

Evaluating Medical AI Model Claims When Model Weights and Data Are Not Available

M. Antony Ewing

mewing4@uic.edu

University of Illinois Chicago <https://orcid.org/0009-0005-1550-1718>

Research Article

Keywords: medical AI evaluation, information sufficiency, non-identifiability, predictive blindness risk, predictive indeterminacy measure, residual predictive advantage, model-agnostic diagnostics, substrate information limits, clinical decision support, data-processing inequality, deployment validation, proprietary model auditing

Posted Date: August 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10283069/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author is Managing Director of Conquer Medical Health and has financial and intellectual-property interests in the medical-AI frameworks and computational systems described in this manuscript, certain aspects of which are covered by provisional patent applications filed before his current institutional appointment. The study received no external funding. No other competing interests are declared.

Evaluating Medical AI Model Claims When Model Weights and Data Are Not Available

Maurice Antony Ewing, PhD

Professor of Research, Department of Information and Decision Sciences, University of Illinois Chicago, Chicago, IL, USA

Managing Director, Conquer Medical Health

mewing4@uic.edu; mae@conquermedical.health · ORCID: 0009-0005-1550-1718

Abstract

Medical AI can extract predictive signal from complex patient data at a scale and granularity unavailable to earlier clinical instruments, yet whether a reported claim holds in deployment is hard to establish when models are proprietary and their training and calibration are not inspectable. Existing evaluation frameworks—reporting standards, trial and diagnostic-accuracy guidelines, risk-of-bias and governance tools—improve how model studies are conducted and disclosed, but each presupposes access to the model and asks whether a study was done well, not whether the deployment data could support the claim at all. We introduce the AI Evaluation Protocol (AEP), which evaluates the substrate rather than the model. Any model can associate patient data x with a clinical outcome y only to the extent the data locally determine y from x ; where matched states branch into opposing futures, discrimination reflects the dominant branch rather than resolved information, and more data sharpens the estimate without removing the branching. AEP partitions a prospective substrate into identifiable, mixed, and non-identifiable regions using nonparametric, pre-model diagnostics and asks whether performance survives once information content is conditioned on. We apply AEP across four large-scale clinical substrates spanning distinct information geometries—routine longitudinal monitoring (ADNI), acute-care physiology (MIMIC-IV), exploratory tumor imaging (Lumiere/Yale), and population-health follow-up (HRS)—comprising approximately 40,500 unique patients, over 2,010,000 patient-level state transitions, and roughly 586,000 test observations, with a 12-model architecture-agnostic zoo run on ADNI and MIMIC-IV (96 model-claim evaluations) and substrate diagnostics computed on all four. Apparent discrimination ranged from AUC 0.66 to 0.90, yet residual advantage over the local majority floor was at or below zero in every substrate but one; that substrate also carried the highest false-appearance index. The finding is not universal failure but separation: some claims retain residual support, others derive apparent performance chiefly from substrate structure. Medical AI capability, and any trust placed in it, is bounded not only by model sophistication but, absent additional branch-separating information, by the information the deployment substrate contains.

Introduction

Medical AI claims are typically grounded in the reported performance of large, often proprietary models on clinical tasks. These models are benchmarked against the predictions of clinicians or conventional models and—when they exhibit superior performance—alleged to

add clinical value. But the external observer usually cannot inspect these models. Hidden from view is a myriad of the models' inner workings: weights, training substrates, prompts and deployment logic, preprocessing, calibration, and data-generating restrictions. The question then is not whether a model's claim is true in a scientific sense, but whether it translates into clinical reality. Even if a claim is fully valid within the study's setting, it may not answer the clinician's deployment question: given the data I have for this patient, in this situation, is this model's claim supportable?

We present a measurement protocol with which researchers and the clinicians they support can evaluate the potential validity of AI model claims in prospective deployment, even when the model's details are unavailable. Unlike standard benchmarking, which targets the particular model and therefore requires knowledge of its inner workings, the AI Evaluation Protocol (AEP) focuses on the informational limits of the clinical situation in which any model might be deployed. Those limits differ dramatically by situation: a routine checkup, a symptomatic examination, and an emergency visit expose starkly different information at the moment a decision must be made—even for the same patient with the same underlying pathology.

AI models—and all statistical models—share a common denominator: the kernel of prediction. Every model attempts to predict patient state y (say, diagnosis) from data x (biological readings, scans, notes). An architecture h , expressed as $h_{x;\theta}$, uses data x weighted by parameters θ to predict y ; training selects θ by minimizing a loss $\ell_{h,y}$. What makes AI models powerful is that h is not a pre-assumed parametric form: it takes its shape from the data and its computational approach (MoE, transformers, gradient boosting), and it can absorb multimodal inputs. AI models are therefore exceptional at extracting the information within a substrate. But this also means they, like all models, are bounded by the information that substrate contains. A classical model such as logistic regression takes x as a vector of structured features; an AI model can take x in multimodal form—text, image, waveform—and extract the maximum information the substrate affords. But in arriving at y , the model still performs a many-to-one compression: many configurations of x collapse to one response, trading information for a usable answer. A clinician performs a similar association—and also compresses—but conditions on a richer x , incorporating information present in the encounter that never entered the recorded substrate. This is why clinical intuition can outperform a sophisticated model: not because the clinician escapes compression, but because the model's ceiling is the recorded substrate while the clinician's is the encounter itself (Supplementary Appendix B).

Here is the theoretical crux. The primary limitation on h is its ability to establish an accurate, reliable link between x and y . If a configuration x_1 always maps to a single state y_1 in the scenario where data are collected, then y is identifiable from x , and a model powerful enough to discern x_1 can predict y_1 perfectly—the diagnosis is encoded in the data.¹ But if x_1 can be associated with multiple states— y_1 (positive) and y_2 (negative)—the mapping becomes conflicted: the model is effectively confused. We then say y is not identifiable from x . To resolve it, the model weights the observations by their sample frequency: if a fraction p of patients with

¹ Our notion of (non-)identifiability is related to but prior to the causal identifiability of Pearl and others: we ask whether the observed substrate locally determines the outcome at all, before any causal structure is posited.

x_1 show y_1 and $1-p$ show y_2 , the model maps x_1 to the average $y = py_1 + (1-p)y_2$, and if forced to a binary answer, sides with the majority branch. The clinician's patient, however, is a particular patient, not an average, and the difference between y_1 and y_2 is the whole point. Worse: if the pathology itself encodes a direction—most patients with these symptoms trend positive—the model will appear accurate despite its false positives and negatives.

A second form of non-identifiability arises from treatment. Even if h maps x_1 to y_1 statically, y may change with latent factors dynamically. Writing $y = f(x, u)$ with u an intervention level, an administered treatment u' produces $\Delta u = u' - u$ and a resulting change in y . The predictor knows the mapping to y , not to y' , so the effect of treatment is invisible to it. And, as we show below, biophysics intrudes: under treatment y does not merely shift to $f(x, u')$ but follows a new trajectory $f'(x, u') \neq f(x, u)$, so even a model that accounts for treatment may miss a curvature effect that reshapes the x - y relation.

AEP addresses all of this by partitioning a dataset, or a data-collection scenario, into identifiable and non-identifiable regions, enabling a three-way comparison: (1) a model's performance ignoring information limits; (2) its prospective performance on identifiable data alone; and (3) its performance on non-identifiable observations. A model whose performance varies little across the three is reliable—even if poor, it is reliably poor. A model that shines on mixed or identifiable data but is not reliable sides with the majority branch whenever confused, a decision the clinician cannot see.

Existing approaches to evaluating AI model claims

Existing medical-AI evaluation frameworks address important parts of this problem but do not resolve the translation from study to deployment. TRIPOD+AI extends prediction-model reporting for regression and machine-learning models; CONSORT-AI and SPIRIT-AI extend trial reporting and protocols for AI interventions; DECIDE-AI addresses early-stage live clinical evaluation of AI decision support; STARD-AI and CLAIM cover diagnostic-accuracy and imaging-AI reporting; PROBAST+AI addresses quality, risk of bias, and applicability; FUTURE-AI and Good Machine Learning Practice give broader principles for trustworthy development and deployment. These frameworks improve transparency, study quality, risk-of-bias assessment, and governance, but each presupposes access to the model or its validation outputs and asks how a model study should be conducted or reported. As a recent *Nature Medicine* editorial argues, a framework that evaluates a model's claims *in proportion* to what the evidence can support does not yet exist. AEP is such a framework: it asks whether the prospective deployment data contain enough information to support the claim at all. It complements rather than replaces reporting, validation, bias, and governance frameworks.

The disconnect between performance and reliability

The gap AEP targets is not hypothetical. At the frontier, performance and reliability have come apart: the Stanford HAI AI Index, an annual survey of the field, has for six consecutive editions reported gains in benchmark performance decoupled from reliability—the inverse of the error rate—with some periods showing reliability declining as performance rises. Benchmark performance and demonstrated capability, useful for scoping where medical AI might help, do not by themselves close the gap between demonstrated and prospective deployment.

performance. AEP makes that gap measurable at the level of the substrate, before a model is trusted.

AEP makes the limitation measurable with nonparametric, pre-model diagnostics. Predictive Blindness Risk (PBR) measures the local mass of opposing future branches within matched-state neighborhoods—the degree to which similar x admits opposite y . Predictive Indeterminacy Measure (PIM) measures the normalized separation between those branches in outcome units. Residual predictive advantage (RPA) measures the accuracy a model retains above the local majority floor in the non-identifiable region; where RPA is near zero, the apparent skill was momentum—directional performance obtained by following the dominant branch rather than resolving branch membership. These partition the substrate into identifiable, mixed, and non-identifiable regions and yield a Claim Support Score and a categorical verdict (supported, bounded, weak, failed, untested). When model outputs are available, AEP computes these directly; when they are not, it computes the substrate burden—the restricted-space evidence a claim would need to disclose. The logic echoes classical bounds: the data processing inequality caps information nonparametrically, and the Cramér-Rao bound caps estimator precision by the Fisher information of the data.

We demonstrate AEP across four clinical substrates spanning distinct informational geometries—routine longitudinal monitoring (ADNI), acute-care physiology (MIMIC-IV), exploratory tumor imaging (Lumiere/Yale), and population-health follow-up (HRS)—with a 12-model architecture-agnostic zoo. Apparent discrimination and information-constrained support separate cleanly: median AUC ranged from 0.66 to 0.90 across substrates, while residual predictive advantage over the local floor sat at or below zero in every substrate but one. The remainder proceeds as follows. Section 2 and Section 3 present the demonstration and its central finding; Supplementary Appendix A works a closed multimodal LLM claim end to end; Section 7 discusses limitations; and the Online Methods collect the model/substrate/claim formalism, the identifiability theorem, and the diagnostics.

Empirical Demonstration

Substrates

We applied AEP to eight medical prediction tasks from ADNI and MIMIC-IV. The ADNI task was CDR-SB cognitive-direction prediction using CDR subdomain measures and ADAS total score features. The seven MIMIC-IV tasks were direction prediction for lactate, mean arterial pressure (MAP), a derived acute-state trajectory, creatinine, urea nitrogen, bilirubin, and white blood cells, from longitudinal trajectories. One further task (ADNI MMSE) did not construct a valid AEP task under the present configuration and was excluded.

Model zoo

To avoid architecture-specific claims, we evaluated a 12-model architecture-agnostic zoo: logistic regression ($C=0.1, 1, 10$), random forest, shallow random forest, extra trees, decision tree, gradient boosting, histogram gradient boosting, AdaBoost, k -nearest neighbors ($k=15$), and Gaussian naive Bayes. Models were deliberately left at default hyperparameters with no per-substrate tuning, no class weighting, and no ensembling beyond the listed families. The

purpose is not state-of-the-art performance but claim-evaluation behavior across model families: if AEP verdicts were an artifact of a single architecture, they would not recur across logistic, tree-ensemble, instance-based, and generative-Bayes classifiers. The zoo spans those families precisely so that a verdict reflects the substrate rather than the estimator.

Evaluation design

Each task was split by patient or encounter group into training and test folds. Local substrate bins, PBR/PIM, and local majority floors were estimated in the training fold and applied to the test fold. Model performance was evaluated on the full test set and separately in identifiable, intermediate, and non-identifiable regions. The primary restricted-space test used the severe non-identifiability criterion.

Results

Four clinical substrates, four information geometries

Clinical tasks generate different informational geometries, and those geometries—not the algorithm—determine what performance is achievable. The same model family can appear powerful in one setting and collapse in another because the substrate supports different claims. We therefore evaluate AEP across four substrates chosen to span distinct clinical encounter types (Table 1). The 12-model zoo was run on ADNI and MIMIC-IV, where model outputs drive the RPA and verdict computations; Lumiere/Yale and HRS enter through the substrate diagnostics (branching, separability, information floor), which require no model outputs. ADNI represents routine longitudinal monitoring, where patients with similar observed states can branch into different futures. MIMIC-IV represents acute-care physiology, where clinicians actively change the generating process, so the model sees $f_{x,u0}$ while intervention moves the patient to $f'_{x,u1}$. Lumiere/Yale represents exploratory tumor imaging, where similar imaging phenotypes may carry different futures because latent biology is not encoded in the image. HRS represents population-health follow-up: noisy, heterogeneous, long-horizon trajectories in which population averages can mask distinct subgroup transition structures.

Apparent model performance versus information-constrained support across clinical encounter types. Median apparent AUC reports conventional discrimination; branching measures the extent to which patients matched on observed state separate into different futures; PIM measures pre-model separability where signed opposing branches exist (“—” where none are defined); residual predictive advantage measures performance beyond the local majority floor; the false-appearance index flags high apparent performance with limited residual support. High apparent performance coexists with zero or negative residual advantage in every substrate but one. For Lumiere/Yale and HRS, residual support is substrate-diagnostic rather than 12-model direct AEP RPA.

Encounter type	Substrate	Claims	Median AUC	Branching	PIM	RPA / res. support	AEP interpretation
Routine /	ADNI	3	0.655	0.380	—	−0.031	below

Encounter type	Substrate	Claims	Median AUC	Branching	PIM	RPA / res. support	AEP interpretation
longitudinal							information floor
Routine / longitudinal	ADNI	1	0.763	0.452	1.657	0.000	untestable in restricted space
Emergency / acute care	MIMIC-IV	7	0.684	0.803	1.787	0.070	high false-appearance (FAI 0.627)
Exploratory imaging	Lumiere/Yale	1	0.900	0.282	—	−0.012	below information floor
Population follow-up	HRS	2	0.750	0.340	—	−0.011	below information floor

Table 1 aggregates the per-task evaluations reported below into encounter-type rows. Across identical learning architectures, apparent performance varied not because models gained or lost intelligence, but because clinical substrates imposed different information limits on the claims tested. The models are probes; the clinical scenarios are the experiment. The highest-discrimination substrate (Lumiere/Yale, AUC 0.900) carries a *negative* residual advantage: the setting where models look most competent is, on these data, the one where the recorded image encodes least about the transition. Only MIMIC-IV exceeds its floor (RPA 0.070), and it does so while posting the highest false-appearance index in the study. The relevant finding is not universal failure. AEP supports claims when models retain residual advantage over the local floor; the central result is separation—showing which claims survive information conditioning and which derive apparent performance primarily from substrate structure.

Tasks and transitions

AEP constructed eight tasks totalling 2,010,168 patient-level transitions across 40,372 patient or encounter groups:

- ADNI CDR-SB: 2,385 transitions, 1,105 groups.
- MIMIC-IV lactate: 138,265 transitions, 23,702 groups.
- MIMIC-IV MAP: 329,135 transitions, 7,902 groups.
- MIMIC-IV acute-state trajectory: 1,529,472 transitions, 6,242 groups.
- MIMIC-IV creatinine: 2,631 transitions, 385 groups.

- MIMIC-IV urea nitrogen: 3,530 transitions, 452 groups.
- MIMIC-IV bilirubin: 811 transitions, 125 groups.
- MIMIC-IV white blood cells: 3,939 transitions, 459 groups.

Across eight tasks and 12 models, AEP produced 96 model-claim evaluations. The zoo yielded 34 supported, 11 bounded, 10 weak, 14 failed, 15 no-support, and 12 untested verdicts.

Task-level substrate structure

Seven of the eight tasks carried non-identifiable mass under the primary threshold. Non-identifiable test mass was 1.000 for creatinine, 0.840 for urea nitrogen, 0.938 for bilirubin, 0.718 for white blood cells, 0.641 for MAP, 0.513 for the acute-state trajectory, and 0.500 for lactate. Mean non-identifiable-region PBR ranged from 0.816 to 0.934 and mean PIM from 1.748 to 1.906. The ADNI CDR-SB task had weighted PBR 0.452 and weighted PIM 1.657, but no test observations fell into the severe non-identifiable region under the severe non-identifiability criterion; AEP therefore returned an untested restricted-space verdict for all 12 ADNI models.

Discrimination by identifiability regime

Aggregating the 96 evaluations by substrate regime makes the central pattern explicit (Table 2). In identifiable regions, models reach a median AUC of 0.957; on the mixed full substrate, 0.672; and in non-identifiable regions, 0.628—at or near the local majority floor. Residual predictive advantage tells the complementary story: it is essentially zero in identifiable regions (median -0.001 ; there is no majority-branch momentum to exceed once the region is one-sided), rises on the mixed substrate (median 0.064), and is small and dispersed in non-identifiable regions (median 0.027, IQR $-0.008, 0.080$), where roughly a quarter of evaluations fall below the floor.

Discrimination and residual advantage by identifiability regime, across all 96 model-claim evaluations. Identifiable-region statistics derive from the single task (acute-state trajectory) with non-empty identifiable test mass ($n=12$).

Regime	n eval.	Median AUC (IQR)	Median acc.	Median RPA (IQR)
Identifiable	12	0.957 (0.957–0.965)	0.932	-0.001 (-0.004 –0.000)
Mixed (full substrate)	96	0.672 (0.604–0.743)	0.625	0.064 (0.012–0.135)
Non-identifiable	84	0.628 (0.576–0.677)	0.601	0.027 (-0.008 –0.080)

The acute-state trajectory contributes 1,529,472 of the 2,010,168 transitions and is the only task with identifiable test mass, so it alone anchors the identifiable-regime row. Task-level summaries and verdict counts are unweighted by transition count—each model-task pair contributes one evaluation—so the trajectory does not dominate the verdict tallies. Excluding it entirely, the qualitative conclusion is unchanged: across the remaining seven tasks the full-substrate median AUC (0.653) still exceeds the non-identifiable-region median (0.616), and 26

supported restricted-space verdicts remain, confirming that global discrimination and restricted-space support diverge independently of the largest task.

Task-level AEP summary. PBR and PIM are train-defined substrate quantities; non-ID mass is the proportion of the test fold in cells meeting the severe non-identifiability criterion. RPAN is the maximum residual advantage over the local floor in the non-identifiable region.

Task	Models	Max AUC	PBR	PIM	Non-ID mass	Max RPA N	Best CSS
ADNI	12	0.809	0.452	1.657	0.000	–	0.0
CDR-SB							
MIMIC lactate	12	0.684	0.696	1.787	0.500	0.015	22.1
MIMIC MAP	12	0.744	0.729	1.752	0.641	0.063	97.6
MIMIC acute-state traj.	12	0.869	0.555	1.631	0.513	0.119	100.0
MIMIC creatinine	12	0.679	0.888	1.888	1.000	0.099	71.5
MIMIC urea nitrogen	12	0.707	0.903	1.829	0.840	0.161	82.7
MIMIC bilirubin	12	0.682	0.824	1.735	0.938	0.035	44.8
MIMIC WBC	12	0.605	0.803	1.817	0.718	0.033	28.1

Supported restricted-space claims

Four tasks produced supported restricted-space claims for multiple models. MIMIC-IV urea nitrogen supported all 12 models; the best (shallow random forest) had full AUC 0.707, non-identifiable-region accuracy 0.657, local floor 0.496, RPAN=0.161, CSS 82.7. The acute-state trajectory supported 8 of 12; several models reached CSS 100.0, with maximum RPAN=0.119. MAP supported 6 of 12 (best CSS 97.6, RPAN=0.063). Creatinine supported 8 of 12; the best (gradient boosting) had full AUC 0.679, non-identifiable-region accuracy 0.625, local floor 0.527, RPAN=0.098, CSS 71.5. AEP supports claims when models exceed the local substrate floor in high-PBR, high-PIM regions.

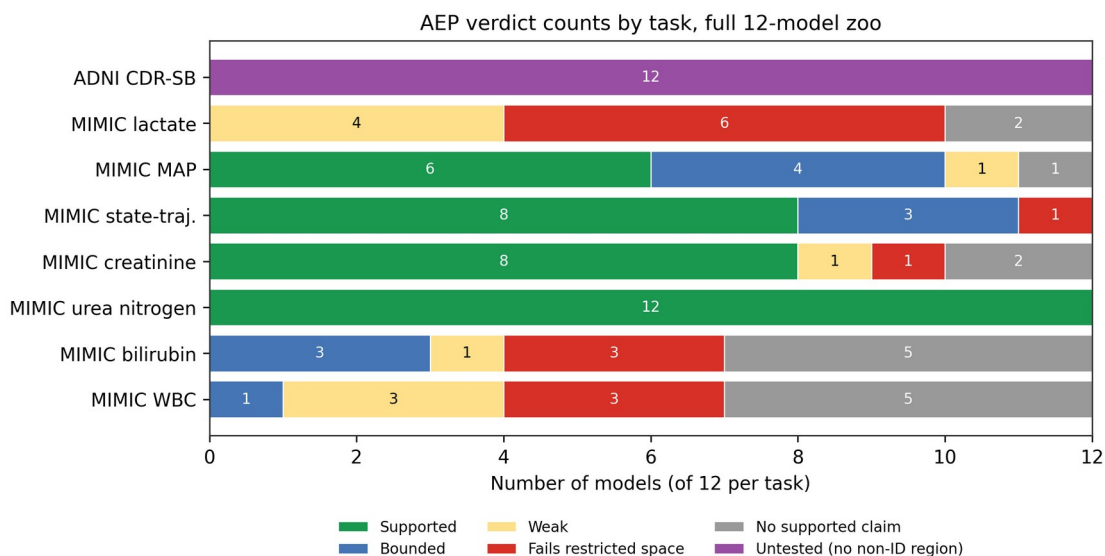
Bounded, weak, and failed claims

Lactate, bilirubin, and white blood cells showed the central AEP phenomenon: global signal did not guarantee restricted-space support. For bilirubin, the best model by CSS (shallow random forest) had full AUC 0.682 and RPAN=0.031 (bounded, CSS 44.8), while other models with global AUC above 0.60 had negative RPAN (failed). For white blood cells, AdaBoost achieved bounded

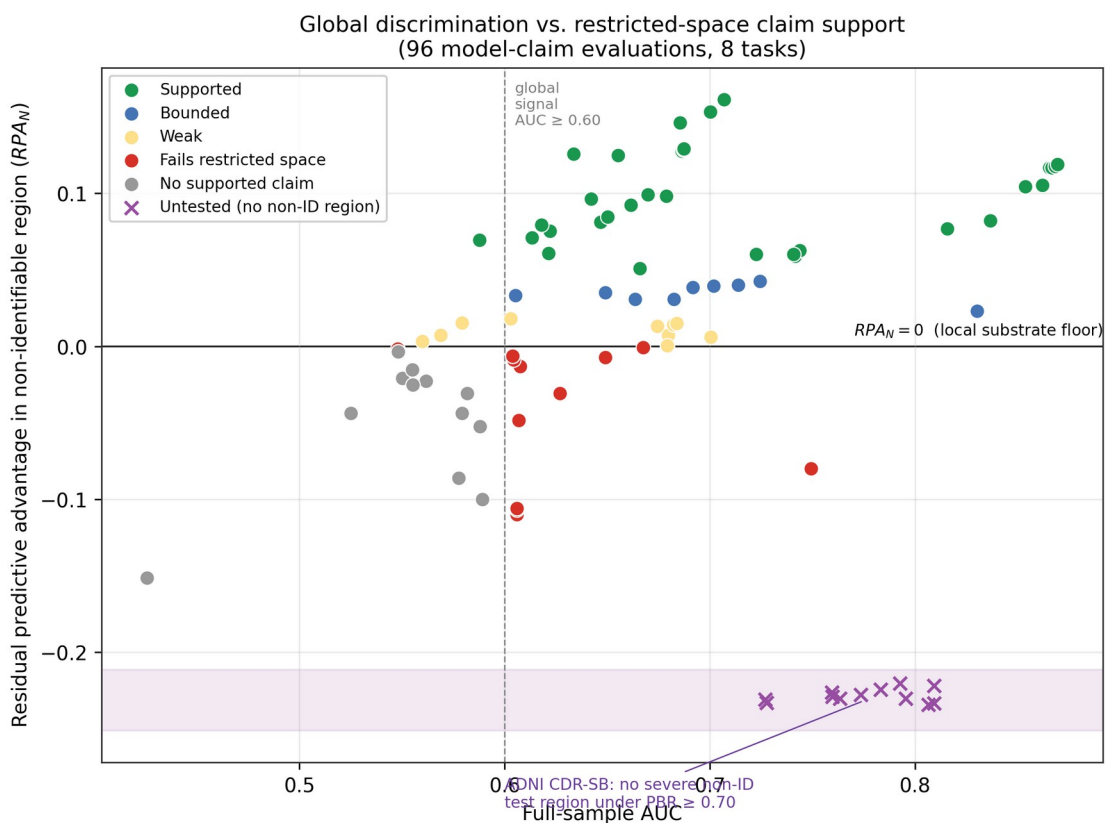
support (AUC 0.605, RPAN=0.033, CSS 28.1) while logistic models with AUC $\approx$ 0.604 had RPAN=-0.006 (failed). Lactate produced no supported verdicts: its best model (histogram gradient boosting) reached only RPAN=0.015 (weak, CSS 22.1), and six of twelve models failed the restricted-space test outright.

Best claim-support result per task. Support is determined by residual advantage over the local floor in the non-identifiable region.

Task	Best model	AUC	Acc.	Non-ID floor	Non-ID acc.	RPAN	Verdict
MIMIC acute-state traj.	Random forest	0.837	0.763	0.596	0.678	0.119	Supported
MIMIC MAP	GBM	0.744	0.678	0.570	0.633	0.063	Supported
MIMIC urea nitrogen	Shallow RF	0.707	0.644	0.496	0.657	0.161	Supported
MIMIC creatinine	GBM	0.679	0.625	0.527	0.625	0.098	Supported
MIMIC bilirubin	Shallow RF	0.682	0.650	0.627	0.658	0.031	Bounded
MIMIC WBC	AdaBoost	0.605	0.596	0.548	0.582	0.033	Bounded
MIMIC lactate	Histogram GBM	0.684	0.657	0.587	0.602	0.015	Weak
ADNI CDR-SB	Logistic C=0.1	0.763	0.705	–	–	–	Untested



AEP verdict counts by task, full 12-model zoo (96 evaluations). Urea nitrogen, creatinine, MAP, and the acute-state trajectory carry predominantly supported claims; lactate, bilirubin, and WBC contain weak, bounded, and failed cases; ADNI CDR-SB is untested (no severe non-identifiable region).



Global full-sample AUC versus restricted-space RPA_N, colored by verdict, across all 96 model-claim evaluations. Models with acceptable global AUC fall at or below RPA_N=0, showing failed

restricted-space support despite adequate global discrimination; the 12 ADNI evaluations appear as an untested strip with no severe non-identifiable test region.

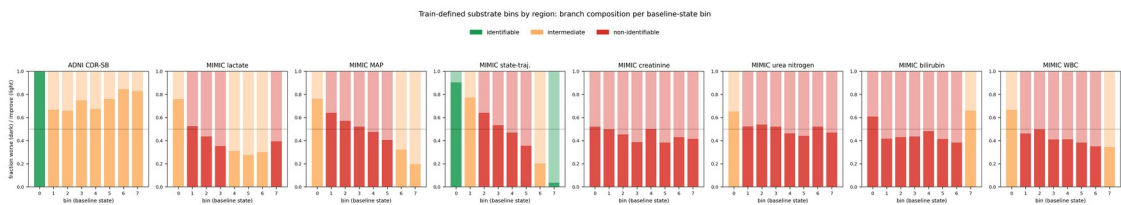
Stability across repeated splits

To assess whether the predictive signal underlying the MIMIC-IV analyses was sensitive to a single train/test partition, each MIMIC-IV task was re-evaluated across 30 independent patient-group splits (test fraction 0.4) using a majority-class control, logistic regression, and a nonlinear random-forest reference model. Global discrimination was stable throughout (Table 3): standard deviations of full-sample AUC ranged from 0.0004 on the acute-state trajectory (n≈382,000 per split) to 0.034 on bilirubin (n≈203), the smallest substrate. A majority-class control held at AUC 0.500 in every split of every task, confirming that the local floor is a property of the substrate rather than an estimation artifact. The largest tasks vary least; the dispersion on bilirubin reflects its sample size, consistent with the narrow information window the substrate metrics already report for it. This analysis supports the robustness of the task-level predictive signal used in the demonstration; it is not a full repeated-zoo re-estimation of every AEP verdict.

Best-model full-sample AUC across 30 repeated patient-group splits, MIMIC-IV tasks. A majority-class control held at 0.500 in all splits (not shown).

Task	Mean AUC	SD
Acute-state trajectory	0.868	0.0004
MAP	0.743	0.0015
Lactate	0.685	0.0026
Urea nitrogen	0.677	0.0168
Creatinine	0.684	0.0161
WBC	0.599	0.0180
Bilirubin	0.604	0.0345

Performance–reliability decoupling



Reliability is separable from apparent performance. Defining a derived reliability score as one minus the range of a model’s AUC across populated information regimes (full, identifiable, non-identifiable, intermediate), computed only where at least three regimes are populated, exposes the split directly (Table 4). The acute-state trajectory model discriminates well (AUC 0.837) yet is the least reliable in the set (score 0.772): its AUC falls from 0.957 in the identifiable region to 0.729 in the non-identifiable one. Several models are stable but flagged for high false-appearance risk; one substrate (creatinine, bilirubin) populates too few regimes to score, and is marked as such rather than assigned a spurious perfect reliability. This diagnostic is not a

ninth core AEP output but a summary showing that apparent performance and supportable reliability are separable.

Performance–reliability decoupling across AEP information regimes. Reliability is defined only when at least three regimes are populated. A model can have high apparent discrimination while collapsing in non-identifiable regions, yielding low reliability despite high AUC.

Task	Model	AUCfull	Regimes	Range	Reliability	RRC†	Reliability verdict
Acute-state traj.	rf	0.837	4	0.228	0.772	0.221	high-performance unstable
MAP	gbm	0.744	3	0.143	0.857	0.130	high false-appearance
Urea nitrogen	rf	0.707	3	0.030	0.970	0.029	bounded /stable
Creatinine	gbm	0.679	2	—	—	0.008	insufficient coverage
Bilirubin	rf	0.682	2	—	—	0.070	insufficient coverage
WBC	adaboost	0.605	3	0.056	0.944	0.043	high false-appearance
Lactate	histgbm	0.684	3	0.100	0.900	0.077	high false-appearance
ADNI CDR-SB	logit	0.763	3	0.066	0.934	0.000	no residual advantage

†RRC (regime residual contrast) is the regime-range reliability diagnostic, not the primary restricted-space RPAN used for AEP verdicts (Table 2).

Support-adjusted branching and scale-invariance

Branching estimates depend on sample size, and the “more data fixes it” objection deserves a direct answer. Each matched-state cell’s branching estimate carries an exact 95% interval; cells

whose interval is narrow are treated as resolved and wider ones as undersampled (Table 5). Separately, branching is recomputed over increasing subsamples: estimator variance shrinks toward zero while the converged branching estimate remains bounded away from it (Table 6). Sample size stabilizes the estimate; it does not remove the branching.

The interval construction, the resolution criterion, and the estimator internals are part of a protected framework and are not reproduced here (see Implementation Boundary). The substantive point is the one the subsampling establishes: as sample size grows, estimator variance shrinks toward zero while the converged branching estimate remains bounded away from it. Where matched-state cells are refined as the sample grows, additional data may simply populate a finer set of ambiguous cells rather than resolve them.

Support-adjusted branching. Branching estimates carry exact 95% intervals; cells with narrow intervals are treated as resolved.

Substrate	Task	Bins	n	Weighted PBR	Max CI width	Status
ADNI	ADAS	8	1,470	0.721	0.157	undersampled
ADNI	CDR-SB	6	892	0.762	0.281	undersampled
ADNI	MMSE	5	1,230	0.603	0.148	resolved
Lumiere/Yale	modality bins	32	5,957	0.514	0.178	resolved

Scale-invariance of PBR. Estimator SD falls to zero at full n while the converged branching estimate stays bounded away from zero.

Substrate	Task	N	Converged PBR	SD at n/8	SD at full n
ADNI	ADAS	1,470	0.721	0.056	0.000
ADNI	CDR-SB	892	0.762	0.073	0.000
ADNI	MMSE	1,230	0.603	0.052	0.000
Lumiere/Yale	modality bins	5,957	0.514	0.030	0.000

Discussion

What AEP does and does not establish.

AEP determines whether a prospective substrate contains enough information to support a claim in its non-identifiable regions, and whether an available model's performance survives conditioning on that information. It does not certify a diagnosis, validate a model prospectively, or prove that a claim is false; a verdict of "untested" or "below floor" means the recorded substrate does not support the claim as posed, not that no substrate could. Nor does AEP

measure clinical utility, calibration, or causal effect. It measures one axis—substrate information sufficiency—and is silent on the others by construction.

AEP provides a protocol for evaluating medical AI claims when model internals are inaccessible. It does not require access to weights to ask whether the published evidence is adequate for the claim. If the substrate is locally non-identifiable, then a broad predictive claim requires restricted-space evidence. If that evidence is absent, weak, or negative, the claim must be bounded or rejected. The demonstration shows three important properties. First, AEP supports claims when the substrate carries them. Urea nitrogen and creatinine claims were supported in severe non-identifiable regions. These tasks had high PBR and PIM, but several models still retained positive residual advantage over the local substrate floor. AEP therefore permits strong claims when the evidence supports them. Second, AEP identifies overbroad claims. In bilirubin and WBC tasks, some models had global AUC near or above 0.60 but failed to beat the local substrate floor in non-identifiable regions. A global result that looks acceptable by ordinary reporting becomes insufficient under AEP because the claim does not survive the restricted-space test. Third, AEP can return untested rather than forcing an answer. The ADNI CDR-SB task had global AUC up to 0.809, but the severe non-identifiable test region was absent under the prespecified threshold. AEP did not treat the global result as restricted-space support. It returned an untested restricted-space verdict. This is a strength, not a weakness: the protocol distinguishes unsupported, supported, and not testable under the chosen evidence rule. The practical implication is direct. A closed medical AI model may claim broad predictive competence, but if the substrate contains high local non-identifiability, the claim requires evidence of residual advantage in that region. If the paper reports only global AUC, calibration, or benchmark accuracy, AEP can say:

The model may or may not work under undisclosed conditions, but the published evidence does not support the claim as stated.

This is the appropriate standard for inaccessible models. The burden is not on external evaluators to reconstruct hidden weights. The burden is on claim authors to show that their model exceeds the substrate floor where the substrate makes the claim hardest.

Relationship to Conventional Evaluation

AEP does not replace discrimination, calibration, external validation, or clinical utility evaluation. Instead, it supplies a prior evidentiary layer. Conventional metrics ask whether a model performs well on average. AEP asks whether the claim is supported in the substrate regions where local information structure allows or blocks the claim. AUC is therefore treated as a global signal, not as claim support. Accuracy is compared not only with the global majority baseline but also with the local substrate floor. A model can have acceptable global AUC and still fail AEP if it has no residual advantage in the non-identifiable region. This distinction is especially important for models trained or evaluated on broad clinical substrates. The model may perform well in easy regions where state already determines direction. But the scientific claim often concerns hard regions: ambiguous disease trajectories, early deterioration, treatment response, recovery, reversal, or diagnosis under uncertainty. AEP forces evaluation to occur where the claim is actually informative.

Use with Inaccessible Models

When a model is fully accessible, AEP can be run directly on its predictions. When a model is inaccessible, AEP can still evaluate the claim burden. The evaluator can compute the substrate's PBR, PIM, temporal structure, and local floors. If the published paper does not report restricted-space residual advantage, then the claim is not externally supportable as stated. AEP therefore has two modes:

- **Direct AEP:** model outputs are available, so restricted-space RPA and CSS are computed directly.
- **Substrate-burden AEP:** model outputs are unavailable, so AEP estimates what the claim would need to demonstrate on the substrate and determines whether the published evidence supplies it.

In the inaccessible-model setting, AEP should not say that the model cannot work. It should say that the claim is not supported by the evidence disclosed. This distinction is central.

Limitations

This demonstration uses an architecture-agnostic model zoo, not proprietary closed models. That is intentional: the purpose is to show the protocol's behavior when claims are tested against substrate structure. The same framework can be applied to closed models when outputs are available, or used to evaluate claim burden when outputs are unavailable. The non-identifiability threshold the severe non-identifiability criterion is prespecified and conservative. Sensitivity analyses should report alternative thresholds such as $PBR \geq 0.60$. The current run included eight constructed tasks; one attempted ADNI MMSE task did not construct a valid AEP task under the present configuration and was excluded. The local floor is a directional benchmark. It does not capture calibration, clinical utility, treatment value, or causal effect. The protocol also depends on correct target orientation, support thresholds, and patient-level splitting. Finally, AEP does not identify the missing latent variable ϕ ; it identifies when the observed substrate x does not support the claim without additional branch-separating information.

Conclusion

AEP evaluates medical AI claims by testing them against the information structure of the substrate. If the substrate is locally non-identifiable, a model must beat the local substrate floor in that region. If it does not, the claim is unsupported there, regardless of global AUC. If the non-identifiable region is unavailable under prespecified thresholds, the claim is untested rather than assumed supported. In 96 model-claim evaluations across ADNI and MIMIC-IV tasks, AEP supported some claims, bounded others, rejected several restricted-space claims, and refused to over-interpret a globally strong ADNI result without a severe non-identifiable test region. This establishes the central role of AEP: not to rank models in aggregate, but to determine whether a medical AI claim is supported, bounded, failed, or untestable when model internals are inaccessible.

Methods

The formal apparatus of AEP—the model/substrate/claim distinction, the identifiability theorem, and the substrate diagnostics—is collected here. Readers interested only in the empirical demonstration may skip to the Results.

Methods 1. Substrate and claim formalism

AEP separates three objects that are often conflated:

- **The model:** a fitted or deployed mapping from observed inputs to predictions.
- **The substrate:** the observed data state space on which the claim is made.
- **The claim:** the statement that the model can predict, classify, diagnose, recover, triage, or otherwise support a medical decision for a target population.

A model can perform well globally because easy regions dominate the sample. A substrate can contain matched-state neighborhoods where patients move in opposite future directions. A claim can be too broad if it is not supported in those locally non-identifiable neighborhoods. AEP therefore proceeds in five stages:

- Construct patient-level state-to-future transitions from the substrate.
- Estimate local non-identifiability using PBR and PIM.
- Partition the substrate into identifiable, intermediate, and non-identifiable regions.
- Evaluate model performance globally and inside the non-identifiable region.
- Assign a claim-support verdict and score.

AEP claim-evaluation pipeline: claim → substrate transitions → PBR/PIM estimation → train-defined substrate regions → model zoo evaluation → local floor comparison → AEP verdict.

Methods 2. The identifiability theorem

Prediction with latent context

Let x denote the observed state. Let ϕ denote unobserved biological, clinical, temporal, institutional, or treatment context. Let the future change be generated by

$$\Delta y = f(x, \phi).$$

The model observes x , or some finite representation of x , but not necessarily ϕ . A predictive claim assumes that the observed substrate contains enough information to recover the relevant future direction or outcome. AEP asks whether that assumption is supportable. If two different latent contexts ϕ' and ϕ'' are both compatible with the same observed state x , but imply different futures, then the substrate is locally non-identifiable:

$$f(x, \phi') \neq f(x, \phi'').$$

If the two futures have opposite signs,

$$\text{sign}f_x, \phi' \neq \text{sign}f_x, \phi'',$$

then a single-valued prediction from x alone cannot be generally correct in that local state region.

Local branching

For finite data, exact equality of x is replaced by a local matched-state cell or neighborhood b . Let $D \in \{-1, +1\}$ be the direction of future movement. In cell b ,

$$nb^+ = \#\{i \in b: D_i = +1\}, \quad nb^- = \#\{i \in b: D_i = -1\}, \quad nb = nb^+ + nb^-.$$

Predictive Blindness Risk (PBR) measures, within a matched-state cell, the local mass of opposing future branches: it is at its minimum when the cell is locally one-sided and at its maximum when the cell is evenly divided between opposite directions. Predictive Indeterminacy Measure (PIM) measures the separation between those opposing branches in outcome units; when PIM is large, the matched-state region admits opposite-signed futures with separation larger than the dominant branch magnitude. The operational estimators for both quantities are part of a protected framework and are not reproduced here (see Implementation Boundary); they are available to editors and reviewers under confidentiality.

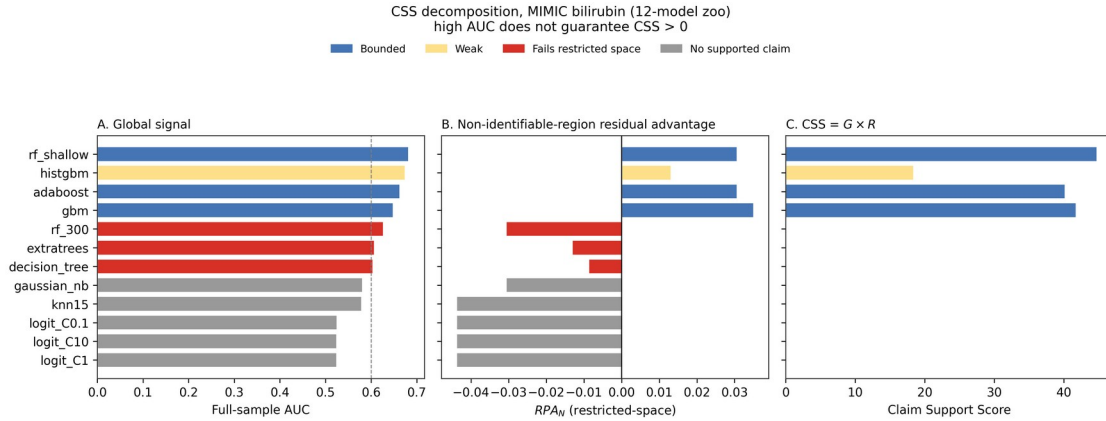
Substrate information sufficiency theorem

Assumption 1 (Observed-state restriction). *A model claim is made using an observed substrate representation x . Within a matched-state cell b , the model does not observe a branch-separating variable sufficient to distinguish ϕ' from ϕ'' . Therefore, for purposes of evaluating the claim on that substrate, the model's prediction is evaluated against the local branch structure induced by x .*

Theorem 1 (Local substrate floor bound). *Within a matched-state cell, any prediction rule that cannot distinguish branch membership is capped by the cell's majority direction. The resulting local substrate floor is therefore determined by the cell's branch composition alone: the more evenly a cell divides between opposing futures, the lower the ceiling on any model operating on that cell, and a maximally balanced cell admits no directional advantage at all.*

Proof. A branch-blind rule inside the cell must commit to one direction without observing the branch-separating latent context; its expected accuracy is therefore the frequency of whichever direction it commits to, and the best such rule achieves the frequency of the majority direction. The closed form relating this floor to the branching quantity is withheld (see Implementation Boundary). ■

For each training cell, AEP estimates branching and indeterminacy and classifies the cell as identifiable, intermediate, or severely non-identifiable. The primary analysis applies a conservative severe-non-identifiability criterion, prespecified before evaluation; sensitivity analyses vary this criterion. The numeric cut-points and their calibration are part of a protected framework (see Implementation Boundary).



Train-defined substrate bins by region, all eight tasks. Non-identifiability is defined from the substrate’s branch composition, not from fitted-model error.

Local substrate floor

For each training cell, the local floor is the cell’s majority direction — the best directional accuracy available to a rule that observes only the matched state. A model must exceed this floor in the non-identifiable test region to support a broad claim; merely exceeding the global majority baseline is not enough. The floor’s construction is part of a protected framework (see Implementation Boundary).

Residual predictive advantage

For model m , region R , and local floor FR , AEP computes

$$RPAR_m = AR_m - FR,$$

where AR_m is the model’s accuracy in region R . In the non-identifiable region N ,

$$RPAR_m = AN_m - FN.$$

This is the central restricted-space quantity. Positive $RPAR$ means the model adds information beyond the local substrate floor. Zero or negative $RPAR$ means the claim is not supported in the non-identifiable region.

Global signal

AEP still records global model performance. In the demonstration, global signal is defined by either:

$$AUC_{full} \geq 0.60$$

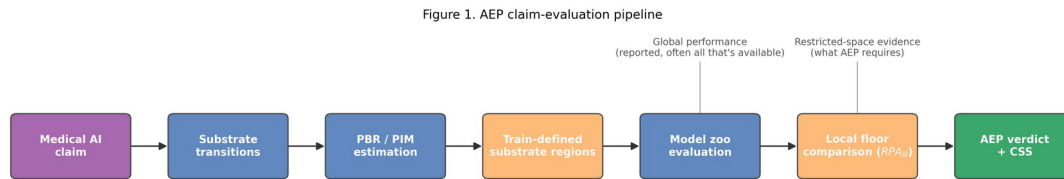
or

$$Accuracy_{full} - Floor_{global} \geq 0.02.$$

But global signal is not sufficient for claim support. It only establishes that the model may have learned a global ranking or global imbalance.

Claim Support Score

AEP reports a continuous Claim Support Score (CSS), but the categorical verdict is primary. CSS combines a global performance signal with restricted-space support, and is constructed so that a model retaining no advantage over the local substrate floor in the non-identifiable region scores zero even when its full-sample AUC is high. This is deliberate: a global signal does not support a broad claim if the model fails where the substrate is locally non-identifiable. The score's functional form and its component scalings are part of a protected framework (see Implementation Boundary).



CSS decomposition for the MIMIC bilirubin task (12-model zoo): global AUC, restricted-space residual advantage, and the resulting Claim Support Score. High AUC does not guarantee a supported claim.

AEP verdicts

The categorical verdicts are: **supported restricted-space claim** (the model retains meaningful advantage over the local substrate floor where the substrate does not resolve the outcome); **bounded support**; **weak restricted support**; **fails restricted space** (a global signal exists but no advantage survives in the non-identifiable region); **no supported claim** (neither a global signal nor restricted-space advantage); and **untested restricted space** (insufficient severely non-identifiable region exists under the prespecified criterion). The numeric bands separating these verdicts are prespecified and are part of a protected framework (see Implementation Boundary).

Step 1: Define the claim target

Let $D \in \{-1, +1\}$ denote whether the patient deteriorates over a specified horizon. Let x denote the observable deployment substrate: the multimodal patient state available at the time of prediction. The claim is therefore:

$$x \mapsto D.$$

The authors may report global AUC, accuracy, or clinician comparison. AEP treats those as global signal, not as claim support.

Step 2: Estimate substrate non-identifiability. Using a deployment-like validation substrate, matched-state regions are formed and their branching and indeterminacy estimated. Where branching is high, patients with similar observed multimodal states go on to deteriorate and to recover in substantial opposing proportions; where indeterminacy is high, the same observed state admits opposite-signed future trajectories. This step answers a question the global model-evaluation literature does not: does the deployment substrate itself locally support the claimed prediction?

Step 3: Define the restricted-space test. The substrate is partitioned into identifiable and severely non-identifiable regions, and each non-identifiable region carries a local substrate floor — the best directional accuracy available to a rule that observes only the matched state. The relevant question is not whether the model beats clinicians globally; it is whether it beats this local floor where the substrate does not resolve the outcome. The estimators, partition criteria, and floor construction are part of a protected framework (see Implementation Boundary).

Step 4: Evaluate the LLM if outputs are available

If the closed LLM can be queried or if the authors provide patient-level predictions, its restricted-space residual predictive advantage is

$$RPAN = ANLLM - FN,$$

where ANLLM is the LLM's accuracy in the non-identifiable region and FN is the corresponding local substrate floor. A broad deployment claim is supported only if

$$RPAN > 0,$$

and strongly supported only if RPAN remains positive and stable across repeated group splits, deployment sites, and plausible substrate partitions. If the LLM has high global AUC but

$$RPAN \leq 0,$$

the appropriate finding is:

The model may have global discrimination, but the broad claim is not supported in the non-identifiable region of the deployment substrate.

Step 5: Evaluate the claim if outputs are unavailable

If the LLM is inaccessible and no patient-level predictions are provided, RPAN cannot be computed directly. The evidentiary burden of the claim can still be evaluated. If the deployment substrate has high PBR and PIM, a global AUC or clinician-comparison statistic is insufficient. The authors must show one of the following:

- The model reduces local non-identifiability by using additional branch-separating information not available to the comparator.

- The model retains positive residual advantage over the local substrate floor inside the high-PBR region.
- The claim is restricted to identifiable regions where the substrate supports prediction.

Without such evidence, the conclusion is not that the LLM cannot work. It is that the claim is not externally supportable as stated.

Step 6: Interpret clinician comparison correctly

Suppose the LLM beats clinician accuracy globally:

$A_{fullLLM} > A_{fullclinician}$.

This does not by itself establish deployment support. The question is whether the comparison also holds in the non-identifiable region:

$A_{NLLM} > A_{Nclinician}$,

and whether the LLM exceeds the local substrate floor:

$A_{NLLM} > FN$.

These are different tests. A clinician may be worse globally but better in specific ambiguous regions, or both clinician and LLM may fail to exceed the local floor. Three claims are therefore kept separate:

- The LLM beats clinicians globally.
- The LLM beats clinicians in non-identifiable regions.
- The LLM exceeds the local substrate floor in non-identifiable regions.

Only the third establishes substrate-local claim support.

Step 7: Deployment verdicts

Finding	Deployment interpretation
High global AUC and positive RPAN	The broad deployment claim is supported in the non-identifiable region.
High global AUC but weak RPAN	The claim is bounded; deployment should be limited or qualified.
High global AUC but $RPAN \leq 0$	The global claim fails restricted-space support.
High PBR/PIM but no patient-level model outputs	The claim is not externally supportable unless restricted-space evidence is disclosed.
No severe non-identifiable region under prespecified thresholds	The restricted-space claim is untested, not assumed supported.

Conclusion of the example

The evidentiary question this example turns on is not whether the LLM can achieve an impressive global score. It is whether the claim remains supported where the deployment substrate is locally ambiguous. For a closed multimodal LLM, the appropriate conclusion is:

The model may be useful, but the published claim is supportable only if the authors demonstrate residual advantage over the local substrate floor in the non-identifiable region, or restrict the claim to regions where the substrate is identifiable.

This is the role of AEP in deployment governance: identifying when the evidence supplied for a medical AI claim is incomplete, without requiring external access to the hidden weights behind it.

Supplementary Appendix B: Why the Clinician Is Not an Exception

The clinician who outperforms a model does not violate the substrate bound; the clinician conditions on a different and richer x . At the level relevant to clinical reliability, a transformer, CNN, random forest, and logistic regression are all computing $y=hx;\theta$, differing only in the richness of h , the size of θ , and the dimensions of x they exploit. Modern AI is extraordinary because it massively expands the feasible class $h \in H$ and finds relations humans cannot. It still cannot condition on variables absent from x .

The clinician's advantage is not magic intuition but unrecorded information acquisition. The record holds

x_{EHR} =labs, imaging, medications, diagnoses,

while the encounter exposes

x_{clinical} = x_{EHR} , behavior, appearance, family report, adherence, subtle change, context.

Two patients with identical recorded state, $x_A=x_B$, need not share a true state, $x_A \neq x_B$, because the hidden state carries information the substrate omits.

Consider two diabetic patients, both recorded at HbA1c 9.0, same age, BMI, and medication. The model sees $x_1=x_2$ and learns $PY|x$, which branches. The clinician observes that the first understands the regimen, exercises, and had a temporary disruption, while the second does not take the medication and lacks support. The clinician has expanded $x \rightarrow x^*$, and under the richer conditioning the branch contracts:

$HY|x^* < HY|x$.

The entropy falls not because humans escape AEP but because they changed the substrate.

"Intuition" is then approximately $h_{\text{clinician}}x^*;\theta_{\text{experience}}$, a learned model over variables the database never captured—less an oracle than a better-provisioned x .

This is why the encounter-type table matters: the visit is an information-generating process. A routine visit, an emergency admission, and an exploratory oncology consult are not three noisy samples from one distribution; they create different x -spaces,

xroutine≠xemergency≠xexploratory, and therefore different achievable prediction boundaries—fixed before any model is trained. AI capability determines how much information can be extracted from a substrate; AEP asks how much the substrate contains to extract. The field has heavily optimized the first axis. AEP measures the second.

Competing interests

The theoretical frameworks, methods, and computational systems described herein are the subject of prior intellectual property predating this manuscript and the author's current institutional appointment. The author declares no other competing interests.

Implementation Boundary

This study evaluates the identifiability structure of a dataset as a scientific metric. It reports the conceptual workflow, validation strategy, and empirical findings. It does not disclose production implementations, automated hypothesis extraction systems, agent architectures, ranking procedures, prompt systems, substrate-selection engines, or deployment infrastructure.

The information-sufficiency metrics presented here are computed as pre-model diagnostics of local separability, residual entropy, temporal placement, trajectory departure, reversal potential, and residual advantage over local substrate structure. Full operational estimators are part of a protected framework and are available for review under appropriate conditions, including to editors and reviewers under confidentiality.