

Diagnostic Performance Evaluation through Proteomics and Biomarker Signatures in Precision Medicine

Sadia Tasmeeem

Jul 12, 2026

Abstract

The article evaluates proteomic diagnostic performance as a multidimensional problem within precision medicine. It argues that sensitivity and specificity alone are insufficient because a useful diagnostic signature must also demonstrate calibration, reproducibility, prevalence-aware predictive value, subgroup stability, analytical precision, and clinical benefit. High-dimensional protein platforms can discover rich molecular patterns, but unrealistic case-control designs, sample-handling differences, technical drift, and overfitting can create inflated results. The discussion uses tumor-associated autoantibody panels to illustrate both the promise and limitations of multianalyte signatures. Stronger evidence requires truly independent validation cohorts, standardized preanalytics, orthogonal measurement, repeated-sample reliability, transparent handling of missing values, and comparison with simpler clinical alternatives. Decision-curve analysis and prospective workflow studies are emphasized because clinical value depends on changing management safely. The article concludes that proteomic diagnostics become credible only when discovery, analytical validation, external validation, and decision utility are demonstrated together in realistic patient populations.



Proteomic diagnostics are usually at their best when they are discovered and a molecular fingerprint is compared to a narrowly drawn group of cases and controls. The more difficult question with precision medicine is whether the same signature is predictive enough to make a sound clinical decision. Diagnostic performance isn't a single number! It is a set of properties which involve discrimination, calibration, reproducibility, prevalence sensitivity, subgroup stability, analytical precision, and the net clinical benefit.

Protein based signatures are appealing due to the fact that proteins represent active biology and not just the DNA sequence. Proteins in circulation may capture inflammatory events, tumor growth, tissue damage or immune recognition as autoantibodies, post transcripts or pathway fragments. Many candidates can be measured using platforms like mass spectrometry or affinity arrays or programmable protein arrays. This width is good for discovery power, but also for statistical risk—a lot of candidate features can produce seemingly powerful classifiers purely by chance.

The condition to be diagnosed should be clearly defined, along with a target population and goals. A test used to detect advanced cancer is not necessarily used to detect cancer in patients who have vague symptoms, benign disease or early lesions. Spectrum bias occurs when the validation data sets are more easily distinguished from one another than real-life patient sets. The same confounders, comorbidities, disease stages and treatment exposures are expected to apply at deployment as in a precision medicine signature.

Sensitivity and specificity are still useful but each requires a decision threshold. A low

threshold will pick up more cases but results in more false positives, while a high threshold will miss infections but will have fewer interventions. The ROC curve summarizes that tradeoff over thresholds. Precision and recall are helpful for rare conditions in terms of providing useful perspective, as positive predictive value may drop dramatically despite high sensitivity and specificity.

Prevalence alters the interpretation of the assay, but not the assay itself. The use of a proteomic panel created in a specialty clinic will have a higher positive predictive value due to the increased prevalence of the diseases. Many false alarms can be produced by the same panel that is used for population screening. Diagnostic performance should then be expressed in terms of the numbers of true positive, false positive, true negative and false negative expected in the intended setting. Those numbers directly relate to clinical workload and patient outcomes.

It is also crucial to calibration. A model could also rank individuals properly, but its probabilities would be systematically too high or too low. Risk thresholds are becoming a more common way of using precision medicine for imaging, biopsy, surveillance or treatment, even if discrimination is acceptable, inaccurate probabilities can mislead decisions. Calibration intercepts and slopes, observed event frequencies, and stability across clinically important subgroups should be assessed by external cohorts, and not just the area under a curve.

Protein microarray studies of tumor associated autoantibodies are a useful example of what can be achieved and what needs to be overcome. Multianalyte antibody signatures were identified in breast cancer, ovarian cancer and colorectal cancer research using work with nucleic acid programmable protein arrays. Panels can work better than single antibodies, because the different tumors evoke different immune responses. However, the more markers added the more flexible the model, and claims of performance must be supported by some form of validation and have some level of independence before they can be translated to real-world scenarios.

Analytical performance is below clinical performance. Protein abundance can be affected by variables encountered during sample processing, storage conditions, freeze thaw history, hemolysis, choice of anticoagulant, batch effects, variation in the protein in the reagents used and instrument drift. If there are differences in the way cases and controls are handled in the lab, a classifier may inadvertently be trained in the handling of the disease, not the biology of the disease. There is a need for standard operating procedures, blinded processing, quality controls, and reproducibility studies between instruments and/

or laboratories of precision medicine.

External validation should be authentic. A new dataset that is processed by the same people and uses the same data extraction procedure can test some overfitting but not transportability. Stronger evidence is provided by an independent institution, different collection period, different operator environment, or population with different clinical characteristics. If performance is lost during such testing, it is informative as it suggests hidden dependence on local practice, sample composition or assay conditions.

A key additional layer is provided by the subgroup analysis. Protein levels may change according to age, sex, ethnicity, kidney, liver function, body composition, inflammatory status and drug exposure. A good signature on average may be a bad signature in a group with some diagnostic inequity. Such issues can be identified prior to deployment using prespecified subgroup estimates, interaction testing, and calibration assessment. External validation should be sufficiently representative to be meaningful, not just aesthetically pleasing.

Biological confidence is provided by orthogonal verification. Once a candidate signal has been identified via protein array, there is the option to retest via a targeted mass spectrometry or an immunoassay. Similar results from different principles of measurement decreases concern about the association being driven by increasing the number of assay architectures. Dissatisfaction can reveal epitope dependence, protein isoforms, interference or unstable post-translational states. Those differences can not only help narrow down the biomarker, but also help to validate it.

Decision curve analysis links statistical and clinical values. Assume that a signature predicts the likelihood of malignancy of an undetermined lesion. Whether or not the score distinguishes between malignant and benign cases is not all that matters; whether or not a change in management based on the score results in a proliferation of unnecessary procedures that is not offset by unacceptable false-negative results is also important. Net benefit methods formalize that balance across decision thresholds, and can indicate whether a model is providing added value beyond the existing clinical predictors.

The ultimate stress test is a prospective evaluation. In real clinical workflows, there are missing data, delayed samples, changing prevalence, changing treatment profiles and clinicians making decisions based on molecular results, history, and imaging. A signature which keeps calibration and utility under those conditions is no longer a promise of the past. Operational issues like turnaround time, cost, assay failure or ambiguous reports that retrospective accuracy cannot identify can also be found by prospective studies.

Repeatability at the patient level should also be investigated in diagnostic studies. One sample taken twice under identical conditions provides information on analytical variation, and repeat sampling over several days will give information on biological fluctuation. The accuracy of the measurement may be acceptable at the population level and yet be misleading on an individual basis if the classification fluctuates wildly around the clinical border. Reliability coefficients and within person variation can thus be used to guide the question of repeat testing.

Prespecification is needed with regard to missing data. Proteomic platforms may not be able to quantify low abundance proteins, and when quantifying the proteins in a sample may result in values that are below detection. Nonsensible imputation can bias associations and the exclusion of participants with missing values can lead to selection bias. The validation procedure should simulate the exact missing data procedure to be used in the clinical setting and it should report on the frequency of results that are not available.

The diagnostic performance also should be compared to other simpler alternatives. A high dimensional signature should be superior to some small number of traditional markers in order to be worth the extra cost and complexity. Nested models can test incremental information and decision analysis can convert that benefit to avoided procedures or earlier diagnosis. Precision medicine only has merit if it translates into an improvement in care, but complexity only has merit if it leads to an improvement in care.

Shift in labelling should be expected following deployment. The percentage and type of disease that is seen by the assay may change because of changes in diagnostic criteria, screening intensity, or treatment. This drift can be detected with time monitoring of the calibration. Any recalibration must maintain an open history of the patient that allows for performance auditing.

The validation of diagnostic performance in proteomics is thus a multidimensional diagnostic validation problem. Discovery leads to the possibility, analytical validation leads to the reliability of measurement, external validation leads to transportability, and decision analysis leads to usefulness. All four layers are essential for the value of precision medicine. A proteomic signature becomes clinically credible when it is accurate across realistic populations, generates interpretable probabilities, is not technically variable and increases a defined decision (as opposed to just producing an attractive performance statistic).