MACHINE LEARNING FOR HEALTH

Performance Evaluation: Beyond Accuracy

Metrics, calibration, external validation, and equity in clinical models

TECHNICAL HANDOUT · SESSION 05

Professor Flávio Luiz Seixas
Graduate Program in Computing

Academic teaching material based exclusively on the PDFs available in the project

Table of Contents

Introduction

Evaluating a clinical classifier requires relating errors to prevalence, threshold, and consequence. Aggregate accuracy may appear high even when the model fails to identify the relevant event, especially in imbalanced problems. [1]

This session connects the confusion matrix, sensitivity, predictive value, curves, calibration, validation, and subgroup analysis. The question shifts from "which number is larger?" to "what property was measured, and what decision can it support?" [1]

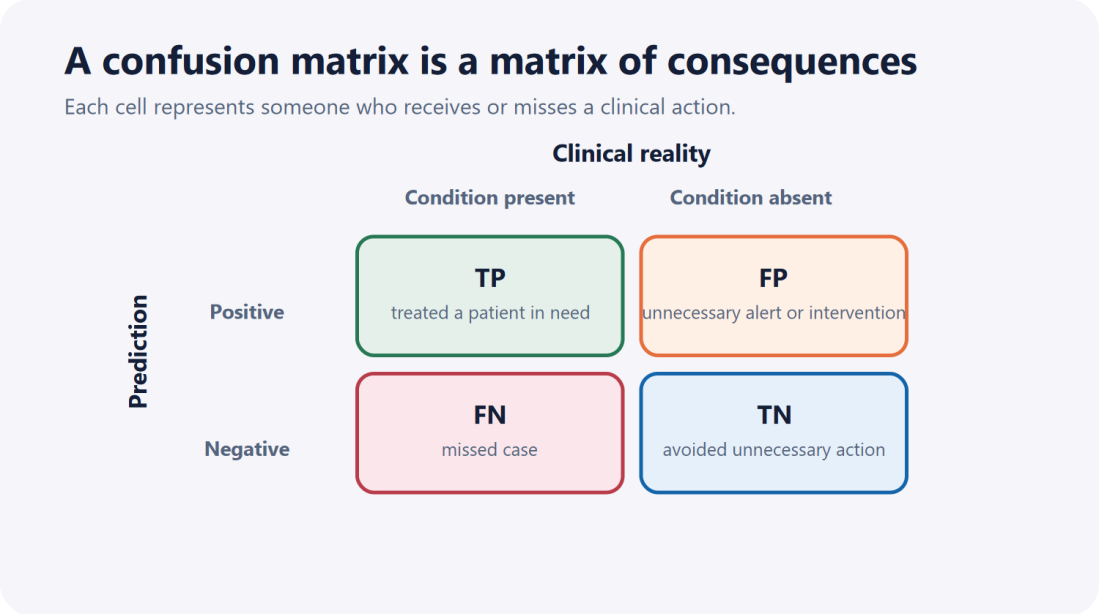

Figure 1 — Confusion matrix interpreted as the consequences of classification. Source: original illustration based on [1].

1. Confusion matrix and conditional measures

The confusion matrix separates true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Sensitivity is TP/(TP+FN): among cases with the condition, how many are detected. Specificity is TN/(TN+FP): among cases without the condition, how many receive a negative result [1].

Positive predictive value—equivalent to precision in classification terminology—is TP/(TP+FP). Unlike sensitivity and specificity, predictive values depend on prevalence in the population. The same test may therefore produce different proportions of true positive results when the context of use changes [1].

KEY CONCEPT
The confusion matrix separates true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

2. Thresholds, discrimination, and imbalance

Changing the threshold shifts the balance between false positives and false negatives. The ROC curve summarizes sensitivity against the false-positive rate across thresholds; the precision–recall curve emphasizes the behavior of the positive class. No area under a curve can choose the clinical threshold on its own because costs and the capacity to intervene remain external to the metric [2].

The F1 score combines precision and sensitivity through their harmonic mean. It is useful when both matter, but it does not directly incorporate specificity, calibration, or clinical utilities. Metrics should be chosen from the task and its consequences [5].

3. Calibration and validation

Discrimination ranks individuals by risk; calibration compares predicted probabilities with observed frequencies. Calibration is central when probabilities guide personalized decisions. It may deteriorate as the population, incidence, and practice change [2].

Internal validation estimates optimism within the development process. Temporal and external validation test transportability to other periods and institutions. A complete clinical evaluation moves from statistical performance to integration, use, and impact [3].

4. Equity across subgroups

An overall average may conceal differences between groups. Subgroup auditing should compare data coverage, prevalence, calibration, and relevant error rates. A disparity requires investigation of the data, formulation, model, and use because bias can arise in every one of these phases [4].

Session summary

Confusion matrix and conditional measures: the confusion matrix separates true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn).
Thresholds, discrimination, and imbalance: changing the threshold shifts the balance between false positives and false negatives.
Calibration and validation: discrimination ranks individuals by risk; calibration compares predicted probabilities with observed frequencies.
Equity across subgroups: an overall average may conceal differences between groups.

References

[1] Owens, D. K.; Goldhaber-Fiebert, J. D.; Sox, H. C. Biomedical Decision Making: Probabilistic Clinical Reasoning. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/III-Biomedical Decision Making- Probabilistic Clinical Reasoning.pdf · PDF pages used: 12, 13, 14, 17, 20.

[2] Bellazzi, R.; Dagliati, A.; Nicora, G. Predicting Medical Outcomes. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XI-Predicting Medical Outcomes.pdf · PDF pages used: 20, 21, 22, 23, 24.

[3] Shortliffe, E. H.; Sepúlveda, M.-J.; Patel, V. L. Framework for the Evaluation of Clinical AI Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XVII-Framework for the Evaluation of Clinical AI Systems.pdf · PDF pages used: 6, 10, 15, 20.

[4] Korngiebel, D. M.; Solomonides, A.; Goodman, K. W. Ethical and Policy Issues. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/XVIII-Ethical and Policy Issues.pdf · PDF pages used: 7, 8, 9, 14.

[5] Subramanian, D.; Cohen, T. A. Machine Learning Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VI-Machine Learning Systems.pdf · PDF pages used: 8, 9.