BJPsych The British Journal of Psychiatry (2025)
227, 4, 723–724. doi: 10.1192/bjp.2025.113

Letter

Diagnostic accuracy for multiple categories: statistical advice

Jialiang Li, Wangcheng Li and Lei Feng

Keywords
Diagnostic medicine; ROC manifold; HUM; MCI; PTSD.

Many diagnostic tasks require the accurate assignment of patients into more than two conditions. Not assigning patients to the correct category may result in sub-optimal treatment. To achieve good classification performance we may use relevant diagnostic tests or combine them into a risk prediction model. For statistical analysis we need to describe the accuracy of the tests or models. Though recent medical literature offered various practical tools to assess the quality of predictive models,[1] we noticed that specific discussions on multi-category diagnostic outcome are limited.

One-vs-rest may be misleading

A common method for evaluating test accuracy is to use the area under the ROC curve (AUC) for binary classification. Many researchers extend this to a "one-versus-rest" (OVR) framework for multiclass problems. For a 4-class diagnostic problem, the one-vs-rest (OVR) approach produces four AUC values, one for each class against the rest. However, this approach is often criticized for two reasons. First, sometimes the "rest" does not represent a homogeneous class and may not make any sense. Second, there is no single measure to provide an overall assessment of accuracy for differentiating all classes. For example, consider using a diagnostic test for a 4-class analysis where classes I to IV represent normal, mild, moderate, and severe disease conditions in a progressive manner. If the AUC values of a diagnostic model for each class versus the rest are 0.76, 0.69, 0.53, and 0.78 respectively, the result for class III versus the rest, 0.53, is particularly misleading. This is because "the rest" constitutes a highly heterogeneous group including patients with both milder and more severe conditions. Furthermore, the 4 AUC values do not offer a single summary of the model's overall accuracy. A similar criticism applies to the one-versus-one approach, where, for a 4-class analysis, there are $\binom{4}{2} = 6$ pairwise AUC values.

A more appropriate measure

In the statistical and machine learning literature, a more suitable accuracy measure for multi-class diagnostic tests is the hypervolume under the ROC manifold (HUM). HUM extends the concept of AUC and quantifies how often a diagnostic model can differentiate among all classes simultaneously. A higher HUM value indicates superior classification performance, and when the number of categories is two, HUM reduces to AUC. Similar to AUC, HUM can be used to evaluate the accuracy of a single biomarker or a risk prediction model involving multiple markers. The computation of HUM is supported by several R packages. For single biomarkers or diagnostic tests, the HUM and mcca packages can both be used to compute the desired results, while the mcca package is also appropriate for evaluating accuracy in statistical or machine learning models such as decision trees, logistic regression, or neural networks. See refs.[2,3] for recent reviews.

An illustration

We consider a proteomics dataset from Kuan et al[4] which consisted of 276 proteins targeting neurobiological process, cellular regulation, immunological function, cardiovascular disease, inflammatory processes, and neurological diseases on $n = 154$ responders involved in the September 11, 2001 World Trade Center (WTC) terrorist attack. The WTC responders suffered from clinically significant post-traumatic stress disorder (PTSD) and mild cognitive impairment (MCI). Thus, it is an important task to identify effective biomarkers which can detect responders at risk of developing PTSD and MCI, as well the disease burden characterized by co-occurrence of PTSD-MCI, which can be targeted for intervention and inform treatment efforts. We thus considered a three-class discrimination analysis. Among all responders, 81 are in the control group, 39 have PTSD only and 34 have both PTSD and MCI. The dataset is extracted from Supplementary Table 1 in Kuan et al[4] which is available under open access at https://doi.org/10.1038/s41398-020-00958-4.

We present the following sample R code that demonstrates the calculation of HUM value for a single biomarker:

```
#### 1. Preparation
# install.packages(c("mcca","readxl"))
## Import dataset from the excel file:
PTSD=readxl::read_excel
("./41398_2020_958_MOESM4_ESM.xlsx")
PTSD=PTSD[PTSD$Group %in% c("Control",
"PTSD-only","PTSD-MCI"),]
protein_expr = PTSD[,-1] # Matrix(154x276):
The expression information of 276 proteins on
154 participants
categories = factor(PTSD$Group) #
Vector(154): The categories of 154
participants
protein_name = mapply(function(x){x[1]},
strsplit(colnames(protein_expr),"_")) #
Vector(276): The names of 276 proteins
#### 2. Calculate HUM of each protein
hum.value=c()
for (i in 1:276){
  hum.value=c(hum.value, mcca::hum(catego-
  ries,protein_expr[,i])$measure)
}
dt = data.frame(Name = protein_name,
HUM=hum.value)
head(dt,3)
#> Name HUM
#> 1 IFNL1 0.2739884
#> 2 EIF4B 0.1812841
#> 3 CRADD 0.1981267
```


https://doi.org/10.1192/bjp.2025.113 Published online by Cambridge University Press

The output above displays three protein markers along with their HUM values. After evaluating the induvial HUM values we can easily screen out markers with low classification accuracy.

We notice that using a single marker may not be sufficient to provide adequate prediction for multi-class outcome. Next we demonstrate how to calculate the HUM for a model involving several biomarkers. To this end, we select 12 proteins through a forward selection strategy and use them in a regression model to predict the three-class outcome. In the following R code, we first fit the multinomial logistic model and predict the probability vectors. Then, we use mcca package to calculate HUM value based on the estimated probability vectors:

```
#### 3. calculate HUM of the fitted model, taking multinomial logistic model as an example
## Select 12 proteins into the model:
sub.protein_expr = protein_expr[,c(16,18, 20, 24, 96,107,132,150,167,229,238,253)]
df = data.frame(cbind(sub.protein_expr, categories))
## Fit the model:
model.multimomial = nnet::multinom(categories~.,data = df)
## Predict the probability vectors:
prob.estimated = predict(model.multimomial, df,type = "prob")
## Calculate the HUM:
hum.manual = mcca::hum(categories, prob.estimated,method = "prob")$measure
print(hum.manual)
#> 0.7864179
```

The R output indicates that the model achieves a HUM value 0.7864, almost five times better than a random guess,[2] suggesting that the model can simultaneously differentiate the three groups accurately with such a high probability.

In general, the `mcca` package can accommodate all kinds of statistical and machine learning programs for classification, in addition to multinomial logistic regression. Users can substitute `prob.estimated` with probability assessment vectors derived from other packages, or simply modify the "method" option in the above code.

Jialiang Li, Department of Statistics & Data Science, National University of Singapore, Singapore; **Wangcheng Li**, School of Statistics, Beijing Normal University, Beijing, China; **Lei Feng**, Department of Psychological Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Correspondence: Jialiang Li. Email: jialiang@nus.edu.sg.

First received 8 Dec 2024, final revision 8 Feb 2025, accepted 22 Mar 2025

Author contribution

LJ: formal analysis, writing initial draft; **LW**: data implementation, reviewed the manuscript; **FL**: study conceptualization, reviewed the manuscript.

Funding

This study received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interest

There is no conflict of interest.

References

1 Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019; **170**: W1–W33.

2 Li J, Gao M, D'Agostino R. Evaluating classification accuracy for modern learning approaches. *Stat Med* 2019; **38**: 2477–503.

3 Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 2022; **12**: 5979.

4 Kuan PF, Clouston S, Yang X, Kotov R, Bromet E, Luft BJ. Molecular linkage between post-traumatic stress disorder and cognitive impairment: a targeted proteomics study of World Trade Center responders. *Transl Psychiatry* 2020; **10**: 1–15.

https://doi.org/10.1192/bjp.2025.113 Published online by Cambridge University Press