

When accuracy isn't enough: A machine-learning case in healthcare decision-making

David J. Ranney

Global Health Executive Mentoring

Author Note

ORCID iD's

Contact Information: david.ranney@afchca.org

When accuracy isn't enough: A machine-learning case in healthcare decision-making

Can a less accurate machine-learning model make a better healthcare decision? No, right? I mean, seriously, why would any decision-maker choose a model that correctly classifies 85% of cases over another that correctly classifies 87%? As it turns out, the answer depends on what the model gets wrong, and what happens when it does.

A neural-network exercise using COVID-19 data provides a useful example. The model used patient characteristics including cough, fever, sore throat, shortness of breath, headache, age, sex, and known contact to classify cases as COVID-positive or COVID-negative. After data preparation, the dataset contained 21,936 observations.

The first model achieved 86.83% accuracy. But accuracy told only part of the story. Its recall was 74.14%, meaning the model was substantially less successful at identifying positive cases than its overall accuracy might suggest. The exercise specifically identified false negatives as a concern.

There was also an important imbalance in the data: 15,420 negative cases compared to 6,516 positive cases. To address that imbalance, the model was retrained using class weights that gave greater importance to the less common positive class.

The updated model results initially look worse:

Table 1

Initial model vs Weighted model

Measure	Initial Model	Class-weighted model
Accuracy	86.83%	84.98%
Recall	74.14%	86.11%
Precision	80.03%	70.13%
F1 Score	76.97%	77.30%

So, which model is better?

And now, we've arrived at the intersection of data science and decision-making. A data scientist can explain what happened. By changing the importance assigned to the positive cases during training, the model became more sensitive to positive cases. That reduced one kind of error while increasing another. Healthcare leaders must ask what those errors mean once the model leaves the notebook.

The Tradeoff

Errors may have very different consequences depending on how the model is being used. The confusion matrix is a simple yet effective visual for model errors.

Figure 1

Confusion Matrix (provided by ibis-project.org)

		Prediction	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

The confusion matrix is an easy-to-read and interpret visual guide for understanding actual versus predicted results. About the matrix:

- A true positive: both actual result and predicted result are TRUE
- A false negative: actual result is positive while the predicted result is negative
- A false positive: actual result is negative while the predicted result is positive
- A true negative: both actual result and predicted result are NEGATIVE

Digging deeper

False positive. The model classifies someone who is actually negative as positive. Put yourself in the following position: you feel fine, but during routine screening, you're told 'the system' has classified you as COVID-positive. Eish. Depending on your location and

occupation, that result could mean isolation, missed work, additional testing, and concern among family or colleagues. All while not actually having COVID.

False negative. Now reverse the situation. You're feeling a bit under the weather, but 'the system' classifies you as COVID-negative. The result provides reassurance when caution may actually be warranted. You return to work, visit family, or simply become less cautious because you've been told you're negative. But you actually have COVID.

The Fallout

As discussed briefly above, false positives and negatives are consequential. If the model were part of a screening process during an infectious-disease outbreak, missing a positive case might allow an infected person to enter a workplace or clinical environment without additional precautions. Adjusting the model to increase sensitivity to positive cases could also have consequences: additional testing, isolation, use of staff time or protective equipment, delays, anxiety, and consumption of limited resources.

The appropriate tradeoff could therefore differ between a well-resourced hospital, a community screening program, and a health system operating with severe shortages of testing capacity, staff, beds, or protective equipment.

That context isn't contained in any accuracy score.

The Decision Behind the Model

Machine learning can quantify performance. It cannot, by itself, determine which consequences a healthcare organization should accept. That requires another conversation.

The data scientist needs to explain how the model behaves, where uncertainty exists, and what changes when the model is adjusted. Clinicians, healthcare executives, financial leaders,

operational leaders, and others then need to translate those technical tradeoffs into consequences for patients and the health system.

The important question therefore isn't simply which model is more accurate, it's: which errors can we tolerate for this particular healthcare decision, and what are the consequences when we're wrong?

That distinction becomes increasingly important as healthcare organizations adopt machine learning and artificial intelligence. Technical performance matters, but a model does not create value merely because its metrics look impressive.

The real test begins when model output becomes somebody else's decision.