Confusion matrix

Shengping Yang PhD, Gilbert Berdine MD

I am evaluating the sensitivity and specificity of an assay for COVID-19 diagnosis, and our team is developing a confusion matrix for this analysis. Could you explain the key considerations when using a confusion matrix for this purpose?

In biomedical research, particularly when evaluating diagnostic tests or predictive models, performance metrics are essential for assessing the effectiveness of assays or classification systems. One commonly used tool is a contingency table, which displays the frequency distribution of categorical variables. A confusion matrix is a specialized form of a contingency table used to assess the performance of classification algorithms by showing the actual versus predicted outcomes. It is particularly useful for evaluating key metrics such as sensitivity, specificity, accuracy, and precision, which are crucial for interpreting the performance of diagnostic tests.

1. The Confusion matrix

A confusion matrix is a specific type of two-dimensional contingency table used to evaluate the performance of a classification model. Its two dimensions, "actual" and "predicted," represent identical sets of "classes" (e.g., disease positive and disease negative), allowing for a direct comparison between actual and predicted outcomes.[1,2]

Specifically, in a confusion matrix, each row represents an actual class, while each column represents a predicted class (or vice versa). The diagonal cells represent correctly predicted outcomes, while the off-diagonal cells represent misclassifications. The matrix provides a clear visualization of where the model confuses different classes, which is why it is called a confusion matrix.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.Yang@pbrc.edu
DOI: 10.12746/swrccc.v12i53.1391

Table 1. An Example Confusion Matrix

		Predicted	
		Positive (PP)	**Negative (PN)**
Actual	**Positive (P)**	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Table 1 provides an example of a confusion matrix, where the rows represent actual conditions, and the columns represent predicted conditions. The matrix contains four key components: True Positives (TP): The model predicts positive, and the actual condition is positive; False Negatives (FN): The model predicts negative, but the actual condition is positive; False Positives (FP): The model predicts positive, but the actual condition is negative; True Negatives (TN): The model predicts negative, and the actual condition is negative. Furthermore, the sums of these components define: Actual positive cases (P) = TP + FN; Actual negative cases (N) = FP + TN; Predicted positive cases (PP) = TP + FP; Predicted negative cases (PN) = FN + TN.

While the confusion matrix presents data in a straightforward manner, several important metrics can be derived from it to assess the performance of a diagnostic test or predictive model.

2. Key metrics derived from the confusion matrix

Commonly used metrics that can be derived from a confusion matrix include.[2,3]

- *Sensitivity* (or True Positive Rate; TPR): The proportion of actual positives that are correctly classified as positive, calculated as TP/P. A high sensitivity is particularly important in detecting diseases in which missing a positive case (FN) could have severe consequences. For example, in cancer screening,

prioritizing high sensitivity is vital to ensure that as many true positive cases as possible are detected early, enabling timely treatment and improving the patient's prognosis.

- *Specificity* (or True Negative Rate; TNR): The proportion of actual negatives that are correctly classified as negative, calculated as TN/N. A high specificity is crucial in situations in which it is important to avoid classifying healthy individuals as diseased (FP). For example, in HIV testing, while both high sensitivity and high specificity are important, high specificity ensures that healthy individuals are not wrongly diagnosed as HIV-positive. This helps prevent the significant consequences of a false positive diagnosis, including unnecessary emotional stress, social stigma, discrimination, legal complications, and unwarranted medical interventions.
- *False positive rate* (FPR): Measures the likelihood of a false alarm (predicted positive) among those classified as negative, calculated as FP/N = 1-TN/N = 1-specificity.
- *Accuracy*: The overall proportion of correctly classified instances is calculated as (TP + TN)/(P + N). While commonly used, accuracy can be misleading in the context of imbalanced datasets, in which the number of observations in different classes varies greatly. For example, imagine you are evaluating a diagnostic test for a disease that affects 1% of the population. In a sample of 1,000 individuals, only 10 (P = 10) people actually have the disease, while the remaining 990 (N = 990) do not. The test predicts all 1,000 individuals as negative, meaning it misses all the true positives but correctly identifies all the true negatives (Table 2). Therefore, accuracy can be calculated as: (TP + TN)/(P + T) = (0 + 990) / 1,000 = 99%. However, despite the high accuracy of 99%, the test is completely ineffective at identifying individuals with the disease (sensitivity = 0). In this case, the accuracy gives a false sense of the test's effectiveness, as it is driven by the large number of true negatives in an imbalanced dataset. More appropriate metrics in this situation would be sensitivity or precision.
- *False discovery rate* (FDR): The FDR is the proportion of false positives out of all predicted positives, calculated as FP / (TP + FP). While a lower FDR is generally desirable, a high FDR can sometimes be misleading.[4] For example, consider a rare disease affecting 1% of the population. In a sample of 10,000 individuals, only 100 (P = 100) actually have the disease, while 9,900 (N = 9,900) do not. Assume the test has a sensitivity of 90% and a specificity of 99%. This would result in: 90 true positives (90% of 100 diseased individuals), 9,801 true negatives (99% of 9,900 healthy individuals; Table 3). Though the test performs well in terms of sensitivity (90%) and specificity (99%), the FDR calculation reveals a different picture: FDR = FP / (TP + FP) = 99 / (90 + 99) ≈ 52.4%, meaning that over half of those who test positive are actually false positives. This high FDR is somehow misleading because, despite the test's effectiveness (90% sensitivity and 99% specificity), the low prevalence of the disease (1%) makes false positives more noticeable. It is important to note that while the FDR appears high in low-prevalence populations, the test could still be valuable in higher-prevalence settings or when sensitivity is prioritized, such as in early screening. Nevertheless, in low-prevalence contexts, precision may be a better metric to assess alongside sensitivity and specificity.
- *Precision* (Positive Predictive Value): Proportion of predicted positives that are true positives and can be calculated as TP/PP, or equivalently, 1-FDR.

Table 2. An Example of an Imbalanced Dataset

		Predicted	
		Positive (PP)	**Negative (PN)**
Actual	**Positive (P)**	0	10
	Negative (N)	0	990

Table 3. An Example of Test Results for a Rare (Low Prevalence) Disease

		Predicted	
		Positive (PP)	**Negative (PN)**
Actual	**Positive (P)**	90 (100 × 90%)	10
	Negative (N)	99	9,801 (9,900 × 99%)

This metric is particularly important when false positives can be costly. Precision answers the question, "Of all the individuals the test identified as positive, how many actually have the disease?" In low-prevalence settings, precision becomes especially crucial because false positives may outnumber true positives simply due to the low number of actual cases. Now, let's revisit the previous example, and the precision would be Precision = TP / (TP + FP) = 90 / (90 + 99) ≈ 47.6%. This means that 47.6% of individuals who test positive actually have the disease. While this may seem low, it offers a clearer understanding than FDR regarding the usefulness of the test for identifying true positives. It is worth noting that by combining sensitivity, specificity, and precision, clinicians and researchers can better evaluate the likelihood that a positive test result accurately indicates the presence of disease. Particularly, in low-prevalence situations, precision is key for assessing the clinical utility of a positive test result, while sensitivity and specificity evaluate the overall accuracy and reliability of the test. Together, these metrics provide a comprehensive picture of test performance.

There are other metrics, such as the F1 Score, the Fowlkes-Mallows Index, and the Matthews Correlation Coefficient, etc.[5] However, these will not be discussed in detail.

3. Applications in biomedical research

There are many applications of a confusion matrix in biomedical research:

- *Diagnostic tests*: The confusion matrix is often used to evaluate the performance of medical diagnostic tests for diagnosing diseases, helping assess how well tests distinguish between disease and non-disease cases.[6]
- *Predictive modeling*: In areas like disease risk prediction, confusion matrices are used to evaluate classifiers predicting outcomes such as heart disease, cancer, or other conditions.[7]
- *Imaging and segmentation*: In medical image analysis (e.g., MRI scans, histopathology), confusion matrices are useful for evaluating algorithms that classify or segment regions of interest.[8,9]
- *Drug discovery and development*: Models predicting compound efficacy or toxicity are often evaluated using confusion matrices to balance precision and sensitivity in identifying potentially harmful side effects.[10]
- *Genomics and proteomics*: In genomics research, classifiers are often used to predict outcomes like gene expression profiles, locations of enhancers, or disease susceptibility based on high-dimensional biological data. Confusion matrices are critical in assessing the performance of these classifiers by breaking down their predictions into categories that help evaluate diagnostic or predictive accuracy.[11]

4. Challenges in biomedical applications

There are challenges in determining the best application of a confusion matrix in biomedical research.

- *Determination of actual positive and actual negative states:* For many conditions, a "gold standard" test for actual positive and actual negative states may not exist. For example, the diagnosis of pneumonia has no generally accepted inclusion and exclusion criteria. Acute respiratory distress syndrome (ARDS) is another example; even the definition of the acronym ARDS has evolved over time by general consensus. The basic definition is "bilateral pulmonary infiltrates on chest imaging, hypoxia, and exclusion of congestive heart failure." All three points lack precise definitions with universally accepted thresholds. How much opacity is necessary on each side to consider an image to represent bilateral pulmonary infiltrates? What is the definition of hypoxia with an objective numeric threshold? It has been generally accepted that the pulmonary capillary wedge pressure (PCWP) must be measured and be less than some threshold, but the actual threshold has evolved over time depending on whether the concern is how borderline cases are categorized. For many conditions, an "expert" panel of opinion replaces an objective test until some objective test is considered good enough to become the new "gold standard."

Once the new test becomes the "gold standard" how many false positive and false negatives were found in original studies? The "Light's" criteria for categorizing pleural effusion as transudate or exudate is an example of replacing one set of uncertainty with a new set of uncertainty.

- *Imbalanced datasets*: Many biomedical datasets exhibit significant imbalance, often containing fewer positive cases (e.g., rare diseases). In such scenarios, traditional accuracy metrics may be misleading. It is crucial to emphasize the importance of alternative metrics such as precision and sensitivity to provide a more accurate evaluation of model performance.
- *Cost of errors*: The implications of false positives and false negatives can vary considerably from one study to another. For example, in cancer screening, false negatives (missed diagnoses) can have far more serious consequences than false positives (leading to unnecessary testing). A comprehensive understanding of these errors can greatly facilitate the decision-making process.

Additional challenges include threshold tuning and multi-class classification. These issues must be addressed with careful consideration of the specific diseases or conditions being evaluated in clinical practice.

5. Confusion matrix and receiver operating characteristic (ROC) curve

Both the confusion matrix and the ROC curve are essential tools for evaluating classifier performance, but they serve different purposes. The confusion matrix offers insights into performance at specific decision points, detailing metrics such as precision, sensitivity, and accuracy at a particular threshold. In contrast, the ROC curve plots sensitivity against the FPR (1–specificity), at various classification thresholds, which summarizes performance across all possible thresholds, providing a broader perspective on how effectively the classifier distinguishes between positive and negative classes.[12] Importantly, both sensitivity and FPR used in ROCs can be derived from a confusion matrix. In this sense, the confusion matrix and the ROC curve complement each other in evaluating the performance of a diagnostic tool.

In summary, the confusion matrix is an important tool for evaluating diagnostic tests/tools performances, particularly in healthcare and biomedical research. While it provides straightforward metrics for assessing the performance of such tests/tools, the interpretation of these metrics can sometimes be misleading, especially in the context of rare diseases. Nevertheless, confusion matrices are increasingly used in many areas of biomedical research including personalized medicine, explainable artificial intelligence modeling, etc.

Article citation: Yang S, Berdine G. Confusion matrix. The Southwest Respiratory and Critical Care Chronicles 2024;12(53):75–79
From: Department of Biostatistics (SY), Pennington Biomedical Research Center, Baton Rouge, LA; Department of Internal Medicine (GB), Texas Tech University Health Sciences Center, Lubbock, Texas
Submitted: 10/9/2024
Accepted: 10/10/2024
Conflicts of interest: none

References

1. Chapter 2 Contingency Tables (uchicago.edu). (last access: Oct. 5, 2024)
2. Confusion matrix – Wikipedia. (last access: Oct. 5, 2024)
3. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas) 2021 May 16;57(5):503. doi: 10.3390/medicina57050503
4. Guesné SJJ, Hanser T, Werner S, Boobier S, Scott S. Mind your prevalence! J Cheminform 2024 Apr 15;16(1):43. doi: 10.1186/s13321-024-00837-w
5. Chicco D, Jurman G. A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes-Mallows index. J Biomed Inform 2023 Aug; 144:104426. doi: 10.1016/j.jbi.2023.104426

6. Lin D, Liu L, Zhang M, et al. Evaluations of the serological test in the diagnosis of 2019 novel coronavirus (SARS-CoV-2) infections during the COVID-19 outbreak. Eur J Clin Microbiol Infect Dis 2020 Dec;39(12):2271–7. doi: 10.1007/s10096-020-03978-6. Epub 2020 Jul 17.
7. Pal M, Parija S, Panda G, Dhama K, et al. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars) 2022 Jun 17;17(1):1100–13. doi: 10.1515/med-2022-0508
8. Gull S, Akbar S, Khan HU. Automated detection of brain tumor through magnetic resonance images using convolutional neural network. Biomed Res Int 2021 Nov 30; 2021: 3365043. doi: 10.1155/2021/3365043
9. Klauschen F, Goldman A, Barra V, et al. Evaluation of automated brain MR image segmentation and volumetry methods. Hum Brain Mapp 2009 Apr;30(4):1310–27. doi: 10.1002/hbm.20599
10. Adeluwa T, McGregor BA, Guo K, et al. Predicting drug-induced liver injury using machine learning on a diverse set of predictors. Front Pharmacol 2021 Aug 18;12:648805. doi: 10.3389/fphar.2021.648805
11. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015 Jun;16(6): 321–32. doi: 10.1038/nrg3920
12. Confusion Matrix: How To Use It & Interpret Results [Examples] (v7labs.com). (last access: Oct. 5, 2024)