

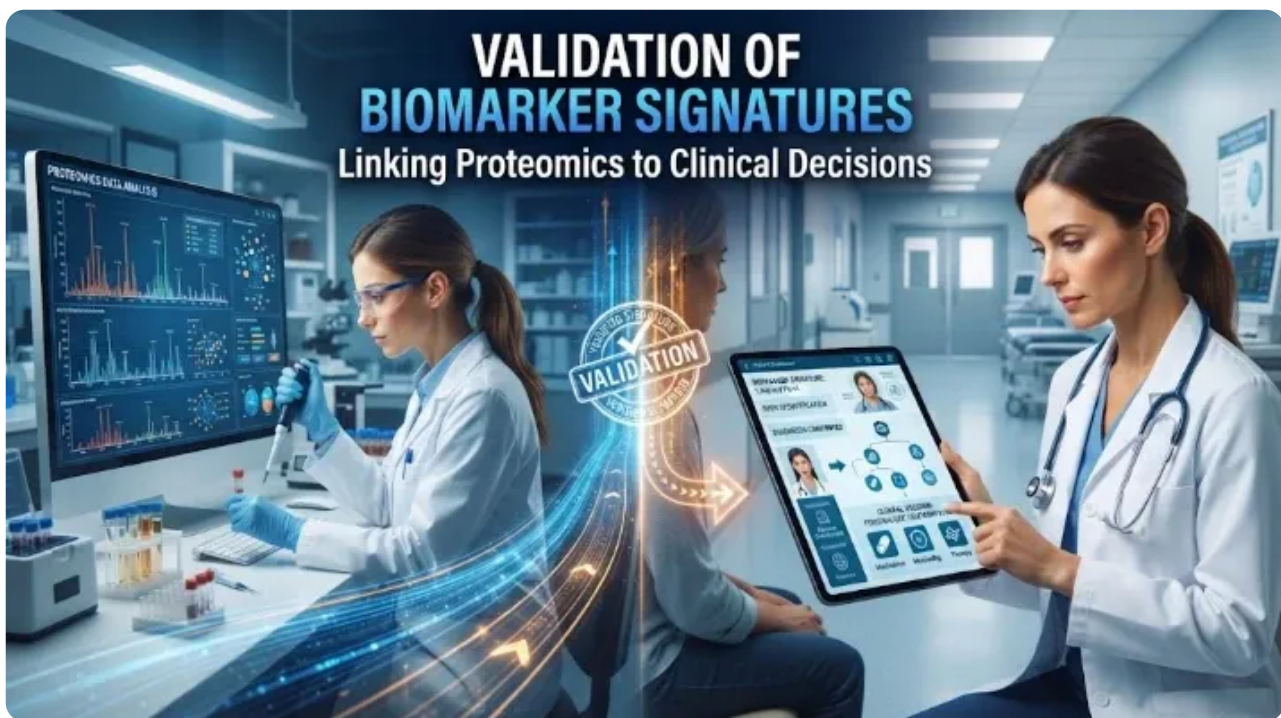
Validation of Biomarker Signatures Linking Proteomics to Clinical Decisions

Sadia Tasmeeem

Jul 19, 2026

Abstract

The article focuses on how proteomic biomarker signatures should be validated before they influence clinical decisions. It begins by distinguishing different intended uses, such as screening, diagnosis, prognosis, and biopsy selection, because performance requirements depend on the decision being made. Multi-marker proteomic panels can capture biological heterogeneity but are vulnerable to overfitting, especially when developed from small high-dimensional datasets. The article emphasizes independent cohorts, realistic disease prevalence, calibration, preanalytical control, subgroup performance, and orthogonal verification across assay technologies. Decision-curve analysis is presented as a way to determine whether a signature improves management rather than simply distinguishing cases from controls. Prospective evaluation is considered the strongest test because real workflows introduce missing data, varying populations, turnaround constraints, and interpretation challenges. The article also highlights locked thresholds, comparison with established clinical models, ongoing post-deployment monitoring, and clear reporting as essential steps for turning a proteomic pattern into a dependable clinical tool.



the molecular profiles that can be created with proteomics are rich, a statistically impressive panel can fail in screening, diagnosis, prognosis, or in deciding on a treatment. The process that distinguishes a repeatable biological signal from the delicate model is called validation. The best programs regard that process as a series of increasingly realistic tests rather than a single separate data set.

The first problem is figuring out what you're going to use it for. The signature, which is used to differentiate cancer from healthy controls, is designed to answer a different question than the test used to determine whether an indeterminate lesion should undergo biopsy. The purpose of a prognostic panel predicting recurrence is quite different. If the clinical role is unclear, then the sensitivity and specificity can be tricky to interpret because the costs of having a false positive result and false negative result vary by decision.

In most cases, proteomic signatures are not based on a single strong marker, but rather are derived from a series of weak signals. This approach can encompass biological heterogeneity particularly in diseases with multiple molecular subtypes. Consider auto antibody panels. A panel of tumor associated autoantibodies is often used alone to detect a subset of cases, and can also be used together to increase sensitivity by combining different immune responses. The benefit is that there is a risk of overfitting as there are more candidates to be tested in a limited number of samples.

The following study using a programmable protein microarray is a good example of both sides of this problem in early breast cancer. Thousands of antigens were screened, and a

number of candidate autoantibodies were selected and tested in an independent serum sample set using a multi marker classifier. The performance was still meaningful but not perfect, and again it was learned that while discovery strength may be great, realistic validation is still required. A test designed to augment imaging should demonstrate effectiveness in a population with current screening dilemmas.

The same level of advancement was achieved in studies of ovarian cancer with high density protein arrays. Large antigen libraries facilitated identification of tumor associated autoantibody signatures, which were then validated independently of each other. These types of studies show the power of proteomics in reducing a large search space to a small panel. It also illustrates why validation cohorts must be distinct from discovery cohorts, as multiple rounds of feature selection on the same samples will give the illusion of accuracy.

Prevalence is used in clinical decision making. Sensitivity and specificity reflect test performance in specific settings and positive and negative predictive value vary according to the frequency of the disease. Many false positives can be produced in general population screening when using a biomarker with an acceptable performance in a high risk referral clinic. It is therefore important that validation is done in a population that is close to the intended use. If not, reported performance numbers can be mathematically correct, and clinically misleading.

The other thing that is neglected is calibration. A model can be used to rank patients according to risk level, and systematically overestimate the probability of it being true. This could result in a good receiver operating characteristic curve and yet mislead treatment thresholds. Calibration plots, calibration slopes, and observed vs. predicted event rates provide information as to the validity of risk estimates. If shared clinical decisions are required, accuracy of the probability counts as much as the ranking.

For proteomics, too, preanalytical control is of equal importance. Proteins may be affected by the time before processing, the storage temperature, the number of freeze thaw cycles, the type of anticoagulant used, hemolysis, or the method of sample collection. A model can learn a history of the laboratory instead of the biology of the disease if cases and controls are treated differently. These conditions should be reproduced and/or intentionally changed to gauge robustness in the process of validation. A panel must have standard operating procedures before it can be used routinely.

It is important that subgroup performance be explicitly tested. Proteomic profiles can vary with age, sex, ancestry, renal function, inflammatory status, exposure to medication and

disease subtype. A good average-signature may fail in a clinically important group. They can be identified by pre-specified interaction analyses or by stratified calibration. This is critical for precision medicine, where average accuracy is not the objective but accurate decision making across diverse patients in a biological sample.

Orthogonal validation decreases reliance on a single assay technology. Once a protein array identifies a protein of interest, a new cohort of candidates can be measured by either an immunoassay or a targeted mass spectrometric method. The direction and magnitude of the marker are consistent, which increases confidence in the marker's biological underpinning. When the failure to reproduce occurs across platforms, it could indicate epitope specificity, interference, and/or protein form instability. This should be investigated, and not simply ignored by selectively reporting.

The last step is decision analysis. Imagine a real proteomic signature, which estimates the likelihood of a pulmonary nodule being malignant. The important question is not if the cases can be shown to have higher scores than controls, but whether or not a particular threshold encourages fewer unnecessary biopsies with no unacceptable missed cancers. Net benefit methods make comparisons with other policies, such as testing everyone or testing no one. This makes the measure of the molecule directly relevant to the clinical impact.

The most demanding test is prospective studies where the investigators are no longer in control of the selection of all aspects of the sample. There are missing values, workflows wander, patients' populations shift, and clinicians interpret the results in the context of imaging, symptoms and previous history. If a panel is valid under those circumstances, it has passed a major milestone from research artifact to clinical tool. Prospective evaluation can also provide insight into whether the test result turnaround is suitable for the clinical practice scenario, and whether the test is cost-effective and interpretable.

With regulatory and implementation issues comes the validation. Assay manufacture, quality control, laboratory accreditation, assay reporting format, and interpretive guidance all affect performance in deployment. Technical accuracy is not enough if there are misinterpretations of probability estimates by clinicians, or if results are obtained after the decision is made relevant to the problem. Implementation science needs to therefore start before, not after, problems of the adoption of the product arise.

The threshold selection should not be coupled with the model development. If you take a look at the validation cohort, and select a cutoff, you are effectively redoing the training process. This leakage is prevented by a prespecified threshold, based on the intended

clinical action. This can then be reported at that locked point, thus providing a clearer picture for clinicians of likely consequences, including sensitivity, specificity, predictive values, and net benefit.

A comparison with established clinical models is also crucial. A proteomic signature can be correct when measured alone but may be of little value when age, imaging, symptoms and traditional laboratory markers are taken into account. It is possible to evaluate incremental value using likelihood measures, reclassification analyses, and decision curves. The fundamental question is whether molecular data will modify the risk assessment sufficiently to change management, as opposed to just being associated with disease.

Monitoring should be ongoing following the clinical introduction, and should be continued for a minimum of 24 months. Sensitivity and calibration may vary with time due to reagent lots, population mix, referral patterns, and treatment standards. False positive rates, false negative events, subgroup performance and assay failures can be tracked within a learning health system. If certain conditions are met that were predetermined, periodic re-calibration might be warranted but any recalibration should be validated before replacing the locked clinical version.

Clinical reporting must be interpretable as well. A multianalyte score should convey the meaning of probability, the population on which it is grounded, and the action threshold was validated for. Reports without decision context can generate false confidence as they present the molecular complexity. Human factors testing with clinicians can uncover confusing displays prior to deployment, and can minimize the difference in how the same information affects care.

That clarity is important for patients and for laboratories because there is a validated score for a defined clinical question and it will not be applied in places that weren't intended.

Proteomics moves clinical decision making towards the molecular state but translation has to be held back. Validation asks the following questions about a locked signature: Is it reproducible, calibrated, transferable, clinically useful, and operationally feasible? Every stage decreases the level of uncertainty associated with a failure mode. If all these layers superimpose, a proteomic biomarker is no longer a mere molecular pattern and it becomes real evidence for a clinical decision.