



**INTERNATIONAL JOURNAL OF
PHARMACEUTICAL SCIENCES**
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: <https://www.ijpsjournal.com>



Review Article

Artificial Intelligence-Enabled Biomarker Discovery in Clinical Biochemistry: Emerging Biomarkers, Machine Learning and Barriers to Clinical Translation

Dr. K. Narendra Kumar Reddy, K. Ravi*, Pedavegi. Mounika, Guntaka. Charita Sree

Department of Pharmacy, Vikas Group of Institutions, Nunna, Vijayawada, Andhra Pradesh- 521212, India.

ARTICLE INFO

Published: 15 Oct 2026

Keywords:

Artificial intelligence, Machine learning, Biomarkers, Clinical biochemistry, Extracellular vesicles, MicroRNAs

DOI:

10.5281/zenodo.23391974

ABSTRACT

Clinical biochemistry now produces large, multidimensional datasets from automated analysers, laboratory information systems, mass spectrometry and molecular assays. Artificial intelligence (AI), and machine learning (ML) in particular, offers tools to combine such measurements into multivariable patterns, to monitor laboratory quality and to prioritise candidate biomarkers. This review examines how AI is used across the pre-analytical, analytical and post-analytical phases of laboratory work and how it is being applied to four groups of emerging biomarkers: pentraxin-3, extracellular vesicle cargo, circulating microRNAs and adipokines. It describes the main learning approaches, the discovery pipeline, the failure modes that most often inflate reported accuracy (overfitting, data leakage, batch effects and confounding) and a staged validation pathway running from analytical validation to post-deployment monitoring. The literature indicates that AI is most defensible where it supports established laboratory processes and combines established tests with emerging markers. High discrimination in a development cohort does not establish biological validity or clinical usefulness. Standardised measurement, leakage-free modelling, calibration, external and prospective validation, transparent reporting and human oversight are the conditions under which AI-assisted biomarker research can move into routine practice, including in resource-variable settings such as India.

INTRODUCTION

Biochemical measurements underpin diagnosis, risk stratification, therapeutic monitoring and screening. Over the past two decades the volume and variety of laboratory data have grown sharply, because high-

throughput analysers, middleware, electronic records and omics platforms now generate results that are too numerous and too interrelated for simple threshold rules to exploit fully.^{1,2}

***Corresponding Author:** K. Ravi

Address: Department of Pharmacy, Vikas Group of Institutions, Nunna, Vijayawada, Andhra Pradesh- 521212, India

Email ✉: ravikorrapati11@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



Traditional biomarker research tests one analyte, or a few, against a defined outcome. This approach has produced most of the tests in routine use, but many diseases involve interacting inflammatory, metabolic, vascular and molecular processes, so useful information is often spread across several modestly informative measurements instead of a single strong one. Machine learning addresses this by estimating relationships between many variables and an outcome, grouping patients without predefined labels, flagging unusual results and ranking candidate markers.³

AI has also entered the laboratory itself. A European survey of laboratory professionals reported growing interest and use of AI tools alongside uneven readiness and a perceived need for training and governance.⁴ Reviews of clinical chemistry applications describe uses ranging from specimen assessment and quality control to result interpretation.^{1,2}

The enthusiasm needs tempering. A model can absorb bias from its training data, learn laboratory-specific artefacts, overfit a small cohort, or report excellent internal accuracy that disappears in a new population.^{3,5} In the laboratory setting, guidance stresses that validation, implementation and monitoring of ML tools are separate, ongoing responsibilities and not one-off statistical exercises.^{5,6}

This review has three aims: (i) to explain the ML concepts that matter for biochemical data; (ii) to map AI applications across the laboratory workflow; and (iii) to examine four emerging biomarker groups as case studies of the analytical and biological problems that AI cannot remove by itself. Throughout, the emphasis is on what evidence would be needed before an AI-assisted marker or tool is used in patient care.

2. LITERATURE SEARCH APPROACH

This is a narrative review and not a systematic review. Sources were identified through targeted searches of PubMed-indexed and publisher-hosted literature and of official publications from the World Health Organization and the United States Food and Drug Administration, using combinations of the terms artificial intelligence, machine learning, clinical chemistry,

laboratory medicine, biomarker, pentraxin-3, extracellular vesicles, microRNA, adipokines, validation and governance. Recent reviews, methodological and reporting guidance, laboratory-practice papers and regulatory documents were prioritised because the aim was conceptual integration and not pooled estimation. The reference set is therefore selective and representative, not exhaustive, and no quantitative performance figures have been pooled or compared across studies.

3. MACHINE LEARNING CONCEPTS RELEVANT TO BIOCHEMICAL DATA

3.1 Terminology

Artificial intelligence is the broad field of computer systems that perform tasks associated with pattern recognition, prediction or decision support. Machine learning is the subset in which algorithms learn relationships from data and not from hand-written rules, and deep learning uses multilayer neural networks that are most useful for unstructured data such as images and signals.¹

3.2 Learning paradigms

Supervised learning requires labelled outcomes and is used for disease classification, outcome prediction and estimation of continuous variables. Unsupervised learning looks for structure without labels, for example clustering patients into metabolic phenotypes or reducing dimensionality for visualisation. Semi-supervised approaches may suit laboratories in which results are abundant but confirmed diagnoses are scarce.^{1,3}

3.3 Choosing model complexity

Complex models are not automatically better. Penalised logistic regression and other transparent models are strong baselines, particularly with limited sample sizes, and often match more flexible methods once calibration, stability and maintenance are considered.^{3,5} Tree ensembles, support vector machines and neural networks become attractive where interactions, non-linearity or very high dimensionality are present, but they demand more data and more careful validation (Table 1).



Table 1: Common Machine Learning Approaches in Biochemical Biomarker Research

Approach	Typical use	Strengths	Principal risks
Penalised / logistic regression	Risk models; panels with few markers	Interpretable; stable in small samples; easy to recalibrate	May miss interactions; needs careful handling of correlated predictors
Tree ensembles (random forest, gradient boosting)	Tabular laboratory and clinical data	Capture non-linearity and interactions; handle mixed variables	Overfitting if tuned repeatedly; less transparent; miscalibration
Support vector machines	High-dimensional omics classification	Effective when features outnumber samples	Sensitive to scaling and kernel choice; limited probability output
Neural networks / deep learning	Images, spectra, signals; large multimodal data	Learn complex representations	Large data requirement; opacity; vulnerability to dataset shift
Clustering and dimensionality reduction	Phenotype discovery; exploratory analysis	No labels needed; reveals structure	Clusters may not be clinically meaningful or reproducible

Compiled by the author from general ML methodology as discussed in references ³ and ⁵.

3.4 Feature selection and explainability

In biomarker discovery the candidate features (proteins, metabolites, nucleic-acid species) often far exceed the number of participants, so noise can be fitted very closely. Feature selection, scaling and imputation must therefore be performed inside each training fold and never on the full dataset before evaluation.³ Explainability methods, which show how much each input contributed to a prediction, are useful for model inspection but do not demonstrate causation: a variable can be influential because it tracks a confounder or a technical batch.^{3,5}

3.5 Data preparation and harmonisation

Laboratory data are not a neutral record of biology. Results depend on analyser platform, calibrator lot, reagent generation, unit conventions and local reference intervals, and a model trained on pooled data can therefore learn the laboratory of origin. Preparation steps such as unit conversion, platform-specific scaling, handling of values below the limit of detection and correction of batch effects must be documented and, importantly, fitted only on training data so that information about the test set does not enter the model.^{3,5}

Missing data need particular thought. In clinical practice a test is usually ordered because a clinician suspects a condition, so the absence of a result is itself informative and cannot be treated as random. Simple deletion or mean imputation may bias estimates, and

the chosen strategy (for example model-based imputation fitted within training folds, or an explicit indicator for "not tested") should be stated and its influence examined.⁷

Repeated measurements from the same person create a further trap. If several samples from one patient are split between training and test sets, the model can recognise the individual and not the disease, giving falsely high accuracy. Splitting must be done at the level of the patient, and where drift is expected a temporal split, with the most recent period held out, gives a more realistic estimate than random partitioning.^{5,6}

3.6 Performance metrics and their interpretation

No single metric describes a model. Sensitivity and specificity characterise a test at a chosen threshold but do not say how likely a positive result is to be correct, which depends on prevalence. As a hypothetical illustration, a test with 90% sensitivity and 90% specificity applied where the condition affects 1% of people would give a positive predictive value of only about 8%, because most positive results would come from the large unaffected group. This is why screening-type applications must be evaluated in the population of intended use and not in enriched case-control samples.

The area under the receiver-operating-characteristic curve summarises ranking ability across thresholds but is insensitive to calibration, meaning the agreement between predicted and observed risk. Calibration should be shown graphically and summarised, for example by calibration slope and intercept, and the Brier

score combines discrimination and calibration in one number. For rare outcomes, precision-recall curves are often more informative than ROC curves. Decision-curve analysis estimates net clinical benefit across a range of decision thresholds and helps answer whether using the model would do more good than harm compared with treating everyone or no one.^{5,7}

For laboratory quality-control tools the relevant measures differ: false-alert rate, positive predictive value of alerts, time to detection of a true shift and the workload generated.⁸ Whatever the application, point estimates should be accompanied by confidence intervals, and the comparison should always be against a sensible reference such as current practice or a simple regression model, since a complex model that merely equals a simple one has not earned its additional maintenance burden.⁵

3.7 Interpretability and explainability tools

Clinicians, laboratory managers and regulators are reluctant to rely on a model whose reasoning is entirely opaque, so interpretability is both a scientific and a practical requirement. Two broad strategies exist. The first is to use inherently transparent models, such as penalised regression or small decision rules, whose coefficients or branches can be read directly. The second is to apply post-hoc explanation methods to a more complex model, for example permutation importance, which measures how much performance falls when one input is scrambled, partial-dependence or accumulated-local-effect plots, which show the average effect of an input on the prediction, and attribution methods that apportion an individual prediction among its inputs.^{3,5}

These tools must be used with caution. Correlated inputs, which are common among laboratory analytes, share importance in ways that are easily misread; an input may appear unimportant because a correlated partner carries the same information. Importance rankings can also change between resamples of the same dataset, so their stability should be examined before any marker is highlighted as key. Most importantly, an explanation describes the model and not the underlying biology: a high attribution to a marker may reflect confounding or a technical batch, and only independent experimental or clinical evidence can show that the marker is involved in the disease process.³

For routine laboratory use, explanations should be matched to the user. A laboratory scientist reviewing a flagged result may need to know which related analytes and which earlier results prompted the flag, whereas a clinician needs to know the intended use, the main limitations and the level of uncertainty. Presenting a single probability with spurious precision encourages automation bias; presenting a range, the reason for the alert and a clear statement of what the tool was not designed to do is safer.⁶

4. AI ACROSS THE LABORATORY WORKFLOW

Laboratory errors and opportunities differ by phase, so applications are best organised as pre-analytical, analytical and post-analytical (Table 2).^{1,2}

4.1 Pre-analytical phase

The pre-analytical phase accounts for many laboratory errors and covers test ordering, patient preparation, collection, transport and processing. ML tools can screen test requests for unusual combinations, flag samples at risk of being unsuitable and support identity checks. A recent report described a two-step AI model for automated serum quality assessment, a task traditionally dependent on subjective visual inspection for haemolysis, icterus and lipaemia.⁹ Such systems must be validated across instruments, tube types, lighting and local sample populations, because performance in one site does not guarantee performance in another. Their safest role is to prioritise review and add information and not to override established specimen-rejection criteria.



4.2 Analytical phase

Internal quality control and external quality assessment remain the foundation of analytical assurance. AI adds a complementary layer. Patient-based real-time quality control uses the continuous stream of routine patient results to detect shifts in assay performance, and ML versions can examine multivariable patterns to identify analytical or pre-analytical error.⁸ A particular difficulty is that patient populations change with new clinical services, referral patterns and seasons; a monitoring algorithm must separate such expected variation from genuine analytical deterioration. Anomaly detection can also find results that are individually plausible but inconsistent with related analytes or with the patient's own history, which may reveal specimen mix-ups, interference or entry errors. Evaluation should report alert burden and positive predictive value as well as sensitivity, because frequent unhelpful alerts cause fatigue and wasted investigation.^{6,8}

4.3 Post-analytical phase

Rule-based autoverification is already common. ML could extend it by considering previous results, related

analytes and temporal patterns, with the aim of reducing manual review without releasing implausible results. Decision-support systems can highlight results that merit clinical attention or suggest follow-up testing, and a review of laboratory decision support has examined their clinical, organisational and financial impacts, with the strongest emphasis on whether patient outcomes actually change.¹⁰ Longitudinal interpretation is a particular opportunity: a change from a person's own baseline may matter even when the value remains inside the population reference interval, but the model must be able to distinguish biological change from a change of method, instrument or medication.

Two cross-cutting risks apply. Automation bias, in which users over-trust a precise-looking probability, is addressed by displaying uncertainty and keeping human accountability explicit. Equally, technical accuracy is not a sufficient reason for deployment; implementation studies should show benefit in turnaround time, appropriateness of testing, missed abnormalities or workload.^{6,10}

Table 2: Summary of Laboratory AI Applications, Benefits and Safeguards

Phase	Application	Potential benefit	Key safeguard
Pre-analytical	Serum quality assessment; test-re-request review; identity checks ⁹	Fewer unsuitable samples; consistent assessment	Local validation; established rejection rules retained
Analytical	Patient-based real-time quality control; anomaly detection ⁸	Earlier detection of drift or error	Used with, not instead of, internal QC and EQA; alert-burden monitoring
Analytical	Predictive instrument maintenance ²	Less downtime	Maintenance records audited
Post-analytical	ML-assisted autoverification; decision support ¹⁰	Reduced manual review; better interpretation	Validation against expert review and safety rules; uncertainty displayed
Post-analytical	Longitudinal trend analysis	Detection of change within reference range	Adjustment for method and medication changes

Compiled by the author from the cited references.

4.4 Introducing a laboratory AI tool: shadow operation and monitoring

A responsible implementation does not begin with automatic action. A common and prudent approach is to run the tool in shadow mode, in which it produces outputs that are recorded but not acted on, while the existing process continues unchanged. This allows its

decisions to be compared with those of experienced staff and with final outcomes, its alert rate and workload to be measured, and thresholds to be adjusted before it influences any result that reaches a patient.^{5,6}

Once live, performance has to be monitored as carefully as the analytical methods themselves. Useful measures include the proportion of results flagged, the proportion of flags confirmed on review, the distribution of the model's inputs and outputs over time, and



agreement with manual review in periodic audits. A sustained change in these indicators may signal drift in the patient population, a new reagent lot or method, or a software change, and should trigger investigation. Pre-specified rules for recalibration, retraining and withdrawal, together with version control and an audit trail of which model version produced which output, are essential.⁶

Staff training and communication complete the picture. Laboratory personnel need to understand what the tool does and does not do, how to override it and how to report problems, and the laboratory should keep documented accountability for every decision so that responsibility never passes silently to the algorithm. Lorde and colleagues, in their review of machine learning for patient-based real-time quality control, emphasise that such methods are complementary to, and not a substitute for, conventional quality-control procedures, and this principle extends to the other AI applications described here.⁸

5. EMERGING BIOMARKERS AS CASE STUDIES

An emerging biomarker is not necessarily a newly discovered molecule; it is a candidate with biological plausibility that has not reached routine clinical use. Their main attraction for AI is that panels of moderately informative markers may outperform any single one if their errors are not strongly correlated. The decisive question is incremental value: does the marker improve diagnosis, risk classification or management beyond established tests and clinical variables? The four groups below illustrate distinct problems (Table 3).

5.1 Pentraxin-3 (PTX3)

PTX3 is a long pentraxin produced by endothelial cells, fibroblasts, myeloid cells and other cell types in response to inflammatory signals. It differs from the classical short pentraxin C-reactive protein in its cellular sources and kinetics, which is why it has been proposed as a marker of local vascular and tissue inflammation.¹¹ It has been studied in coronary syndromes, heart failure and vascular disease, where associations with severity and outcome have been

described, although these data do not support treating PTX3 as a replacement for established markers.¹¹

The role most consistent with the evidence is as a component of multimarker models. An ML framework could evaluate PTX3 together with cardiac troponin, natriuretic peptides, renal function, conventional inflammatory markers and clinical variables, and the analytic question would be whether PTX3 adds information after those predictors are in the model. Confounding is the principal hazard: PTX3 rises with infection, systemic inflammation and other comorbidities, so a model trained without suitable comparator groups may simply learn non-specific inflammation. Assay platform, calibration and sample stability must also be controlled, because differences in measurement can propagate into model performance. Unsupervised methods could additionally explore whether inflammatory phenotypes exist in which PTX3 is particularly informative.

5.2 Extracellular vesicles (EVs) and exosomal cargo

EVs are membrane-bound particles released by cells into biological fluids, carrying proteins, lipids and nucleic acids that reflect their cell of origin, which makes them attractive for minimally invasive diagnostics.¹² The central obstacle is methodological heterogeneity. Isolation by ultracentrifugation, precipitation, size-exclusion chromatography or affinity capture recovers different populations, and sample handling and storage change the measured profile. Consensus guidance (MISEV2023) calls for rigorous characterisation and transparent reporting,¹³ and recent work argues for integrating EV-specific standards with general molecular-measurement standards for nucleic-acid biomarker studies.^{14,15} Even basic concentration measurements by flow cytometry have required dedicated standardisation efforts.¹⁶

For AI this matters directly. If cases and controls were collected or processed differently, an algorithm can classify the processing pattern and not the disease. ML is most helpful after analytical quality is controlled, for example in prioritising a small panel from high-dimensional proteomic or nucleic-acid data for targeted validation in an independent cohort using a predefined assay.^{3,15} The final panel must also be realistic for a



routine laboratory in terms of turnaround time, precision and cost.

5.3 Circulating microRNAs

MicroRNAs are short non-coding RNAs that regulate gene expression and can be measured in blood and in EVs, giving interest in non-invasive diagnosis and monitoring across cancer, cardiovascular, metabolic, infectious and neurological disease.^{17,18} Because several microRNAs often change together, signatures are well suited to multivariable modelling, whereas individual microRNAs frequently lack specificity.

Measurement is the weak point. Collection tube, haemolysis, RNA extraction, reverse transcription, amplification platform, endogenous controls and normalisation strategy can each alter observed values, and between-laboratory differences can exceed the biological effect of interest.¹⁸ Strict quality control, technical replicates, transparent normalisation and batch assessment should precede model building, and feature selection must occur only within training data. Linking selected microRNAs to a coherent biological pathway can increase plausibility, although pathway enrichment does not establish causality. Multi-centre external validation is particularly important because signatures that work in one centre often fail in another.

5.4 Adipokines

Adipose tissue is an endocrine organ secreting adipokines such as adiponectin, leptin, resistin and omentin, which have been investigated in obesity, insulin resistance, diabetes, cardiovascular disease and metabolic syndrome.^{19,20} Their appeal for AI is that they capture components of metabolic and inflammatory physiology that routine glucose, lipid, liver and renal tests do not, so a model can integrate them with anthropometric and clinical variables.

Concentrations are strongly influenced by age, sex, body composition, nutritional state, renal function, inflammation and medications.^{19,20} These influences generate confounding: a model may link an adipokine to diabetes only because both are related to obesity. Comparator groups must match the intended use; a diagnostic model for diabetes should be tested against obese individuals without diabetes, not only lean controls. Clustering may reveal metabolic phenotypes defined by combinations of insulin resistance, inflammation and adipose dysfunction, but a cluster is not automatically a disease subtype unless it is reproducible, biologically interpretable and linked to outcomes that matter to patients.

Table 3: Emerging Biomarker Groups: Rationale, Confounders and Ai-Specific Considerations

Biomarker	Rationale	Main confounders / analytical issues	AI-specific consideration
PTX3 ¹¹	Inflammatory protein produced by vascular and myeloid cells	Infection, systemic inflammation, comorbidity; assay platform	Test incremental value over troponin, natriuretic peptides and CRP with realistic comparators
EV cargo ^{12,13}	Cell-of-origin proteins and nucleic acids in biofluids	Isolation method, storage, handling; particle characterisation ¹⁶	Models may learn processing batch; require independent cohort and predefined assay
MicroRNAs ^{17,18}	Regulatory RNAs with multi-marker signatures	Haemolysis, extraction, platform, normalisation	Leakage-free feature selection; multicentre external validation
Adipokines ^{19,20}	Endocrine signals from adipose tissue	Age, sex, adiposity, renal function, drugs	Match comparators to intended use; validate clusters clinically

Compiled by the author from the cited references.

5.5 Combining the marker classes: a staged, multi-modal perspective

The four groups discussed above differ greatly in cost, turnaround time and maturity. Routine chemistry, glycated haemoglobin, lipid profile, renal and liver tests

and established cardiac markers are inexpensive, standardised and widely available, whereas EV isolation, microRNA profiling and some adipokine immunoassays are costlier, slower and less standardised. A practical way to use AI is therefore as a staged system in which a model built on routine data identifies patients in whom the added information from an expensive



emerging assay is most likely to change management. In this arrangement the emerging biomarker is not applied to everyone, and its incremental value is judged within the subgroup where it is actually used. This is the author's perspective and not an established clinical pathway, and it would require prospective testing like any other strategy.

5.6 Illustrative design for testing the incremental value of PTX3

To make the preceding principles concrete, Table 4 outlines how an incremental-value study of PTX3 could be designed so that most of the hazards described in this review are addressed in advance. It is an illustrative design and not a protocol that has been carried out.

Table 4: Illustrative Design Elements for a PTX3 Incremental-Value Study

Element	Proposed approach
Clinical question	Does PTX3 improve prediction of a predefined cardiovascular outcome at a fixed follow-up beyond clinical variables, troponin and natriuretic peptide?
Population	Consecutive patients with suspected acute coronary syndrome or heart failure from at least two centres, recruited without selecting only extreme cases
Comparator and confounder groups	Include patients with infection or systemic inflammation and with renal impairment, so that the model cannot succeed by detecting non-specific inflammation ¹¹
Pre-analytical and analytical control	Standardised collection and storage, one validated assay platform, randomised placement of samples across assay runs, documented precision and stability
Baseline model	Penalised regression using clinical variables and established markers; PTX3 then added as the only change
Comparison of models	Change in discrimination, calibration, reclassification and net benefit, each with confidence intervals
Sample size	Planned from the expected number of outcome events relative to the number of candidate predictors, before recruitment ⁷
Validation	Temporal hold-out within the development centre and a fully independent external centre, with the model locked before testing
Reporting	TRIPOD+AI checklist; assay details reported in sufficient detail for replication ⁷

Prepared by the author as an illustration of the principles discussed in this review.

6. MACHINE LEARNING PIPELINE FOR BIOMARKER DISCOVERY AND ITS PITFALLS

A defensible discovery pipeline starts with a defined clinical question and outcome, followed by data curation, partitioning into development and independent test data, preprocessing and feature selection within the training framework, model fitting with tuning inside nested resampling, and finally a single evaluation on data untouched during development.^{3,5} Reviews of ML for omics-based biomarker discovery highlight that deviations from this sequence are the commonest reason for over-optimistic reports.³

Overfitting occurs when a model fits noise specific to the development data; it is promoted when many algorithms and hyperparameters are tried and the best result

is reported without a held-out test set. Data leakage arises when information from validation or test samples influences development, for example when the most predictive markers are selected on the entire dataset before cross-validation. Batch effects arise when technical factors, such as run date, operator or reagent lot, correlate with the outcome. Spectrum bias arises when controls are healthy volunteers although the clinical problem is differentiation from benign or inflammatory conditions. Table 5 summarises these failure modes and countermeasures, which are also among the points that reviewers and examiners most often raise.

Sample size and data handling deserve explicit attention. Discovery cohorts of a few hundred participants cannot support a very large number of candidate features and a flexible model at the same time, so the planned number of predictors, the expected number of outcome events and the intended model complexity should be reconciled before data collection and not



afterwards. Missing values should be handled by pre-specified methods applied within training data, and imbalance between outcome classes, which is typical for rare diseases and screening, should be addressed in both modelling and metric choice, because accuracy alone can look high when the outcome is uncommon. Where possible, the full analysis code and pipeline should be shared or at least documented so that others can reproduce each step.^{3,7}

Interpretation of results should remain proportionate. A narrow confidence interval around a cross-validated estimate describes the development data only; it says little about performance in a different centre. Likewise, a panel of selected markers should be described as hypothesis-generating until it has been measured again, by a locked assay and a locked model, in new participants.⁵

Table 5: Common Failure Modes in Ai-Assisted Biomarker Studies and Mitigation

Failure mode	How it arises	Mitigation
Overfitting	Many features, few samples; repeated model selection	Penalisation; nested cross-validation; independent test set; sample-size planning
Data leakage	Selection, scaling or imputation using full dataset	All preprocessing inside training folds; code and pipeline audit
Batch effect	Cases and controls processed differently	Randomised processing; balanced batches; batch-adjusted analysis; pre-analytical standardisation
Spectrum bias	Unrealistic comparator groups	Include clinically relevant differential diagnoses and the intended-use population
Confounding	Marker tracks age, adiposity, renal function or drugs	Adjust and stratify; report incremental value over clinical model
Miscalibration	Good ranking but inaccurate probabilities	Report calibration; recalibrate locally
Dataset shift / drift	Changes in assay, population or practice after deployment	Monitoring thresholds; scheduled revalidation; change-control plan

Compiled by the author from the cited methodological literature^{3,5,6}.

6.1 A leakage-free workflow in practice

The safest way to avoid leakage is to define the order of operations before any modelling begins and to treat the test set as sealed. Table 6 sets out a sequence that

can be followed in most discovery projects. The key idea is that every data-dependent decision, including scaling, imputation, feature selection and tuning of hyperparameters, is made using training data only, usually inside nested cross-validation, and that the final model is evaluated once on data that played no part in those decisions.^{3,5}

Table 6: Sequence of Steps for a Leakage-Free Biomarker Modelling Workflow

Step	Action	Common error to avoid
1	State the clinical question, outcome and intended use	Defining the outcome after seeing the data
2	Set aside a test set at patient level (and by time or site if possible)	Splitting samples at random so one patient appears in both sets
3	Fit scaling, imputation and batch correction on training data only	Normalising the whole dataset before splitting
4	Select features inside each cross-validation fold	Choosing the top markers on all samples, then cross-validating
5	Tune hyperparameters with nested re-sampling	Reporting the best of many tuned models without correction
6	Compare with a reference model (clinical variables and established markers)	Comparing only with chance or with healthy controls



Step	Action	Common error to avoid
7	Evaluate once on the sealed test set; report discrimination, calibration and intervals	Re-opening the test set after disappointing results
8	Lock the model and test it externally	Retraining on the external data and calling it validation
9	Report according to TRIPOD+AI and relevant laboratory standards ^{7,13}	Omitting preprocessing and exclusion details

Compiled by the author from the methodological literature cited in this review.

7. VALIDATION, REPORTING AND CLINICAL TRANSLATION

Validation is best viewed as a sequence and not as a single statistical step. Analytical validation asks whether the biomarker can be measured reliably, covering precision, accuracy, measurement range, interference, stability and reproducibility, and is especially critical for emerging markers whose assays are not yet standardised. Biological plausibility addresses whether the marker reflects the disease process. Clinical validation then evaluates discrimination between clinically relevant groups, external validation tests transportability to other institutions, platforms and populations, and prospective evaluation tests behaviour in the real workflow.^{5,6}

Model evaluation should report discrimination (for example the area under the receiver-operating-characteristic curve) together with calibration, since a model can rank patients well while giving inaccurate absolute probabilities, which matters when probabilities guide decisions. Incremental value over a baseline model of clinical variables and established markers should be demonstrated explicitly. External validation should use

genuinely independent data, preferably from another laboratory, analyser platform or population.⁵

Reporting should follow TRIPOD+AI, the updated guidance for studies that develop or evaluate prediction models using regression or machine learning; it asks for clear description of participants, data handling, missing data, model development, validation and performance so that readers can judge credibility.⁷ For EV and nucleic-acid work, MISEV-based reporting should be added.^{13,14}

Clinical utility is a further step again. A tool is useful only if it changes decisions or processes for the better, with outcomes such as earlier diagnosis, fewer unnecessary tests, shorter turnaround time, less manual review, improved risk classification or better patient outcomes.^{6,10} High predictive performance is therefore necessary for many applications but is never sufficient evidence of usefulness.

7.1 A checklist for authors, reviewers and examiners

Because the same weaknesses recur across AI-based biomarker papers, a short checklist can help authors design studies and help reviewers and examiners appraise them. Table 7 lists questions that can be applied to almost any report, with the aspect of the translational pathway to which each relates.

Table 7: Questions for Appraising an Ai-Assisted Biomarker Study

Question	Aspect	Relevant guidance
Is the assay analytically validated (precision, range, interference, stability)?	Analytical validity	Laboratory validation practice ^{5,6}
Were cases and controls collected, processed and measured in the same way?	Batch and pre-analytical bias	MISEV for EV studies ¹³
Do the comparator groups reflect the real differential diagnosis?	Spectrum bias	TRIPOD+AI ⁷
Was all data-dependent processing confined to training data?	Leakage	Good modelling practice ³
Is the number of candidate features compatible with the sample size?	Overfitting	TRIPOD+AI ⁷



Question	Aspect	Relevant guidance
Is calibration reported, and is there a comparison with a baseline model?	Model performance	TRIPOD+AI ⁷
Was the model tested in an independent centre or platform?	External validity	Laboratory ML validation ⁵
Are subgroup results (age, sex, setting) shown?	Fairness	TRIPOD+AI ⁷
Is there evidence that use of the tool changes decisions or workflow?	Clinical utility	Implementation guidance ^{6,10}
Is there a plan for monitoring drift and recalibration?	Post-deployment safety	Laboratory ML monitoring ⁶

Compiled by the author from the cited reporting and laboratory guidance.

7.2 Research evidence versus regulatory evidence

It is useful to distinguish the evidence that satisfies a research journal from that needed when a tool will be used on patients. A published model may be a perfectly good research finding yet fall far short of what is needed for use in care, where the intended use, input data, performance, limitations and procedures for change must be specified and controlled. Regulatory and governance documents emphasise a lifecycle approach, with attention to data management, representativeness of data, independence of test sets, transparency to users and a plan for managing modifications after release.²¹⁻²³ WHO guidance adds an ethical frame in which safety, accountability, inclusiveness and human oversight are required alongside technical performance.^{24,25}

For students and early-career researchers this distinction has a practical consequence. A discovery study should state clearly that its aim is hypothesis generation or early validation and should avoid language implying readiness for clinical use. Doing so is more honest and also less likely to attract the reviewer objection that conclusions go beyond the data.

8. CLINICAL APPLICATION DOMAINS

8.1 Cardiovascular and metabolic disease

Cardiovascular disease is a natural setting for multi-marker modelling because myocardial injury, haemodynamic stress, renal function, inflammation and metabolism interact. Established markers such as cardiac troponins and natriuretic peptides have well-defined roles, so the relevant test of any AI model including an emerging marker like PTX3 is whether it improves on current clinical and laboratory practice,

not merely whether it separates cases from healthy controls.¹¹ Specificity limits of inflammatory markers mean that comparator groups should reflect the real differential diagnosis.

In metabolic disease, routine glucose, glycated haemoglobin, lipid and liver tests can be combined with adipokines and clinical characteristics to identify risk patterns, and longitudinal models may detect deterioration before a conventional threshold is crossed.^{19,20} Treatment effects complicate prediction because a laboratory value reflects both disease biology and drug exposure; models should include treatment information where relevant and be retested after changes in standard care. Models built in specialist clinics should not be assumed to generalise to primary care.

8.2 Oncology, inflammation and liquid biopsy

Oncology is a very active area for AI because tumour biology produces high-dimensional molecular signals, and liquid biopsy offers circulating nucleic acids, proteins and EVs without tissue sampling.^{12,17} The principal danger is cohort selection: a classifier trained against healthy controls can look excellent and then fail against benign or inflammatory conditions that resemble the target disease. Pre-analytical variation in collection, processing time, storage and extraction can also create differences that the model learns in place of disease.¹⁵ For screening, evaluation must be done in the population where the test will be applied, where prevalence is low and false positives carry real consequences. The same logic applies to inflammatory and autoimmune disease, where many markers lack disease specificity.



8.3 Illustrative scenarios

Three short hypothetical scenarios show how the principles in this review apply in practice. They are constructed for teaching and do not describe real patients or studies.

Scenario 1: a quality-control alert. A patient-based quality-control algorithm signals a possible upward shift in a routine analyte on one analyser late in the evening. The laboratory scientist does not simply accept or dismiss the alert. Internal quality-control results, calibration records and recent reagent or maintenance events are reviewed, the pattern is compared with the other analyser, and a change in the case mix (for example a large batch of samples from one ward) is considered as an alternative explanation. The algorithm has therefore functioned as a prompt for investigation and not as an independent decision-maker, which is the role supported by the literature.^{6,8}

Scenario 2: an inflammatory marker in a mixed population. A multimarker model that includes PTX3 gives a high-risk output for a patient admitted with a respiratory infection and mild renal impairment. Because PTX3 can rise with infection, inflammation and comorbidity, the output cannot be interpreted as evidence of a cardiovascular event unless the model was developed and validated in a population that included such patients as comparators. A reviewer would ask for the composition of the training data, the incremental value of PTX3 over troponin and natriuretic peptide, and external validation before accepting the result as meaningful.^{5,11}

Scenario 3: an adipokine panel in obesity. A research group reports that a panel of adipokines separates people with type 2 diabetes from healthy controls with high accuracy. On inspection, the controls were lean and the cases were mostly obese. The panel may largely be measuring adiposity. A more convincing design would compare people with diabetes with obese people without diabetes, adjust for age, sex and renal function, and test whether the panel adds information beyond glucose, glycated haemoglobin and lipids.^{19,20}

In each scenario the correct response is the same: ask what the model was trained on, what it was compared with, whether it adds anything beyond what is already

measured, and whether the result has been confirmed outside the development setting.

9. ETHICS, REGULATION, DATA GOVERNANCE AND THE INDIAN CONTEXT

AI systems process sensitive health information and influence clinical decisions, so privacy, transparency, accountability, fairness and human oversight are core requirements. The WHO guidance on ethics and governance of AI for health sets out principles including protection of autonomy, promotion of safety and the public interest, transparency and explainability, accountability, inclusiveness and sustainability,²⁴ and its later guidance on large multi-modal models addresses risks such as inaccurate output and inappropriate data use.²⁵

Regulators now consider the whole life cycle of AI-enabled medical software. Good machine learning practice principles issued jointly by the US FDA, Health Canada and the UK MHRA emphasise data management, model development, testing and monitoring,²¹ a further joint document sets out transparency principles for ML-enabled devices,²² and FDA guidance describes how a predetermined change control plan can be used to manage planned modifications to AI-enabled device software.²³ These documents address medical devices rather than laboratory-developed research tools, but their principles of documented data, validation, transparency and controlled change are directly useful to laboratory implementers.

Practical implementation also depends on integration with laboratory information systems, middleware and electronic records, on cybersecurity measures such as access control and audit trails, and on economic feasibility, since software, computing, validation, training and maintenance add cost and emerging biomarkers add reagent and instrument costs.⁶ An accurate model whose required assay is unaffordable in the intended setting has limited value.

In India, laboratories vary widely in infrastructure, analyser platforms and digital maturity. In the author's view, models developed in high-income healthcare systems should not be assumed to perform identically in Indian populations, where body composition, disease prevalence, nutrition and pre-analytical conditions can



differ, and local validation and affordability analysis should therefore accompany any adoption. Governance is best multidisciplinary, involving laboratory scientists, clinicians, pharmacists, statisticians, data scientists, quality managers and information-technology staff.

9.1 Data protection, consent and data sharing

Laboratory results linked to demographic and clinical information are sensitive personal data. Studies should collect the minimum data needed, de-identify records before analysis, obtain ethics approval and, where required, consent for secondary use, and keep an audit trail of who accessed which data.²⁴ Sharing data across centres is essential for external validation but raises governance questions; approaches in which models are trained or evaluated without moving patient-level records between institutions are one option under discussion. Researchers and laboratories in India should also consult the current national data-protection legislation and the national guidance on the use of AI in biomedical research and healthcare, because requirements evolve and these documents were not within the verified set of sources used for this review.

9.2 Quality systems, accreditation and local validation

Many Indian laboratories operate under quality systems based on ISO 15189 and seek accreditation through the national accreditation body. In the author's view, any AI component that influences how a result is verified, flagged or released should be introduced through the laboratory's existing change-control, validation and documentation procedures in the same way as a new analyser or a revised autoverification rule, and not as an informal add-on.⁶ Local verification should include comparison against expert review, assessment on the laboratory's own specimen mix and clear criteria for switching the tool off. Laboratories also differ greatly in size, platform, connectivity and staffing, so a model validated in a tertiary-care centre should not be assumed to work in a smaller or semi-urban laboratory without checking.

9.3 Role of the pharmacist and the pharmacy student

Pharmacists already use laboratory data to monitor renal and hepatic function, adjust doses, review therapeutic drug monitoring results and watch for adverse drug effects. As decision-support tools become more common, pharmacists are well placed to contribute to their governance, for example by checking that medication effects on laboratory values are represented in a model, by interpreting alerts in the light of the patient's drug history and by helping to define safe use limits. Pharmacy education should therefore include the basics of data quality, model validation and critical appraisal of AI claims, so that graduates can take part in, and not merely receive, these developments.⁴

10. CHALLENGES, RESEARCH GAPS AND FUTURE PERSPECTIVES

Several challenges recur across the topics above. Data quality is the first: pre-analytical conditions, assay method, calibration and reference intervals can become hidden predictors. Small and unrepresentative datasets produce unstable models and false discoveries. Bias from demographic imbalance, referral patterns and missing data requires performance to be assessed across relevant subgroups. Interpretability matters when outputs influence care, although an explanation of a prediction is not evidence of biological mechanism. Model drift after deployment requires pre-specified monitoring and recalibration.^{5,6}

A persistent research gap is the imbalance between technical performance and demonstrated clinical utility. Many reports present strong discrimination, but far fewer show that a model changes management or improves outcomes. Priorities include larger multicentre cohorts with standardised pre-analytical handling, leakage-controlled analysis, calibration assessment, subgroup analysis, prospective implementation studies that compare AI-assisted pathways with current practice, and health-economic evaluation. For EV and nucleic-acid markers, current standardisation guidance should be built into study design from the outset.^{13,14}

Likely future directions include multimodal integration of routine chemistry, immunoassays, molecular markers, omics and longitudinal clinical data, and development of compact panels that balance performance



with cost, availability, turnaround time and interpretability. Large language models may help with literature synthesis, education and structured reporting, but should be restricted to validated sources and not allowed to generate unverified clinical decisions.²⁵ AI literacy should become part of pharmacy and laboratory-science training, including overfitting, leakage, sensitivity and specificity, calibration, external validation and explainability, so that future professionals can scrutinise performance claims critically. In pharmacy practice, laboratory data linked to medication information could support monitoring of renal and hepatic function and detection of possible adverse effects, under appropriate clinical governance.

10.1 Limitations of this review

This review has limitations that readers should bear in mind. It is a narrative synthesis with a selective set of references, so relevant studies may have been missed. The references are predominantly reviews, reporting guidance and regulatory documents, and few primary studies are discussed; consequently, no study-level performance figures are tabulated and no comparison

of the accuracy of different algorithms is attempted. Several conclusions, particularly those on staged testing, study design and the Indian context, represent the author's reasoning from the cited literature and not findings reported in the sources. The field is also moving rapidly, and guidance documents and standards are updated frequently, so the current versions should always be consulted.

10.2 Synthesis across the biomarker classes

Taken together, the four marker classes show that the barriers to translation are different in kind, even though the computational methods are similar. PTX3 is a single, biologically interesting protein whose main problem is demonstrating incremental value and specificity against a background of non-specific inflammation. EV cargo and microRNA signatures are high-dimensional and therefore the most exposed to overfitting and batch effects, and their greatest need is measurement standardisation. Adipokines are measurable with conventional immunoassays but are tightly bound to adiposity, age, sex and renal function, so confounding is their central difficulty. Table 8 condenses this comparison.

Table 8: Dominant Translational Bottleneck for Each Marker Class

Marker class	Data dimensionality	Dominant bottleneck	Most useful AI role
PTX3 ¹¹	Low (single protein)	Incremental value and specificity	Multimarker models; subgroup identification
EV cargo ^{13,14}	High	Isolation and reporting standardisation	Panel prioritisation after quality control
MicroRNAs ¹⁸	High	Pre-analytical variation and normalisation	Signature building with leakage-free selection
Adipokines ^{19,20}	Low to moderate	Biological confounding	Phenotype discovery; adjusted prediction

Compiled by the author from the cited references.

10.3 Practical priorities for early-career researchers

Researchers beginning in this field, including final-year pharmacy students undertaking a mini project, can contribute meaningfully without large resources if they choose questions that match what they can control. First, a small, carefully designed study with standardised sample handling, a clear comparator and transparent analysis is more valuable than a large but uncontrolled dataset. Second, a simple regression baseline should always be reported before more complex methods are introduced. Third, the analysis plan,

including the validation strategy and the primary metric, should be written down before the data are examined. Fourth, code and preprocessing steps should be documented so that another person could reproduce them. Finally, conclusions should be proportionate: a single-centre study that cannot yet be externally validated should say so plainly and present itself as hypothesis-generating.

CONCLUSION

AI is expanding what clinical biochemistry can extract from complex data, at the levels of laboratory operations, multimarker prediction and biomarker discovery.



Emerging biomarkers such as PTX3, EV cargo, microRNAs and adipokines add biological information that may complement established tests, but each carries its own analytical and biological confounders that computation cannot correct. The evidence supports AI as an assistive technology that augments laboratory scientists, clinicians and pharmacists. Progress toward routine use will depend on standardised measurement, leakage-free modelling, calibration, independent external and prospective validation, transparent reporting, local validation, post-deployment monitoring and clear human accountability.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Pharmacy, Vikas Group of Institutions, Nunna, Vijayawada, for providing the academic environment and resources that supported the preparation of this review article.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Bunch DR, Durant TJS, Rudolf JW. Artificial intelligence applications in clinical chemistry. *Clin Lab Med.* 2023;43 (1):47-69. doi:10.1016/j.-ccl.2022.09.005
2. Hou H, Zhang R, Li J. Artificial intelligence in the clinical laboratory. *Clin Chim Acta.* 2024;559:119724. doi:10.1016/j.cca.2024.119724
3. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res.* 2023;394 (1):17-31. doi:10.1007/s00441-023-03816-z
4. Cadamuro J, Carobene A, Cabitza F, et al. A comprehensive survey of artificial intelligence adoption in European laboratory medicine: current utilization and prospects. *Clin Chem Lab Med.* 2025;63 (4):692-703. doi:10.1515/cclm-2024-1016
5. Miller HA, Valdes R. Rigorous validation of machine learning in laboratory medicine: guidance toward quality improvement. *Crit Rev Clin Lab Sci.* 2025;62 (5):327-346. doi:10.1080/10408363.2025.2488842
6. Spies NC, Farnsworth CW, Wheeler S, McCudden CR. Validating, implementing, and monitoring machine learning solutions in the clinical laboratory safely and effectively. *Clin Chem.* 2024;70 (11):1334-1343. doi:10.1093/clinchem/hvae126
7. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378
8. Lorde N, Mahapatra S, Kalaria T. Machine learning for patient-based real-time quality control (PBRTQC), analytical and preanalytical error detection in clinical laboratory. *Diagnostics (Basel).* 2024;14 (16):1808. doi:10.3390/diagnostics14161808
9. Chen W, Xiong Y, Zhang C, et al. Development and evaluation of artificial intelligence-based two-step model for automated serum quality assessment in clinical laboratories. *Ann Lab Med.* 2026;46 (5):613-623. doi:10.3343/alm.2025.0509
10. Esho A, Chong YP, Choy KW, Loh TP. Clinical decision support in laboratory medicine: a review of clinical, organisational and financial impacts with emphasis on patient outcomes. *Pathology.* 2025;57 (3):285-292. doi:10.1016/j.pathol.2024.11.005
11. Ristagno G, Fumagalli F, Bottazzi B, et al. Pentraxin 3 in cardiovascular disease. *Front Immunol.* 2019;10:823. doi:10.3389/fimmu.2019.00823
12. Lin J, Li J, Huang B, et al. Exosomes: novel biomarkers for clinical diagnosis. *Scientific-WorldJournal.* 2015;2015:657086. doi:10.1155/2015/657086
13. Welsh JA, Goberdhan DCI, O'Driscoll L, et al. Minimal information for studies of extracellular vesicles (MISEV2023): from basic to advanced approaches. *J Extracell Vesicles.* 2024;13 (2):e12404. doi:10.1002/jev2.12404



14. Pfaffl MW, Kubista M, Vandesompele J, Bustin SA. MISEV and MIQE: integrating domain-specific and general standards to strengthen extracellular vesicle biomarker research. *Extracell Vesicles Circ Nucleic Acids*. 2025;6 (4):669-676. doi:10.20517/evcna.2025.68
15. Pfaffl MW. How laboratory guidelines promote the validity of circulating extracellular vesicle-associated nucleic acid biomarker signatures in liquid biopsy. *Int J Mol Sci*. 2025;26 (24):12115. doi:10.3390/ijms262412115
16. Bettin BA, Varga Z, Nieuwland R, van der Pol E. Standardization of extracellular vesicle concentration measurements by flow cytometry: the past, present, and future. *J Thromb Haemost*. 2023;21 (8):2032-2044. doi:10.1016/j.jth.2023.04.042
17. Wang J, Chen J, Sen S. MicroRNA as biomarkers and diagnostics. *J Cell Physiol*. 2016;231 (1):25-30. doi:10.1002/jcp.25056
18. Naranbat D, Herdes E, Tapinos N, Tripathi A. Review of microRNA detection workflows from liquid biopsy for disease diagnostics. *Expert Rev Mol Med*. 2025;27:e11. doi:10.1017/erm.2025.2
19. Datta S, Koka S, Boini KM. Understanding the role of adipokines in cardiometabolic dysfunction: a review of current knowledge. *Biomolecules*. 2025;15 (5):612. doi:10.3390/biom15050612
20. Farkhondeh T, Llorens S, Pourbagher-Shahri AM, et al. An overview of the role of adipokines in cardiometabolic diseases. *Molecules*. 2020;25 (21):5218. doi:10.3390/molecules25215218
21. US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. 2021. Available from: <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>
22. US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Transparency for machine learning-enabled medical devices: guiding principles. 2024. Available from: <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles>
23. US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and FDA staff. Federal Register. 2024 Dec 4. Available from: <https://www.federalregister.gov/documents/2024/12/04/2024-28361/marketing-submission-recommendations-for-a-predetermined-change-control-plan-for-artificial>
24. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. Available from: <https://www.who.int/publications/i/item/9789240029200>
25. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2025. Available from: <https://www.who.int/publications/i/item/9789240084759>

HOW TO CITE: Dr. K. Narendra Kumar Reddy, K. Ravi, Pedavegi. Mounika, Guntaka. Charita Sree, Artificial Intelligence-Enabled Biomarker Discovery in Clinical Biochemistry: Emerging Biomarkers, Machine Learning and Barriers to Clinical Translation, *Int. J. of Pharm. Sci.*, 2026, Vol 4, Issue 10, 2332-2347. <https://doi.org/10.5281/zenodo.23391974>

