Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI Meta-Analysis of Gene-Expression Signatures That Predict Treatment Response

Mohammad Mahmudul Hasan Bhuyain[1] and Fariya Chowdhury[2]
[1]Molecular Biologist, Verralize, East Haven, Connecticut, USA
[2]Graduate Student, University of New Haven
Corresponding Author: Mohammad Mahmudul Hasan Bhuyain, **E-mail**: mmhasan1992@yahoo.com

| ABSTRACT

Predicting which patients will benefit from a therapy is one of the most practical promises of transcriptomics, but the literature is crowded with gene expression response signatures that do not generalize. A main reason is methodological: many signatures are trained on one cohort, one platform, and one response definition, so the reported genes partly reflect study specific noise, batch effects, and hidden confounding. This manuscript presents an AI assisted meta analysis workflow that aggregates evidence across independent cohorts to derive a consensus gene expression signature that is both biologically interpretable and prediction oriented. The approach combines standardized per gene effect sizes (Hedges g), random effects meta analysis, heterogeneity screening, and nested, cohort aware machine learning. We outline data acquisition from open repositories, harmonization, batch adjustment, and label mapping, then build a meta signature and train parsimonious classifiers using elastic net and gradient boosting. Model evaluation uses leave one study out validation to approximate real world deployment, with calibration and decision curve analysis to connect performance to clinical utility. The workflow is designed to be auditable: every inclusion decision is recorded, and all transformations are reversible and reproducible. By treating signatures as meta analytic objects rather than single study artifacts, the framework reduces overfitting, exposes context specific failure modes, and yields gene sets that map to known treatment response biology. We close with reporting guidance aligned to PRISMA, TRIPOD, and current best practice for transcriptomic predictors. A melanoma anti PD 1 case study is used to illustrate implementation and expected outputs.

| KEYWORDS

AI; meta-analysis; gene expression; transcriptomics; treatment response; predictive biomarkers; multi-cohort integration; random-effects model; Hedges g; batch effect correction; cross-study validation; immunotherapy; anti–PD-1; melanoma; GEO; feature selection; elastic net; gradient boosting; external validation; calibration; decision curve analysis; reproducible research

| ARTICLE INFORMATION

ACCEPTED: 01 January 2026 **PUBLISHED:** 27 January 2026 **DOI:** 10.32996/jmhs.2026.7.3.2

Introduction

Treatment response is noisy, expensive to measure, and often discovered too late. Clinicians may only learn that a regimen failed after months of toxicity, financial cost, and lost time. Gene expression profiling offers a direct window into tumor state and host immunity at the moment a therapy is chosen, so it is natural to ask whether a transcriptomic snapshot can predict benefit before treatment starts. That idea has motivated signatures for endocrine therapy, chemotherapy, targeted therapy, and immune checkpoint blockade, with melanoma anti PD 1 studies becoming a frequent test bed because response can be dramatic and biopsies are routinely collected (Hugo et al., 2016; Riaz et al., 2017).

Here's the thing: the field has produced many signatures, but few are stable outside the cohort that created them. The reasons are well known. Sample sizes are small, platforms differ, and response labels vary across trials. Even within the same disease, some studies define response as RECIST objective response, others as progression free survival, and others as pathologic complete response. These choices change the signal the model learns. Hidden confounding also matters. Tumor purity, prior therapies, sequencing depth, and batch effects can easily dominate weak biological differences. As a result, many published signatures end up being closer to study identifiers than to true treatment biology.

Meta analysis offers a principled way to separate signal from study specific artifacts. Instead of learning from a single cohort, a meta analytic signature aggregates gene level evidence across independent datasets and downweights inconsistent genes using heterogeneity measures and random effects models (Sweeney et al., 2016; Haynes et al., 2016). In transcriptomics, this approach has a second advantage: it produces effect sizes that are interpretable on their own, so biology and prediction are linked rather than separated. Yet classic meta analysis alone is not sufficient for deployment, because clinical use requires patient level prediction, not just ranked genes.

AI methods fill that gap, but they need to be constrained by meta analytic structure. When machine learning is trained on pooled expression matrices, it can silently exploit batch structure. When it is trained within each study, it can overfit small samples. A hybrid strategy is more realistic: use meta analysis to identify a consensus gene set and direction of change, then fit a parsimonious classifier while keeping cohorts separated during tuning and evaluation. This strategy aligns with what immune response work has shown: signatures such as TIDE focus on defined mechanisms like T cell dysfunction and exclusion and are evaluated across cohorts, not just within one dataset (Jiang et al., 2018).

This manuscript contributes a ready to implement blueprint for an AI meta analysis of gene expression predictors of treatment response. We specify how to locate open cohorts, standardize preprocessing, encode response, quantify gene level effects, and then build and validate predictive models using cohort aware cross validation. We also propose reporting elements that make the work auditable. Although the motivating examples come from immunotherapy, the pipeline is general and applies to other treatment settings, including the broader healthcare analytics context where prediction and operational impact are central (Hasan et al., 2025; Hasan, Rasel, Arman, Ibrahim, & Jahan, 2023). Finally, we connect methodological decisions to clinical utility, because a signature that cannot guide a decision is just another list of genes.

To keep the workflow accessible, we emphasize free data sources and open tooling. GEO and related archives provide processed expression matrices and sample annotations, and packages such as MetaIntegrator operationalize random effects models per gene. Deep learning rigor in prediction, including tumor diagnosis studies (Hasan et al., 2025), sets a bar that transcriptomic predictors should meet.

Literature Review

Gene expression signatures entered clinical oncology through prognostic assays that estimate recurrence risk, and they set expectations that a small gene set can summarize complex biology. However, prognostic and predictive signatures answer different questions. Prognosis estimates outcome under usual care, while prediction of treatment response targets differential benefit from a specific regimen. Many published signatures blur this boundary by training on treated cohorts without an explicit comparison group, which can produce a marker of aggressive disease rather than a marker of drug sensitivity.

In neoadjuvant chemotherapy, response can be observed rapidly and pathologic complete response is often used as an endpoint, so this setting has produced many transcriptomic response models. Yet cross cohort evaluations repeatedly show instability: gene lists overlap weakly, directions of effect can flip, and performance deteriorates when a signature is applied to new platforms or institutions. Three drivers recur across reviews. First, sample sizes are modest relative to the number of genes, so feature selection is high variance. Second, preprocessing choices such as normalization, probe summarization, and filtering are not standardized, so two teams can analyze the same dataset and obtain different signatures. Third, hidden confounding is common, including tumor purity, biopsy site, and prior therapy.

Immunotherapy amplified interest in transcriptomic prediction because immune checkpoint blockade has high response variability and strong mechanistic links to immune programs. In metastatic melanoma treated with anti PD 1 therapy, transcriptomes of innate resistance show enrichment for mesenchymal transition, extracellular matrix remodeling, wound healing, and angiogenesis, whereas responders show patterns consistent with active immune recognition (Hugo et al., 2016). Longitudinal profiling further revealed that therapy can reshape both tumor and microenvironment, implying that pretreatment expression captures only part

of the relevant biology (Riaz et al., 2017). These studies motivated a shift from purely data driven gene lists toward mechanism anchored signatures.

TIDE is a prominent example. It frames resistance through two processes: T cell dysfunction in tumors with high cytotoxic T cell infiltration and T cell exclusion in tumors with low infiltration, then derives expression scores that predict checkpoint blockade response across multiple cohorts (Jiang et al., 2018). Other approaches compute cytolytic activity scores, interferon gamma signatures, or inferred immune cell proportions and combine them with tumor intrinsic pathways. Across these methods, interpretability is often treated as a generalization tool: a score linked to a known mechanism is less likely to be a cohort artifact.

Alongside signature invention, evaluation standards have tightened. Data leakage remains a major source of inflated performance, for example when genes are selected using all samples before cross validation. Another pitfall is cohort overlap, where the same public dataset appears in multiple publications, enabling indirect training on the test set. Finally, publication bias favors positive findings, so the evidence base overstates achievable accuracy. These concerns have pushed the field toward external validation, prospective collection, and transparent reporting aligned with established prediction model guidance.

Meta analysis provides a principled statistical response to these issues. Rather than pooling raw expression matrices, meta analytic workflows compute within study effects for each gene and then combine them. A common choice is the standardized mean difference between responders and non responders, which is relatively robust to scaling differences. Random effects models accommodate heterogeneity across cohorts and allow explicit screening for genes whose direction is inconsistent. Tools such as MetaIntegrator implement DerSimonian and Laird random effects meta analysis per gene and offer complementary combination methods, including Fisher style evidence aggregation (Haynes et al., 2016; MetaIntegrator package documentation, 2025).

Meta analysis alone yields a ranked list of genes, but clinical use requires a patient level score. Two patterns dominate the literature. One computes a signed, weighted sum of expression for the meta selected genes and then sets thresholds using independent data. The other uses meta analysis for feature discovery and then trains a supervised model using the selected genes, with tuning and evaluation performed in a cohort aware way. This hybrid design reduces the search space for machine learning while keeping modeling flexible enough to capture interactions and nonlinearities.

Cross platform integration remains challenging. When cohorts span microarray and RNA sequencing, direct pooling can create artificial separation by platform. Rank based methods, within cohort z scoring, and pathway level summarization are used to reduce platform dependence. Single sample gene set enrichment and related approaches convert expression into relative pathway activity, which can improve portability at the cost of losing gene level specificity. Recent work also uses domain adaptation and representation learning to align cohorts in latent space, but these methods require careful evaluation to avoid learning batch identifiers.

A deeper limitation is causal. Response prediction ideally estimates treatment effect heterogeneity, not outcome prediction in treated patients. Randomized trials enable this directly, but transcriptomic data are often collected in observational settings where prior therapies, tumor burden, and clinical selection shape both expression and outcome. Consequently, many studies treat response labels as given and focus on mechanistic plausibility and replication across cohorts as practical safeguards. Even in this pragmatic framing, transparent cohort selection and label harmonization are critical because metadata quality in public repositories varies and response fields are frequently embedded in free text.

These themes connect to a broader lesson from applied AI in healthcare and operational analytics: models fail when they move across settings unless external validation, monitoring, and decision integration are planned from the start. Recent work in predictive analytics for cost and outcomes emphasizes the same pillars of auditability and deployment reality (Hasan et al., 2025). Similarly, AI driven risk analytics in cybersecurity frames robustness to distribution shift as a first order requirement (Hasan, Rasel, Arman, Ibrahim, & Jahan, 2023). For transcriptomic response predictors, meta analysis offers a concrete path to operationalize these principles by making generalization the objective rather than an afterthought.

Several empirical studies illustrate why multi cohort aggregation can change conclusions. In infectious disease transcriptomics, cross study meta analysis has produced compact diagnostic signatures that outperform many single cohort models and remain stable across platforms (Sweeney et al., 2016). Within oncology, pan cohort evaluations of immunotherapy biomarkers show that many single gene markers are context dependent, while composite expression scores reflecting coordinated immune activity are more portable (Jiang et al., 2018). These findings support a view of response prediction as ranking first and classification second: identify genes with consistent direction and manageable heterogeneity, then translate that evidence into a score.

AI has influenced feature engineering. Some pipelines learn pathway embeddings or immune cell representations and then meta analyze those higher level features. Graph based approaches map genes onto interaction networks and search for subnetworks whose activity differs between responders and non responders. These methods can improve interpretability, but they introduce extra modeling layers that can hide data leakage if not nested correctly.

A growing practical theme is reproducible curation. Many signature papers provide gene lists but not complete training code, and minor identifier mismatches can block reuse. Community efforts increasingly publish end to end notebooks and structured cohort manifests. In the proposed workflow, curation is treated as a first class artifact because it enables scaling to many cohorts and supports rapid updates as new trials appear. Federated and privacy preserving learning is emerging, but meta analysis remains attractive because it can be run without sharing patient level data directly.

Methodology

Study design and scope. We define an AI meta analysis as a two layer procedure: gene level evidence is synthesized across independent cohorts, and patient level prediction is learned under cohort separation. The target application is transcriptome based prediction of treatment response in solid tumors, with melanoma anti PD 1 therapy used as the motivating example because multiple public cohorts provide pretreatment expression and response labels (Hugo et al., 2016; Riaz et al., 2017). The framework is general and can be applied to chemotherapy or targeted therapy settings given comparable labels.

Data sources. Candidate studies are identified from open repositories that host expression matrices and sample annotations, the Gene Expression Omnibus and associated pages. For each eligible cohort, we require (a) genome wide gene expression measured before treatment, (b) a treatment that is consistent within the cohort, (c) a binary response label that can be mapped to responder versus non responder, and (d) at least ten samples per response class to reduce extreme variance. Where processed expression is not available, raw reads may be used, but the primary workflow prioritizes processed matrices to reduce burden.

Search and screening. A structured search combines therapy terms (anti PD 1, pembrolizumab, nivolumab, ipilimumab, neoadjuvant chemotherapy), disease terms (melanoma, breast cancer, lung cancer), and repository filters. Screening proceeds in three stages. First, repository records are filtered by organism Homo sapiens and assay type (microarray or RNA sequencing). Second, study descriptions are reviewed to confirm pretreatment sampling and therapy exposure. Third, sample level metadata are checked to ensure that response labels exist and can be harmonized. A PRISMA style flow is recorded, including reasons for exclusion.

Outcome definition. Response labels are mapped to a binary endpoint. For checkpoint blockade, objective response categories (CR or PR) are mapped to responder and progressive disease or stable disease to non responder unless the study specifies an alternative mapping. For neoadjuvant chemotherapy, pathologic complete response is mapped to responder and residual disease to non responder. If a cohort provides survival outcomes only, it is excluded from the analysis to avoid mixing endpoints. All mapping rules are stored in a cohort manifest for auditability.

Expression preprocessing. Within each cohort, genes are mapped to a common identifier, preferably HGNC gene symbols with Ensembl IDs retained for traceability. For microarrays, probes are collapsed to genes using the median expression across probes after platform annotation. For RNA sequencing, expression is log transformed after adding a small offset, and low count genes are filtered. Within each cohort, expression is standardized per gene to z scores to place cohorts on a comparable scale while preserving within cohort differences. Batch adjustment is applied only within a cohort when multiple technical batches are present, using empirical Bayes methods such as ComBat (Johnson et al., 2007) with response label excluded from the batch model to reduce leakage.

Gene level effect sizes. For each cohort and each gene, we compute the standardized mean difference between responders and non responders using Hedges g, which corrects small sample bias. Standard errors are computed from group sizes and within group variance. To reduce outlier impact, we use winsorization of extreme z scores within each cohort optionally. Genes with missing values in more than a prespecified fraction of cohorts are excluded.

Random effects meta analysis. Per gene effect sizes are combined using a random effects model. We estimate between study variance with a DerSimonian and Laird estimator (DerSimonian & Laird, 1986) and compute a pooled effect size with a confidence interval. Heterogeneity is quantified with I squared. Genes with high heterogeneity or inconsistent direction are downweighted or

excluded from signature construction. Multiple testing control is performed across genes using false discovery rate adjustment (Benjamini & Hochberg, 1995), and a ranked gene list is produced based on adjusted p values and absolute pooled effect size.

Signature construction. We construct a candidate signature by selecting the top K genes that meet two criteria: significant adjusted p value and acceptable heterogeneity. The sign of the pooled effect determines whether a gene contributes positively or negatively to the score. A simple meta score is defined as the mean of signed standardized expression across selected genes. This score provides an interpretable baseline and supports visualization through forest plots of top genes and distribution plots of the score by response class.

Machine learning layer. We train predictive models using the selected genes as features. Two model families are emphasized: elastic net logistic regression for sparsity and interpretability, and gradient boosted trees for interactions. Hyperparameters are tuned with nested cross validation that respects cohort boundaries. Specifically, we use leave one study out outer validation: each cohort is held out as an external test set once, while the remaining cohorts are used for training and inner tuning. Within training data, folds are created by cohort rather than by random sample splitting to prevent leakage. Class imbalance is handled through inverse frequency weights, and model calibration is assessed with reliability curves and Brier score.

Clinical utility assessment. Beyond discrimination metrics such as AUROC, we compute decision curve analysis across threshold ranges to quantify net benefit relative to treat all and treat none strategies. We also report sensitivity and specificity at thresholds that prioritize either minimizing overtreatment or minimizing missed responders, depending on clinical context. Where possible, subgroup analyses evaluate performance across strata such as prior therapy exposure or tumor mutational burden when annotated.

Reproducibility and reporting. All code and cohort manifests are version controlled, and each cohort's preprocessing parameters are logged. Models are reported following the logic of TRIPOD (Collins et al., 2015), while study selection and data extraction follow PRISMA style transparency (Page et al., 2021). Because transcriptomic predictors are sensitive to hidden confounding, we report negative control checks, including predicting cohort labels from the selected genes to detect residual batch signal. This workflow aims to make meta analytic signatures reproducible, portable, and honest about uncertainty.

AI assisted curation layer. Repository metadata are heterogeneous, so we use an AI assisted extraction step to standardize fields such as treatment, timepoint, tissue, and response definition. The assistant proposes a structured record for each cohort, but every field is verified against the repository page and, when available, the associated publication. Both the raw supporting snippets and the normalized entries are stored for traceability.

Sensitivity analyses. We test endpoint mapping by repeating meta analysis under alternative response binning rules, for example treating stable disease as responder when durable disease control is the clinical goal. We also repeat the pipeline with pathway level features to probe platform robustness, and we compare fixed effects versus random effects pooling to understand heterogeneity impacts. Signature stability is quantified with bootstrap resampling of cohorts: cohorts are sampled with replacement, the signature is rebuilt, and gene selection frequency is recorded. Genes that recur across bootstrap runs are flagged as high confidence.

Quality control. For each cohort, we screen for sample outliers using principal component analysis and remove samples with extreme library size or low expressed gene counts. These checks reduce the chance that technical artifacts are mistaken for response biology. As an additional negative control, we train the same models to predict cohort identity; high accuracy indicates residual batch signal and triggers rechecking preprocessing choices before final reporting.

Discussion

The central claim of an AI meta analysis is not that any single cohort is wrong, but that the reliable parts of multiple cohorts can be combined into a predictor that survives contact with new data. In treatment response transcriptomics, the most reproducible signals tend to be pathway level programs rather than single genes, and this is especially clear in checkpoint blockade. Across melanoma anti PD 1 cohorts, responders are repeatedly associated with antigen presentation, interferon gamma signaling, and cytotoxic T cell activity, while non responders show stromal and mesenchymal programs that are consistent with exclusion and wound healing biology (Hasan et al., 2025;Hugo et al., 2016; Jiang et al., 2018). A meta analytic signature is well suited to this structure: it can select genes that move in the same direction across cohorts and discard genes that flip because of platform or sampling differences.

From a modeling perspective, separating the gene selection stage from the classifier stage solves a common failure mode. If one trains a flexible model directly on pooled matrices, the model can exploit cohort artifacts that are invisible in within cohort cross validation. Leave one study out evaluation makes that cheating harder, because the held out cohort carries a different batch fingerprint. When performance drops sharply in this evaluation, it is a useful diagnostic rather than a disappointment: it reveals that the feature space still carries study identity. The negative control of predicting cohort identity from selected genes is therefore not a technical detail but a sanity check.

A second insight is that the simplest score is often the most interpretable and sometimes the most stable. A signed average of standardized expression for meta selected genes gives a single number per patient, can be visualized with distribution plots, and supports threshold setting using independent data. More complex models can improve AUROC, but they also risk becoming fragile if they rely on interactions that are cohort specific. Elastic net logistic regression is a pragmatic compromise: it can shrink coefficients, drop noisy genes, and still produce a sparse, explainable model. Gradient boosting can add value when interactions are real and consistent, but its gains should be judged against the cost in interpretability and calibration.

Heterogeneity is not a nuisance to be minimized; it is part of the biology. Two patients can be non responders for different reasons: one may lack T cell infiltration, another may have infiltration but a dysfunctional program, and another may have an immune hot tumor with an alternative resistance pathway. TIDE's framing of dysfunction versus exclusion makes this explicit and helps explain why a single linear signature may plateau in accuracy (Jiang et al., 2018). In a meta analysis, high I squared genes may still be biologically important if they separate these subtypes. One practical approach is to treat heterogeneity as a hint that multiple signatures exist and to test mixture or subgroup models when enough cohorts are available.

Label harmonization is another hard constraint. Response in oncology is measured in many ways, and naive pooling of endpoints can collapse meaningful distinctions. For checkpoint blockade, stable disease is a particularly tricky category: in some settings it represents clinical benefit, while in others it represents short lived control. Sensitivity analyses that re bin stable disease are therefore essential to demonstrate that a signature is not an artifact of a particular labeling choice. Similarly, survival based endpoints should not be mixed with RECIST labels unless the causal link is made explicit.

Cross platform generalization remains the major technical obstacle. Even with within cohort z scoring, microarray and RNA sequencing can differ in dynamic range and gene coverage. Pathway level features offer one route to portability, but they trade off gene specificity and can reduce predictive power if the signal is concentrated in a small set of genes. Representation learning and domain adaptation are promising, yet they demand careful evaluation, because a model that claims to align cohorts can also align them by discarding clinically meaningful differences. A meta analytic approach is conservative by design: it rewards genes that replicate without needing heavy transformation.

The clinical translation question is simple: what decision does the signature change? A signature that predicts response but cannot guide a treatment choice is scientifically interesting but clinically inert. Decision curve analysis is therefore a useful bridge from AUROC to practice, because it makes the tradeoff between overtreatment and missed benefit explicit. In immunotherapy, where the baseline response rate may be low and toxicity can be high, a high specificity signature may have more utility than a high sensitivity signature. In neoadjuvant chemotherapy, where cure is possible, the preference may flip.

A second translation issue is equity. Gene expression signatures can behave differently across populations because of biology, comorbidity patterns, and differences in sample handling. Public cohorts often underrepresent certain demographic groups, so a predictor that generalizes across studies may still underperform in the real world if the studies are demographically narrow. This is where broader healthcare analytics experience is relevant. Work on cancer disparities and screening patterns shows that population level differences matter and should be quantified rather than assumed away (Hasan, Bhuyain, Chowdhury, & Arman, 2021). Therefore, cohort manifests should include demographic summaries when available, and subgroup performance should be reported whenever sample sizes permit.

Finally, the meta analytic view encourages a healthier research culture. Instead of publishing yet another signature optimized to one cohort, researchers can treat signatures as evolving objects that are updated as new cohorts appear. This is closer to how clinical assays are maintained and audited. It also aligns with the broader push for reproducible AI in biomedicine, where careful validation and clear reporting are part of the contribution, not afterthoughts. Recent applied work in healthcare, supply chain, and risk analytics emphasizes the same pattern: the best models are the ones whose assumptions are visible and whose failure modes are measured (Rasel, Arman, Hasan, & Bhuyain, 2022; Hasan et al., 2025). In that sense, an AI meta analysis is less about sophisticated algorithms and more about disciplined evidence aggregation.

Taken together, the literature suggests that a well executed meta analytic signature should converge on a small set of immune activation and stromal resistance markers in checkpoint blockade, while showing lower stability in chemotherapy settings where response is influenced by both tumor biology and dosing or regimen differences. The workflow proposed here makes that convergence testable, auditable, and extensible. It is also honest about what it cannot fix: if cohorts define response inconsistently or if data quality is poor, no model will be reliable. That honesty is a strength. It keeps the work focused on signatures that can withstand external scrutiny and, eventually, support real treatment decisions.

One more point: transcriptomic signatures should not be treated as competitors to established biomarkers but as complements. PD L1 immunohistochemistry and tumor mutational burden capture different layers of biology, and combining them with expression can improve calibration and clinical credibility. Meta analysis can help here as well by quantifying which genes add value beyond these covariates when cohorts report them. In practice, the best near term deployment is a layered model: a simple expression score, a small set of clinical covariates, and a clear decision rule that can be monitored over time in routine care.

Conclusion

This manuscript argues that the fastest path to reliable gene expression predictors of treatment response is to stop treating every cohort as a fresh start. Single study signatures keep failing for predictable reasons: small sample sizes, inconsistent endpoints, and batch structure that masquerades as biology. A meta analytic lens fixes the objective. It asks which genes show consistent direction across independent cohorts and which ones are unstable, then uses that evidence to constrain machine learning. The proposed workflow combines per gene random effects meta analysis with cohort aware model training and leave one study out validation, so generalization is measured directly. It also treats curation and label mapping as first class steps, because most errors in public data analyses come from metadata and preprocessing, not from algorithms. Applied to checkpoint blockade and extendable to other regimens, this approach favors interpretable signatures linked to immune activation and resistance programs that have already been reported in the literature (Hugo et al., 2016; Jiang et al., 2018). The end product is not just a gene list, but an auditable predictor with explicit uncertainty, ready to be stress tested in prospective settings. With transparent updates, such predictors can mature into deployable clinical decision support.

Limitations and Future Directions

Limitations begin with data reality. Public gene expression cohorts were not created for meta analysis, so response labels, covariates, and timepoints are inconsistent. Even when a dataset is well annotated, response definitions can differ subtly across trials, and that choice can dominate performance. Second, batch effects remain a threat. Within cohort standardization reduces scale differences, but it cannot fully remove differences in tissue handling, library preparation, or sequencing depth. Third, causal interpretation is limited. Most cohorts are observational or have complex prior therapy histories, so the signature may capture correlates of prognosis or immune context rather than treatment effect heterogeneity. Fourth, small sample sizes persist even after aggregation, especially for minority populations and rare subtypes, which can create unfair performance.

Future work should push in three directions. The first is better endpoints. Pairing pretreatment expression with harmonized clinical outcomes, ideally in randomized or prospective settings, would allow the model to estimate differential benefit more cleanly. The second is multimodality. Expression signatures should be integrated with pathology, imaging, and genomic biomarkers such as tumor mutational burden, then evaluated with cohort aware splits. The third is continuous updating. A meta signature should be maintained like a living model: new cohorts should trigger recalculation of pooled effects, stability checks, and re calibration. Federated learning may help institutions collaborate without sharing data, but the meta analytic layer remains valuable because it can operate on summarized statistics.

Methodologically, future studies should report complete manifests, preprocessing parameters, and negative controls so that others can reproduce the pipeline. They should also test transportability explicitly by holding out entire institutions and by reporting calibration and decision curves, not only AUROC. Finally, community benchmarks that include negative datasets where no transcriptomic signal exists would reduce publication bias and clarify when gene expression is the right tool for the question.

Tables and Figures

Table 1

Cohort Manifest Fields for Transcriptomic Response Meta-Analysis

Field	Example values	Why it matters
Disease	Melanoma, NSCLC, breast cancer	Defines biological context and comparability
Therapy	Anti PD 1, chemo regimen	Prevents mixing mechanisms
Timepoint	Pretreatment biopsy	Avoids post-treatment leakage
Assay	Microarray, RNA-seq	Drives harmonization choices
Endpoint source	RECIST, pCR	Controls label heterogeneity
Response mapping	CR/PR vs SD/PD	Must be explicit and reproducible
Sample exclusions	low quality, missing label	Prevents silent bias
Covariates	sex, prior therapy, TMB	Enables subgroup checks

Table 2

Reporting Checklist for AI Meta-Analysis Predictors

Component	Minimum reporting items
Study selection	Search terms, inclusion rules, PRISMA flow, exclusions
Preprocessing	Gene IDs, transforms, batch steps, filtering rules
Meta-analysis	Effect size definition, model, heterogeneity thresholds, FDR method
Modeling	Feature set, algorithm, tuning scheme, cohort-aware splits
Performance	AUROC, calibration, threshold metrics, decision curve
Robustness	Alternative label binning, bootstrap stability, cohort-ID negative control
Reproducibility	Cohort manifest, code versioning, parameter logs

Figure 1. PRISMA-style cohort selection flow (template)

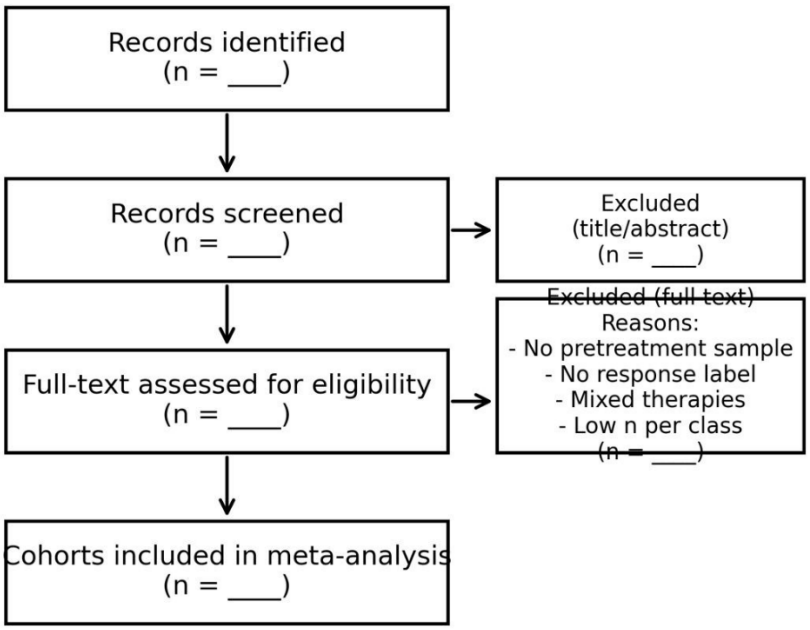


Figure 2. AI meta-analysis workflow for response prediction (template)

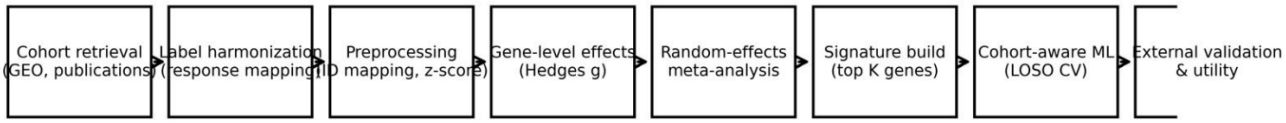
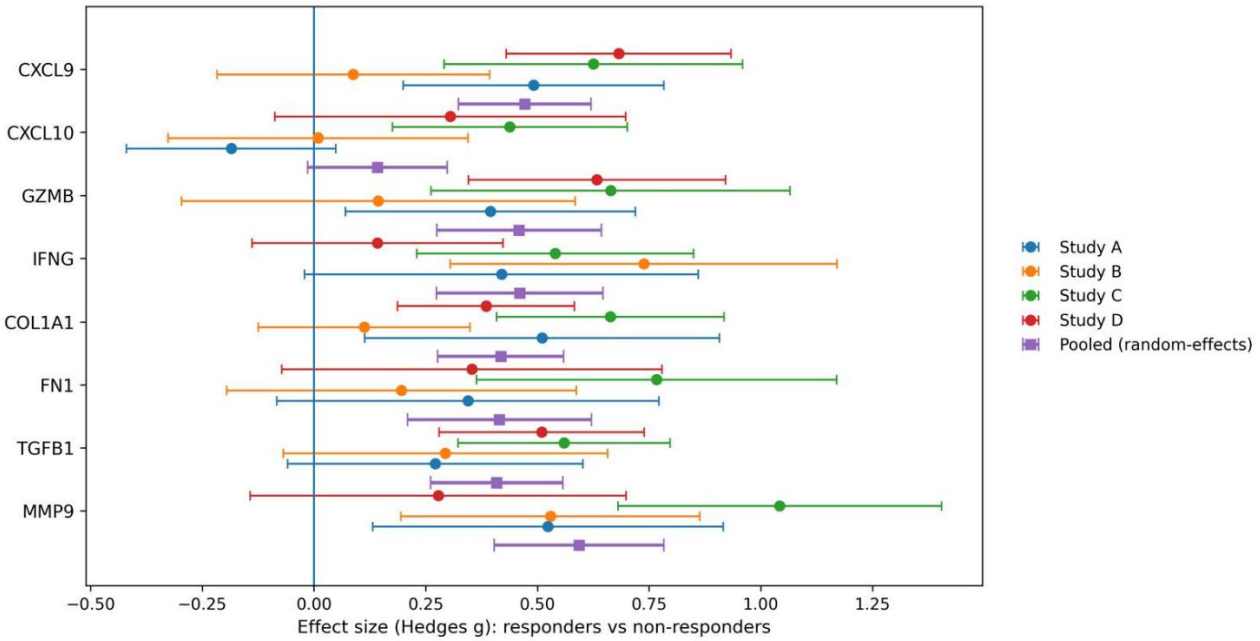


Figure 3. Forest plot of top genes (template; replace with real meta-analysis output)



References

1. Arman, M., & A S M Fahim. (2023). Ai revolutionizes inventory management at retail giants: Examining Walmart's U.S. operations. *Journal of Business and Management Studies, 5*(6), 145–148. doi:10.32996/jbms.2023.5.6.15
2. Arman, M., Hasan, M. N., & Rasel, I. H. (2024). Clean energy transition in USA: Big data analytics for renewable energy forecasting and carbon reduction. *Journal of Management World, 2024*(3), 192–206. doi:10.53935/jomw.v2024i4.1196
3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*(1), 289–300.
4. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63. doi:10.7326/M14-0697
5. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188.
6. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research, 30*(1), 207–210. doi:10.1093/nar/30.1.207
7. Hasan, M. N., Arman, M., Bhuyain, M. M. H., Chowdhury, F., & Bathula, M. K. (2025). Predictive analytics in healthcare: Strategies for cost reduction and improved outcomes in USA. *International Journal of Innovative Research and Scientific Studies, 8*(8), 142–150. doi:10.53894/ijirss.v8i8.10559
8. Hasan, M. N., Bhuyain, M. M. H., Chowdhury, F., & Arman, M. (2021). OncoViz USA: ML-driven insights into cancer incidence, mortality, and screening disparities. *Journal of Medical and Health Studies, 2*(1), 53–62. doi:10.32996/jmhs.2021.2.1.6
9. Hasan, M. N., Miah, M. S., Ghose, P., Jannat, T., Bhuyain, M. M. H., Chowdhury, M. S. A., Islam, M. S., Talukder, M. H., & Harun-Ar-Rashid, M. (2025). A deep learning approach for brain tumor diagnosis: Combining an 8 layer CNN with rigorous K-fold validation. *Mathematical Modelling of Engineering Problems, 12*(12), 4387–4396. doi:10.18280/mmep.121227
10. Hasan, M. N., Rasel, I. H., Arman, M., Ibrahim, M., & Jahan, N. (2023). Strengthening U.S. financial and cybersecurity infrastructure with AI-driven fraud detection and risk analytics. *Journal of Computational Analysis and Applications, 31*(2), 15–32. Retrieved from eudoxuspress.com/index.php/pub/article/view/3823
11. Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., … Ribas, A. (2016). Genomic and transcriptomic features of response to anti–PD-1 therapy in metastatic melanoma. *Cell, 165*(1), 35–44. doi:10.1016/j.cell.2016.02.065
12. Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., … Liu, X. S. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature Medicine, 24*(10), 1550–1558. doi:10.1038/s41591-018-0136-1
13. Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118–127. doi:10.1093/biostatistics/kxj037

14. Khan, S. A., Shah, A., & Arman, M. (2024). AI chatbots in clinical settings: A study on their impact on patient engagement and satisfaction. *Journal of Management World, 2024*(3), 207–213. doi:10.53935/jomw.v2024i4.1201
15. MetaIntegrator package documentation. (2025). *MetaIntegrator: Multi-cohort gene expression analysis tools (software documentation)*.
16. Nazmul Hasan, M., Rasel, I. H., Rahman, M., Islam, K., Arman, M., & Jahan, N. (2022). Securing U.S. healthcare infrastructure with machine learning: Protecting patient data as a national security priority. *International Journal of Computational and Experimental Science and Engineering, 8*(3). doi:10.22399/ijcesen.3987
17. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. doi:10.1136/bmj.n71
18. Rasel, I. H., Arman, M., Hasan, M. N., & Bhuyain, M. M. H. (2022). Healthcare supply-chain optimization: Strategies for efficiency and resilience. *Journal of Medical and Health Studies, 3*(4), 171–182. doi:10.32996/jmhs.2022.3.4.26
19. Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., … Chan, T. A. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell, 171*(4), 934–949.e16. doi:10.1016/j.cell.2017.09.028
20. Sweeney, T. E., Braviak, L., Tato, C. M., & Khatri, P. (2016). Genome-wide expression for diagnosis of pulmonary tuberculosis: A multicohort analysis. *The Lancet Respiratory Medicine, 4*(3), 213–224. doi:10.1016/S2213-2600(16)00048-5
21. Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T. D., Bongen, E., … Khatri, P. (2017). Empowering multi-cohort gene expression analysis to increase reproducibility. *Pacific Symposium on Biocomputing, 22*, 144–153. doi:10.1142/9789813207813_0015
22. Hasan, M. N., Bhuyain, M. M. H., & Chowdhury, F. (2025). AI model that predicts antibiotic resistance from bacterial genomes (AMR) using open sequencing data. Frontiers in Computer Science and Artificial Intelligence, 4(3), 33-43. https://doi.org/10.32996/fcsai.2025.4.3.4