper, R. (1999). Reporting of precision of estimates for diagnostic accuracy: a review. Bmj, 318(7194), 1322-1323.
3. Erdogan, S., & Gulhan, O. T. (2016). Alternative Confidence Interval Methods Used in the Diagnostic Accuracy Studies. Comput Math Methods Med, 2016, 7141050. doi:10.1155/2016/7141050
4. Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 2: predictive values. Bmj, 309(6947), 102.
5. Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. CJEM, 8(1), 19-20. doi:10.1017/s1481803500013336
6. Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. Plast

Reconstr Surg, 126(6), 2234-2242. doi: 10.1097/PRS.0b013e3181f44abc

7. Eghbali, E., Tarzamni, M. K., Shirmohammadi, M., Javadrashid, R., & Fouladi, D. F. (2020). Diagnostic performance of 64-MDCT in detecting ERCP-proven periampullary duodenal diverticula. Radiol Med, 125(4), 339-347. doi:10.1007/s11547-019-01121-w
8. Su, X., Meng, Z.-T., Wu, X.-H., Cui, F., Li, H.-L., Wang, D.-X., . . . Ma, D. (2016). Dexmedetomidine for prevention of delirium in elderly patients after non-cardiac surgery: a randomised, double-blind, placebo-controlled trial. The Lancet, 388(10054), 1893-1902. doi:10.1016/s0140-6736(16)30580-3