QUALITY INTELLIGENCE INSTITUTE

DEFINING THE FUTURE OF HEALTHCARE AI QUALITY.

QUALITY INTELLIGENCE INSTITUTE

From Prediction to Intervention

Why Healthcare Quality AI Must Estimate Who Is Most Likely to Benefit, Not Just Who Is at Risk

Research Briefing #2 | PUB-RB-002

Executive Summary

Healthcare quality programs increasingly use artificial intelligence to identify who is most likely to experience an adverse outcome: who may be hospitalized, remain nonadherent, miss a screening, or fail to close a care gap. That capability is useful. But it answers a different question from the one operational leaders often need to answer: **who is most likely to improve because we intervene?** Outcome-risk prediction and intervention-benefit estimation are related, but they are not interchangeable analytic tasks. [1–4]

The distinction matters because high baseline risk does not, by itself, establish that a particular intervention will produce the greatest incremental benefit. Baseline risk can be an important determinant of absolute treatment benefit, and risk-based approaches can perform well. Yet treatment effects can vary across individuals, and some higher-risk or more severely ill groups may receive less benefit—or even net harm—from a specific intervention. [5–7]

For healthcare quality organizations, the practical implication is not to abandon risk prediction. It is to stop treating risk rank as if it were automatically a benefit rank. Where evidence and data permit, organizations should progressively evaluate modifiability, intervention response, comparative treatment benefit, clinical utility, and the opportunity cost of scarce outreach or care-management capacity. This is an evidence-supported QII recommendation, not a claim that benefit-targeted AI has already been proven superior to risk targeting in Medicare Advantage quality programs. [5,7,8,10,13]

Executive Takeaway

The executive decision is not whether to replace risk models. It is whether the evidence supports using a model to prioritize a specific intervention. QII's position is to match the analytic method to the decision: use risk prediction for prognosis, and require additional evidence before treating a score as proof of intervention benefit or as the basis for scarce-resource allocation. [1,5,7,10,13]

Key Findings

- **Risk and benefit answer different questions.** Risk prediction estimates what is likely to happen. Treatment-benefit estimation compares expected outcomes under alternative interventions and therefore requires causal assumptions and counterfactual reasoning. [1–4]
- **High risk does not universally mean high benefit.** A 2023 randomized-trial reanalysis of therapeutic-dose heparin found that less severely ill patients were more likely to benefit, while sicker patients were more likely to be harmed. This is intervention-specific evidence, not a universal rule about high-risk patients. [6]
- **Risk models remain useful.** In a 2025 scoping review, risk-model HTE claims met credibility criteria more often than effect-model claims among reports that claimed HTE (87% vs 32%). Risk prediction can remain a valuable component of targeting; the limitation is assuming it proves optimal incremental benefit. [5]
- **Benefit models require different validation.** Individual-level treatment effect is not directly observable as a conventional patient-level label. Benefit models require benefit-specific calibration/discrimination and decision-analytic evaluation, with especially careful external validation for flexible direct-effect models. [5,11–13]
- **Clinical AI effects can themselves be heterogeneous.** A 2024 Nature Medicine study across 140 radiologists and 15 chest X-ray tasks found substantial variation in whether AI assistance helped, with common clinician attributes failing to reliably identify who benefited. [14]
- **The Medicare Advantage evidence is promising but incomplete.** A matched observational CHF study found favorable ED-use and outpatient-spending signals from predictive targeting, but it did not compare risk ranking with a CATE/uplift allocation policy and cannot establish that benefit-based targeting is superior in MA quality programs. [8]

1. Prediction Is Not the Same as Intervention

A conventional healthcare prediction model asks a prognostic question: What is likely to happen to this person? For a quality program, that may mean predicting hospitalization, nonadherence, an open preventive-care gap, a missed appointment, or another adverse outcome. Conventional supervised machine-learning methods are primarily optimized to predict observed outcomes from available features. [1,2,9]

Intervention-benefit estimation asks a causal question: What would change if this person received intervention A rather than intervention B, or intervention rather than no intervention? That requires comparison of potential outcomes under alternative treatment conditions. Only one of those potential outcomes is ordinarily observed for a given person, which makes the individual treatment effect fundamentally counterfactual. Machine learning can help estimate such effects, but it does not remove the need for causal identification assumptions, appropriate study design, and careful validation. [1–4,10]

This distinction is operationally important because quality teams can move from a risk score to an intervention decision without explicitly testing whether prognosis and incremental benefit align. A model can identify who is most likely to experience an adverse outcome without establishing who will benefit most from a specific intervention. The risk model predicts prognosis; an allocation policy requires an additional estimate—formal or heuristic—of incremental impact. [5,8,10,13]

2. High Risk Is Not Always High Benefit

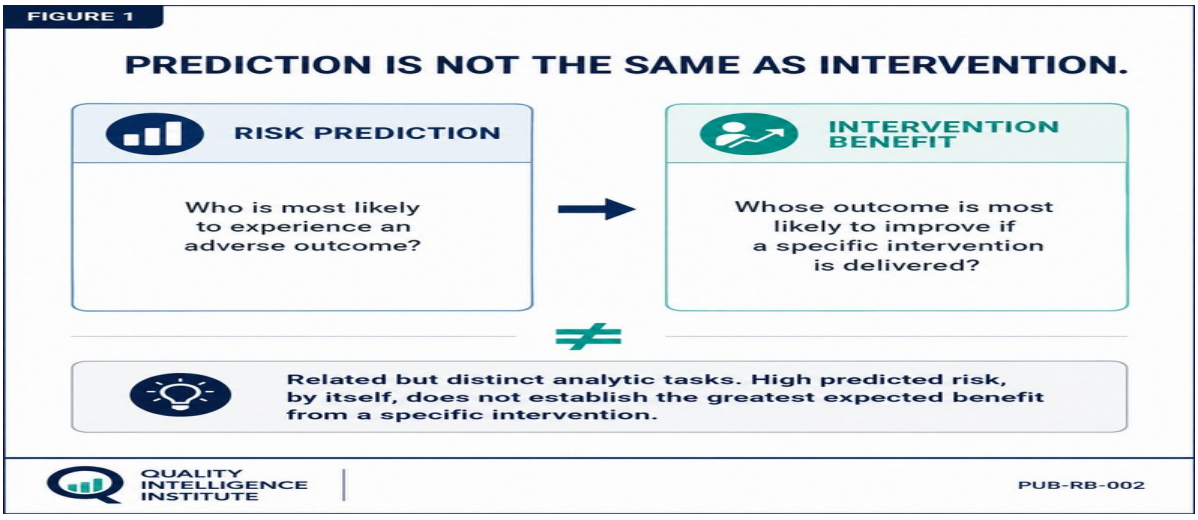
The evidence does not support a simplistic rejection of risk-based targeting. Baseline risk can be a strong determinant of absolute treatment benefit. Simulation work has shown that risk-based treatment-interaction approaches can perform well in many randomized-trial scenarios, particularly when higher baseline risk magnifies absolute benefit. [7]

The problem is the word always. Treatment-effect modifiers can change the relationship between baseline risk and benefit. A patient can be highly likely to experience an adverse outcome yet receive little benefit—or net harm—from a particular intervention because treatment effects vary across clinical states and individuals. [5–7]

A particularly clear example comes from a 2023 exploratory reanalysis of a multiplatform randomized trial of therapeutic-dose heparin in 3,320 patients hospitalized with COVID-19. Across risk- and effect-based analyses, the treatment was more likely to benefit less severely ill patients and more likely to harm sicker patients. [6] This result should not be generalized to all high-risk patients or interventions. It demonstrates a narrower and more important point: **risk severity and intervention responsiveness can diverge.**

The 2025 JAMA Network Open scoping review reinforces that average treatment effects do not always describe every clinically relevant subgroup. Among 65 reports analyzing 162 randomized clinical trials, 24 reports (37%) identified credible, clinically important heterogeneous treatment effects. In trials with an overall favorable treatment result, predictive modeling sometimes identified subgroups expected to receive no benefit or net harm. [5] The 37% figure applies to this selected HTE-reporting literature; it should not be interpreted as the prevalence of clinically important HTE across all trials or interventions.

Figure 1. Prediction Is Not the Same as Intervention.



3. Do Not Replace Risk Models—Use Them for the Right Decision

A benefit-aware quality strategy should build on—not discard—good risk stratification. The strongest contemporary synthesis creates an important guardrail against overstating the maturity of direct treatment-effect modeling. In the 2025 scoping review, among reports that claimed HTE, risk-modeling claims met prespecified credibility criteria in 20 of 23 reports (87%), compared with 10 of 31 reports (32%) for direct effect-modeling claims. [5]

That finding matters for healthcare leaders evaluating AI vendors. A more complex causal or uplift model should not receive a credibility premium merely because it uses newer terminology or more flexible algorithms. Flexible/direct treatment-effect models require especially careful validation, and external validation materially strengthened the credibility of exploratory effect models in the 2025 review. [5,12] In one diabetes treatment-effect study, a penalized regression model calibrated better than a causal forest in external data, illustrating why algorithmic flexibility should not be confused with demonstrated validity. [12]

The target state is therefore not “risk AI versus causal AI.” It is a QII synthesis about matching the analytic method to the decision: use prognosis where prognosis is the question, and require additional evidence when a model is used to make claims about intervention response, treatment selection, or resource allocation. [1,5,7,10,13]

4. From Risk Stratification to Benefit-Aware Allocation

For a healthcare quality program with constrained intervention capacity, QII treats the operational question as two linked questions: who has the greatest need, and where is the available intervention most likely to produce incremental benefit? Those questions can overlap, but they are not identical. [5,7,10,13]

Figure 2. QII Prediction-to-Intervention Ladder™ — QII evidence-supported synthesis.

QII Prediction-to-Intervention Ladder™	Decision Question
1. Descriptive	Who currently has a gap or adverse status?
2. Risk prediction	Who is most likely to experience the adverse outcome?
3. Actionability	Which risks or barriers are plausibly modifiable?
4. Intervention response	Who is most likely to respond to this specific intervention?
5. Comparative benefit	Which available intervention is expected to produce the greatest incremental benefit?
6. Optimized allocation	How should limited capacity be allocated to maximize outcomes, safety, equity, and value?

This ladder is a QII evidence-supported synthesis. It does not imply that every organization should immediately deploy a CATE or uplift model. Benefit estimation requires appropriate treatment variation or study design, defensible causal identification assumptions, adequate data for modeling and validation, and operational capacity to evaluate the intervention in practice. [1–3,5,10,18,19] The practical value of the ladder is governance: it forces the organization to identify which question its model actually answers before using the output to allocate scarce resources.

5. What Contemporary Applications Show—and What They Do Not

Contemporary studies demonstrate technically feasible ML-based treatment-selection and heterogeneous-response approaches in several specific clinical settings, including diabetes, atrial fibrillation, and sepsis. [12,15–17] A 2022 Lancet Digital Health study developed a five-feature treatment-selection algorithm to support choice between SGLT2 and DPP-4 inhibitor therapies and evaluated treatment-response differences in external trial datasets. [15] A 2023 study compared causal-forest and regression-based approaches using randomized diabetes trial data with external primary-care validation. [12]

In persistent atrial fibrillation, a 2024 Scientific Reports post-hoc analysis used uplift modeling to identify subgroups with different responses to more extensive ablation. In the selected model, the higher-uplift group had lower recurrence with the more extensive procedure (HR 0.40; 95% CI 0.19–0.84), while the lower-uplift group did not show a clear benefit (HR 1.17; 95% CI 0.57–2.39). [16] The analysis was exploratory and requires prospective validation; it demonstrates heterogeneous response modeling, not established routine-care superiority.

Similarly, a 2023 Nature Machine Intelligence study developed a causal machine-learning approach to estimate individualized effects of antibiotic timing in sepsis using real-world datasets and synthetic evaluation. [17] In observational data, however, treatment-benefit estimates are only as credible as the assumptions used to account for why different patients received different treatments. Those assumptions include consistency, positivity, and conditional exchangeability; residual confounding can still distort estimated benefit in ways that are difficult to predict. [10]

6. AI Assistance Is Also an Intervention

The prediction-to-intervention distinction also applies when the intervention is AI itself. A 2024 Nature Medicine study examined 140 radiologists across 15 chest X-ray diagnostic tasks and found substantial heterogeneity in the effect of AI assistance. Years of experience, subspecialty, familiarity with AI, and lower baseline performance did not reliably identify which radiologists would benefit; AI errors materially influenced whether assistance helped or harmed performance. [14]

The study cautions against assuming that AI assistance will provide uniform benefit across users and tasks. [14] QII therefore treats AI deployment as an intervention whose clinical utility, workflow effects, and user-level consequences should be evaluated in context rather than inferred from technical performance alone. [18,19]

7. Validation Must Match the Claim a Vendor Makes

A standard outcome-risk model can be evaluated against observed outcomes: a patient was predicted to be hospitalized, and later hospitalization either occurred or did not. Individual treatment benefit is different because the same person cannot ordinarily be observed both receiving and not receiving the intervention under otherwise identical conditions. [2–4,10]

Accordingly, benefit models require performance measures that address treatment benefit itself. Contemporary methods describe calibration for benefit (whether predicted benefit aligns with observed group-level benefit), discrimination for benefit (whether the model meaningfully distinguishes higher- from lower-benefit patients), and decision accuracy around a treatment threshold. [11] Decision-curve approaches can compare the net benefit of model-directed treatment with simpler policies across clinically meaningful thresholds. [13] Outcome-prediction metrics such as ROC-AUC may remain informative for prognostic performance, but they are insufficient by themselves to validate a claim that a model identifies who will benefit most from intervention. [11,13]

Randomized treatment assignment provides a particularly strong foundation for causal treatment-effect estimation because it reduces confounding of treatment assignment, but randomization does not automatically guarantee a valid individualized benefit model. Adequate sample size, appropriately controlled modeling, calibration, external validation, and transportability remain important. [3,5] When a model moves into live clinical use, clinical utility, safety, workflow, and human factors also require evaluation. [18,19] Observational benefit modeling requires additional scrutiny of causal identification and unmeasured confounding. [10]

8. The Medicare Advantage Lesson: Useful Evidence, Not Yet a Head-to-Head Answer

Medicare Advantage offers a practical example of why the distinction should be handled carefully. A 2022 American Journal of Managed Care study evaluated predictive analytics-driven disease management among Medicare Advantage members with congestive heart failure. After propensity-score matching, predictive targeting was associated with a 0.039 lower probability of an emergency department visit (P=.043) and $1,517 lower outpatient spending. The hospitalization difference was 0.032 with P=.075, and total medical spending was not significantly reduced. [8]

The study supports an important conclusion: risk-targeted care management can be useful. [8] It does not answer a different question: whether a policy targeting expected incremental benefit would have outperformed the risk-based policy. [8] The current QII source set does not contain a contemporary Medicare Advantage head-to-head study comparing risk-ranking allocation with a CATE/uplift benefit-maximizing policy. That gap must remain visible in any manuscript, executive claim, or derivative content.

9. What Executives Should Require Before Buying or Deploying

The verified evidence supports a structured set of questions when an AI vendor or internal analytics team claims that a model “targets the right patients.” [1–5,10–13,18,19] Healthcare quality leaders should ask what exactly has been demonstrated:

- **What exactly is the model estimating? In technical terms, what is the estimand: outcome risk, treatment response, comparative treatment benefit, or expected net utility?**
- **What evidence identifies intervention benefit?** Was treatment assignment randomized, observational, or inferred from historical workflow patterns?
- **How was confounding addressed?** For observational treatment-effect models, what assumptions support causal identification and what sensitivity analyses were performed?
- **How was benefit performance validated?** Were calibration for benefit, discrimination for benefit, external validation, or policy-value/net-benefit analyses performed?
- **What is the intervention being optimized?** A phone call, care-management enrollment, digital nudge, transportation support, medication change, clinical AI assistance, or something else?

- **Was the model compared with simpler allocation policies?** Did it outperform risk rank, clinician judgment, treat-all/treat-none, or other realistic operational strategies?
- **Was clinical utility tested live? Did evaluation include safety, workflow, human factors, adoption, and post-deployment monitoring—not only technical performance? [18–21]**
- **As a QII governance safeguard, leaders should also ask who may be harmed or systematically deprioritized. Benefit-aware allocation should be evaluated for equity, access, and unintended exclusion; expected utility is not the only legitimate objective. [19]**

10. The QII Position

QII's position is straightforward: the next stage of healthcare quality AI should not be defined by a race to deploy more complex causal models. It should be defined by a more disciplined match between the decision being made and the evidence supporting it. Risk prediction is appropriate when the task is prognosis. Intervention allocation requires additional reasoning about modifiability, treatment effect, implementation capacity, clinical utility, and competing objectives. [1–5,10,13,18,19]

For constrained healthcare quality interventions, QII recommends that organizations evaluate expected incremental benefit and clinical utility rather than use risk rank as the sole allocation criterion. This recommendation is supported by the broader causal-inference, heterogeneous-treatment-effect, decision-analysis, managed-care, and clinical-AI literature. It remains a QII evidence-supported synthesis—not a claim that benefit-based targeting has already been proven superior in every healthcare setting or in Medicare Advantage quality programs specifically. [5,7,8,10,13]

The governing question should evolve from "Who is most likely to have a bad outcome?" to "For whom can this specific intervention meaningfully change the outcome—and how certain are we?"

Evidence Limitations

- The source base spans methodological reviews, randomized-trial reanalyses, simulations, observational studies, applied ML studies, and reporting/governance guidelines; these designs do not provide identical levels or types of evidence.
- The 37% HTE figure from the 2025 scoping review applies to a selected literature of predictive HTE analyses and should not be generalized as the prevalence of clinically important treatment heterogeneity across all trials. [5]
- In the 2025 scoping review, direct effect-modeling claims met credibility criteria less often than risk-modeling claims among reports that claimed HTE, and external validation strengthened the credibility of exploratory effect models. Flexible ML also did not guarantee better calibration in the externally validated diabetes example. [5,12]
- The current evidence set does not establish that CATE/uplift targeting is superior to risk ranking in Medicare Advantage quality programs.
- Several applied studies are post-hoc, observational, or setting-specific. Feasibility and model validation should not be conflated with prospective improvement in routine-care quality outcomes. [6,8,12,16,17]
- As a QII governance synthesis, equity, access, feasibility, patient preference, and organizational accountability may justify allocation choices that differ from a purely benefit-maximizing policy; trustworthy-AI guidance likewise treats fairness and broader deployment considerations as lifecycle concerns. [19]

References

[1] Feuerriegel S, et al. Causal machine learning for predicting treatment outcomes. Nature Medicine. 2024. doi:10.1038/s41591-024-02902-1.

[2] Curth A, et al. Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities. Clinical Pharmacology & Therapeutics. 2024. doi:10.1002/cpt.3159.

[3] Hoogland J, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Statistics in Medicine. 2021. doi:10.1002/sim.9154.

[4] Ramspek CL, et al. Prediction or causality? A scoping review of their conflation within current observational research. European Journal of Epidemiology. 2021. doi:10.1007/s10654-021-00794-w.

[5] Selby K, et al. Predictive Modeling of Heterogeneous Treatment Effects in Randomized Clinical Trials: A Scoping Review. JAMA Network Open. 2025. doi:10.1001/jamanetworkopen.2025.22390.

[6] Goligher EC, et al. Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19. JAMA. 2023. doi:10.1001/jama.2023.3651.

[7] Rekkas A, et al. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk-based approaches. BMC Medical Research Methodology. 2023. doi:10.1186/s12874-023-01889-6.

[8] Ukert B, et al. On the Impact of Predictive Analytics-Driven Disease Management Interventions. American Journal of Managed Care. 2022. doi:10.37765/ajmc.2022.89275.

[9] Lipkovich I, et al. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. Clinical Trials. 2023. doi:10.1177/17407745231174544.

[10] Xia F, et al. Evaluating treatment benefit predictors using observational data: contending with identification and confounding bias. American Journal of Epidemiology. 2026. doi:10.1093/aje/kwaf239.

[11] Efthimiou O, et al. Measuring the performance of prediction models to personalize treatment choice. Statistics in Medicine. 2023. doi:10.1002/sim.9665.

[12] Venkatasubramaniam S, et al. Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity. BMC Medical Informatics and Decision Making. 2023. doi:10.1186/s12911-023-02207-2.

[13] Chalkou K, et al. Decision Curve Analysis for Personalized Treatment Choice between Multiple Options. Medical Decision Making. 2023. doi:10.1177/0272989X221143058.

[14] Yu F, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nature Medicine. 2024. doi:10.1038/s41591-024-02850-w.

[15] Dennis JM, et al. Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in type 2 diabetes. The Lancet Digital Health. 2022. doi:10.1016/S2589-7500(22)00174-1.

[16] Sato T, et al. Uplift modeling to identify patients who require extensive catheter ablation procedures in persistent atrial fibrillation. Scientific Reports. 2024. doi:10.1038/s41598-024-52976-7.

[17] Liu R, et al. Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis. Nature Machine Intelligence. 2023. doi:10.1038/s42256-023-00638-0.

[18] Vasey B, et al. DECIDE-AI: reporting guideline for early-stage clinical evaluation of AI decision support systems. BMJ. 2022. doi:10.1136/bmj-2022-070904.

[19] Lekadir K, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025. doi:10.1136/bmj-2024-081554.

[20] Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models using regression or machine learning. BMJ. 2024. doi:10.1136/bmj-2023-078378.

[21] Moons KGM, et al. PROBAST+AI: updated quality, risk-of-bias and applicability assessment tool for prediction models. BMJ. 2025. doi:10.1136/bmj-2024-082505.

QUALITY INTELLIGENCE INSTITUTE

DEFINING THE FUTURE OF HEALTHCARE AI QUALITY.

QII Research Briefing #2

From Prediction to Intervention

PUB-RB-002 | Final Release v1.2 — Brand Controlled

qualityintelligenceinstitute.org