Preprints.org

Article

Not peer-reviewed version

Disentangling Retention and Transfer in Resource-Aware Clinical AI Simulation

Gyako Sani Adamu , Augustina C. Amaefule , Godson N Ozioma , Bankole I Oladapo *

Posted Date: 5 November 2026

Keywords: clinical reasoning; artificial intelligence; simulation education; skill transfer; resource accounting

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Disentangling Retention and Transfer in Resource-Aware Clinical AI Simulation

Gyako Sani Adamu [1], Augustina C. Amaefule [2], Godson N Ozioma [3] and Bankole I Oladapo [4,*]

[1] Chemistry Department, College of Science and Engineering, University of Leicester, University Rd, Leicester LE1 7RH, United Kingdom

[2] Department of Management and Business, University of Gloucestershire, UK

[3] Department of Software Engineering, The University of Salford, 43 Crescent, Salford, Greater Manchester, M5 4WT, United Kingdom

[4] School of Engineering and Sciences, Solihull College & University Centre Woodlands Campus, Auckland Drive, Smith's Wood, Solihull, B36 0NF, UK

* Correspondence: bioladapo@abuad.edu.ng

Abstract

Clinical AI education should develop judgement that persists and transfers across systems, yet changing the AI system only at follow-up confounds transfer with forgetting. We developed a crossed assessment framework and evaluated its statistical properties using 5,000 Monte Carlo replications per scenario. Sixty synthetic learners per replication completed four system-by-time assessment cells, with 15% independent attrition, 20 decisions per cell and a shared learner effect. Familiar-system accuracy declined from 0.75 to 0.67; unfamiliar-system penalties ranged from zero to ten percentage points. When the penalty was five points, the diagonal comparison misattributed an additional 4.98 points to forgetting and achieved only 30.42% interval coverage of retention. Within-system retention estimates maintained approximately 95% coverage. A deterministic example showed that improved override alone can conceal complete rejection of safe advice. An accompanying resource audit corrected illustrative expenditure calculations while identifying unsupported energy comparisons. These computational findings support identifiable assessment contrasts and separate acceptance and override outcomes. They do not demonstrate adaptive-training effectiveness, clinical benefit or measured environmental savings; those claims require prospective validation.

Keywords: clinical reasoning; artificial intelligence; simulation education; skill transfer; resource accounting

1. Introduction

Clinical artificial intelligence changes the information environment in which diagnostic decisions are made. A recommendation can supply a useful alternative, direct attention to neglected evidence, or make an incorrect interpretation appear more credible. The educational challenge is therefore to develop selective reliance: accepting useful assistance while detecting advice that warrants verification, rejection, or escalation. This requires observable judgement under uncertainty. Knowledge of model limitations, satisfaction with an AI tool, and performance while assistance is available measure different constructs and should not be treated as interchangeable evidence of clinical competence.

Randomized studies illustrate why this distinction matters. In Goh and colleagues' diagnostic reasoning trial, access to GPT-4 produced no significant improvement over conventional resources among 50 physicians, with an adjusted between-group difference of two percentage points [1]. A subsequent trial involving 92 physicians found an advantage for GPT-4 assistance on management reasoning tasks [2]. Qazi and colleagues reported substantially improved diagnostic reasoning when physicians who had completed a common AI-literacy curriculum received LLM access [3]. These

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

findings establish that assisted performance depends on the task, comparator, population, and interaction conditions. They provide no direct estimate of whether a particular teaching approach creates durable independent skills.

The safety question becomes sharper when advice is deliberately incorrect. Among physicians trained in AI literacy, Qazi and colleagues found lower reasoning accuracy when LLM recommendations contained introduced errors than when recommendations were error-free [4]. Earlier experiments in radiology and pathology likewise showed that incorrect recommendations could impair human diagnostic performance [5,6]. Training that encourages verification is consequently plausible, but its benefit must be demonstrated without rewarding indiscriminate rejection of useful advice. An educational intervention that raises the override rate while reducing appropriate acceptance may merely shift the direction of error.

A motivating ADAPT-AI design proposes to compare didactic teaching, fixed simulation, and adaptive simulation with deliberately fallible assistance, then examine performance after six months. An assessment sequence that changes time, clinical cases, and AI system together creates ambiguity. A lower follow-up score could reflect forgetting, a harder case set, unfamiliar presentation of advice, a different model's error characteristics, or several mechanisms simultaneously. The resulting difference cannot identify a distinct transfer penalty. This is an experimental-design problem that more elaborate regression alone cannot resolve.

This study develops and computationally examines a revised assessment framework that crosses assessment time with AI-system familiarity and preserves separate denominators for correct and unsafe recommendations. Its contribution is methodological: an identifiable way to distinguish retention from transfer, paired measures that expose both overreliance and underreliance, and a resource-accounting boundary that makes platform comparisons interpretable. A reproducible numerical experiment evaluates the consequences of design choices under stated assumptions. It does not estimate the efficacy of adaptive simulation in clinicians. The subsequent human study is specified as a prospective validation route aligned with medical education, patient safety, and healthcare AI research.

2. Literature Review

2.1. Assisted Performance and Susceptibility to Advice

Clinical human–AI performance is heterogeneous. In skin-lesion recognition, Tschandl and colleagues showed that the presentation of AI support and the reader's expertise shaped collaboration outcomes [7]. Kiani and colleagues' pathology experiment found opposing effects according to whether the model's prediction was correct [6]. Such results support evaluating the joint system at the level of individual advice episodes. Average accuracy alone can obscure an institutionally relevant weakness when a relatively small subset of persuasive, incorrect recommendations produces serious errors.

Jabbour and colleagues provide a related caution about explanations. Their multicentre vignette study found that standard predictions could improve diagnosis, whereas systematically biased predictions impaired it; image explanations did not reliably neutralize the adverse effects [8]. Gaube and colleagues manipulated both advice accuracy and its attributed source. Inaccurate advice reduced diagnostic accuracy whether it was described as AI-generated or as coming from a human expert [5]. The latter experiment used human-produced advice, which limits claims about the intrinsic properties of contemporary LLMs. It nevertheless demonstrates that an authoritative source label and the substance of a recommendation are distinct experimental factors.

Evidence also argues against equating caution with adequate calibration. In a randomized trial of 186 medical students, Kıyak and colleagues found that a warning about AI fallibility did not change advice-taking; participants already gave relatively little weight to the advice [9]. Teng and colleagues found that plausible misinformation reduced novice students' accuracy, while correct explanations provided no significant accuracy advantage in their selected tasks [10]. Together, these

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

studies motivate measuring acceptance and override separately. They do not imply that every learner overrelies on AI or that every explanation is harmful.

2.2. Simulation, Reflection, Retention, and Transfer

The educational rationale for targeted practice has precedents beyond AI. Wayne and colleagues used simulation and deliberate practice to bring internal medicine residents to predefined resuscitation mastery standards [11]. Mamede and colleagues found that structured diagnostic reflection could counter availability-related errors in a short experimental setting [12]. These interventions support the plausibility of practising explicit checking behaviours. Their outcomes concern particular clinical tasks, and their effects cannot be imported as expected effect sizes for adaptive AI training.

Retention and transfer require their own assessment. Larsen and colleagues demonstrated better recall beyond six months after repeated testing with feedback than after repeated study [13]. Fraser and colleagues found that students acquired and retained some simulator-trained skills while showing limited transfer between clinical presentations [14]. The implication is a measurement requirement: immediate success, delayed success on a familiar task, and application in an unfamiliar environment are empirically separable. A clinical AI curriculum should specify which of these it seeks to improve and should provide assessment contrasts that identify each outcome.

Cognitive forcing offers a further mechanism. Buçinca and colleagues reported reduced overreliance when interfaces prompted more deliberative engagement, although participants found these designs more effortful and less preferable [15]. This was a non-clinical experiment. It supports testing a mechanism rather than assuming clinical effectiveness. Similarly, a retrieval-augmented medical teaching assistant showed context-dependent student usage across two cohorts [16]; adoption data establish practical interest but do not demonstrate resistance to misinformation or retained reasoning gains.

The broader simulation literature supports feedback, repeated practice and integration with curricula [17,18]. Critical reviews distinguish skill acquisition, maintenance and transfer [19], while comparative evidence on deliberate practice supports an active instructional comparator [20]. These syntheses establish educational design principles, with heterogeneous interventions and outcomes. They cannot supply an expected effect for a fallible-AI tutor. Equating generic simulator efficacy with resistance to persuasive algorithmic errors would bypass the mechanism that the proposed curriculum needs to test.

2.3. The Methodological Gap

The specific opportunity is to evaluate selective reliance across both time and system familiarity while retaining a transparent account of what was measured. Correct and unsafe advice need independent reference labels and comparable assessment opportunities. Parallel case forms require calibration and counterbalancing so that changing case content does not masquerade as transfer. Infrastructure evaluation requires an explicit functional unit, including preparation, expert verification, delivery, and maintenance. Whether local inference offers a resource advantage is an empirical question whose answer depends on utilization and system boundaries. The proposed framework makes these questions testable while keeping simulated design findings separate from future clinical validation.

3. Methodology

3.1. Study Design and Evidential Status

We conducted a computational study of assessment design, complemented by an arithmetic audit of an illustrative deployment budget. No clinicians were recruited, no patient records were analysed, and no educational intervention was delivered. The investigation follows the aims, data-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

generating mechanisms, estimands, methods and performance measures framework for simulation research [21]. Its purpose is to determine what an evaluation can identify before investing in an adaptive training trial. The motivating ADAPT-AI concept included example records and programmed intervention benefits; these were not treated as observations. All numerical assessment-success probabilities introduced here are analyst-selected stress-test inputs, rather than effect estimates from clinical studies.

The proposed future trial would allocate senior medical students and early postgraduate doctors equally to didactic instruction, fixed simulation or adaptive simulation. Each arm would receive four hours of teaching over four weeks, access to the same verified clinical references and equivalent opportunities for debriefing. Fixed and adaptive simulation would draw from the same eligible case pool; selection order and targeting would differ. Allocation would be concealed and stratified by institution, training stage and baseline reasoning performance. Assessors would be blinded to allocation, although learners and teaching staff could not be. Table 1 distinguishes this prospective educational comparison from the computational experiments actually completed.

Table 1. Prospective educational trial and completed computational evaluation.

Component	Proposed educational trial	Completed computational study
Participants	Senior students and early postgraduate doctors	Synthetic learner cohorts only
Arms	Didactic, fixed simulation, adaptive simulation	No randomized intervention arms
Teaching dose	Four hours over four weeks in each arm	No teaching intervention simulated
Assessment	Both systems immediately and at six months	Four generated cells per learner
Inference	Educational effects pending human validation	Bias, RMSE and coverage under assumptions

3.2. Separating Retention from Transfer

Let p(a,t,m) denote the probability of an appropriate decision in arm a, at immediate assessment t=1 or six months t=2, with familiar system m=F or unfamiliar system m=U. For a future evaluation, the assessment crosses both systems with both times, using different, counterbalanced case forms. Familiarity refers to prior training exposure; system U is introduced at the immediate assessment and receives no intervening training. The delayed U contrast therefore measures maintenance on the transfer system after its first assessment exposure. Assessing transfer to a completely new system at six months would require an additional held-out system or a randomized assessment-exposure group. This distinction prevents assessment itself from being mistaken for an absence of exposure.

$$R_a = p_{a,2,F} - p_{a,1,F}\,;\; T_a = p_{a,2,U} - p_{a,2,F}$$
$$D_a = p_{a,2,U} - p_{a,1,F} = R_a + T_a \quad (1)$$

Equation (1) defines retention R within the familiar system, contemporaneous transfer T, and the diagonal comparison D used when the system changes only at follow-up. The identity D=R+T shows why a diagonal estimate cannot separately identify skill decay. This is an identifiability problem: increasing the number of learners does not reveal the two unobserved cells. A positive transfer effect could mask deterioration, whereas a negative transfer effect could exaggerate it. Figure 1 shows the fully crossed assessment. Its fourth cell also permits investigation of a time-by-system interaction. Equivalent forms and randomized system order are needed because crossing alone does not remove practice, fatigue or case-difficulty effects.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

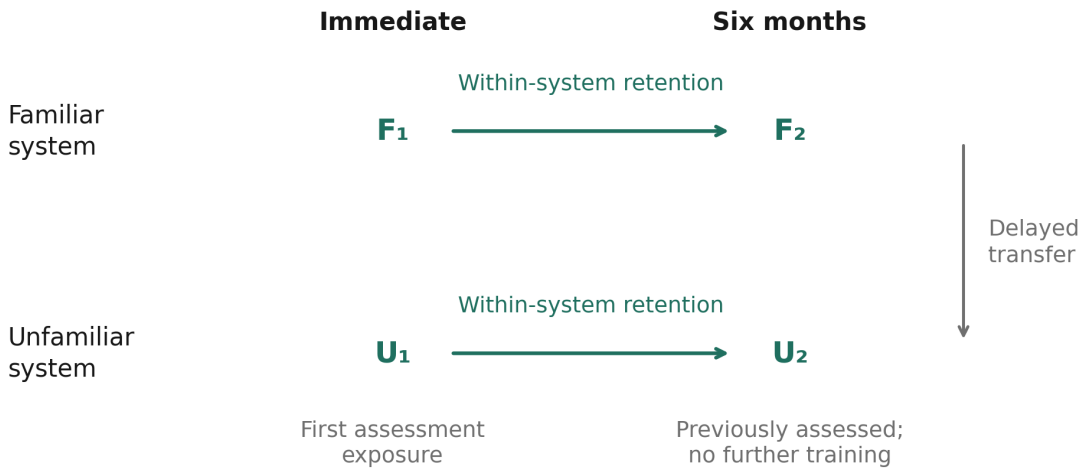

Figure 1. Prospective assessment design crossing time and system familiarity. Retention compares familiar-system performance across time; delayed transfer compares systems at the same time. The unfamiliar system has already been assessed immediately after training. Counterbalanced parallel forms are a future design requirement, not a simulated intervention.

For a future trial, the primary educational estimand would be the adaptive-versus-fixed difference in appropriate decision probability on system U at six months, standardized to an explicitly balanced assessment bank. Retention and transfer contrasts would be secondary estimands, with arm-specific uncertainty and no presumption that the adaptive intervention succeeds. A clinically acceptable loss in appropriate acceptance must be specified before recruitment if a noninferiority safety condition is used; the present study does not select a margin from simulated outcomes.

3.3. Decision Outcomes and Statistical Model

Advice is adjudicated as safe or unsafe before learner exposure by a multidisciplinary panel using dated, jurisdiction-appropriate sources. Appropriate override rate (AOR) is the proportion of unsafe recommendations correctly rejected or corrected. Appropriate acceptance rate (AAR) is the proportion of safe recommendations appropriately used. Rejection alone is insufficient for correctness: the proposed alternative action must also satisfy the case rubric. Requests for additional information and escalation are scored against case-specific criteria, with unresolved responses retained as a separate category rather than silently discarded. Diagnostic reasoning, time to a safe action and potentially harmful alternatives are recorded alongside advice handling.

$$B = \frac{AOR + AAR}{2};\ A(w) = w\,AOR + (1 - w)\,AAR$$

$$L(w) = w\,h_U(1 - AOR) + (1 - w)\,h_S(1 - AAR) \quad (2)$$

Balanced appropriate decision accuracy B gives equal weight to AOR and AAR. Deployment-weighted accuracy A(w) depends on the unsafe-advice prevalence w, which an enriched examination cannot estimate. Neither statistic encodes severity. A complementary loss function assigns prespecified weights hU and hS to accepting unsafe advice and unnecessarily rejecting safe advice. These weights require clinical and patient input; no universal ratio is implied. Both components and their denominators should always accompany a composite result.

A future participant-by-case analysis would use a binomial mixed model with intervention, time, system and their interactions, advice validity, baseline score and training stage as fixed effects, and crossed learner and case intercepts [22,23]. Institution effects would be fixed when sites are few. Random slopes would be prespecified where replication supports their estimation, with convergence and singularity diagnostics reported.

$$logit\ Pr(Y_{ijtm} = 1) = Z^T\beta + b_i + c_j + \eta_{s(i)} \quad (3)$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

Here Z is the fixed-effect design vector, b is learner heterogeneity, c is case heterogeneity, and s indexes institution. Marginal risk differences, obtained by standardizing predictions over the intended case distribution, would be the primary presentation; conditional odds ratios would be secondary. This future model was not fitted to fabricated trial outcomes. The present Monte Carlo experiment deliberately uses transparent paired learner contrasts to isolate the design issue, conditional on a fixed case bank.

Confidence should be elicited as a probability that the submitted action is correct before feedback, including on incorrect decisions. The mean squared probability error is the Brier score [24]. It combines calibration and discrimination and should not be renamed a pure confidence-calibration index. Reliability plots, calibration intercept and slope, and the frequency of high-confidence unsafe actions would complement it [25]. Confidence was not generated from realized correctness in this study, since doing so would build a desired association into the data.

3.4. Monte Carlo Experiments

The main experiment used 5,000 replications per scenario and seed 20261103. Each replication enrolled 60 synthetic learners and independently retained each learner with probability 0.85. All four cell scores were generated for every learner; the same independent follow-up mask restricted all reported contrasts to complete pairs. Immediate unfamiliar-system scores instantiated the design but were not used in the delayed-transfer contrast. No intervention assignment or between-arm training effect was simulated. Retained learners contributed 20 decisions in each of four system-by-time cells, conceptually comprising ten safe and ten unsafe items. Safe and unsafe items shared the same success probability in this deliberately simplified experiment; the separate endpoint illustration below relaxes that equality. A shared normal learner intercept with standard deviation 0.8 induced dependence across repeated decisions. Cell intercepts were calibrated by Gauss–Hermite quadrature so that probabilities were marginal targets, rather than conditional logistic probabilities mistakenly reported as population averages.

$$X_{ic} \mid b_i \sim Binomial(m, \sigma(\alpha_c + b_i)), b_i \sim N(0, 0.8^2)$$

$$E_b[\sigma(\alpha_c + b)] = p_c \; ; \; \sigma(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

Immediate familiar-system performance was set to 0.75 and delayed familiar-system performance to 0.67. Transfer penalties of 0, 0.05 and 0.10 were applied at both times, so the true retention contrast remained −0.08 throughout. These values span interpretable departures from zero transfer and are not calibrated to an Omani population. Conditional on the shared learner effect, decision counts were independently binomial. The simplified generator excludes case heterogeneity, differential forgetting and informative dropout; its confidence intervals therefore concern learners facing the stipulated assessment bank, not all possible clinical presentations.

Table 2. Monte Carlo inputs and provenance.

Input	Specified value	Basis
Replications per scenario	5,000	Precision of computational evaluation
Enrollment per replication	60	Motivating design, one hypothetical arm
Follow-up loss	15%, independent	Assumed MCAR sensitivity setting
Decisions per cell	20 (10 safe, 10 unsafe)	Balanced hypothetical bank
Learner logit SD	0.8	Assumed heterogeneity
Familiar probability T1 / T2	0.75 / 0.67	Assumed marginal assessment success
Unfamiliar penalties	0 / −0.05 / −0.10	Assumed probability differences
Random seed	20261103	Reproducibility setting

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

All inputs are design assumptions; none are estimated ADAPT-AI clinical effects. MCAR: missing completely at random. SD: standard deviation.

For each retained learner, paired differences estimated D, R and delayed T. Two-sided 95% Student t intervals used the empirical standard deviation of the learner differences and the actual retained sample size. Bias, root mean squared error (RMSE), empirical standard deviation and interval coverage were evaluated against each intended estimand. When D was interpreted as retention, its coverage target was R, rather than D itself. This distinction is essential: a statistically precise interval can still answer the wrong question. Monte Carlo standard errors quantified simulation uncertainty. In Equation (5), K is the replication count and $\hat{C}$ is the estimated coverage probability.

$$Bias = \frac{1}{K}\sum_{k=1}^{K}(\widehat{R_k} - R)\,;\; RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\widehat{R_k} - R)^2}$$

$$MCSE(\hat{C}) = \sqrt{\frac{\hat{C}(1-\hat{C})}{K}}\,;\; K = 5000 \quad (5)$$

The simulation retains the motivating cohort size to investigate its consequences, rather than claiming that 60 learners per arm provides adequate trial power. For reference, a conventional independent two-proportion approximation for 0.45 versus 0.65, two-sided alpha 0.05 and 80% power, requires approximately 96 evaluable learners per group, or 113 enrolled after 15% attrition. Repeated cases, baseline adjustment, case clustering and a different primary endpoint change that calculation. A definitive trial requires simulation across clinically justified effect sizes, learner and case variances, and missingness assumptions. The exploratory computational experiment is not a final recruitment calculation.

3.5. Adaptive Selection and Clinical Safeguards

The adaptive policy would select cases using estimated error susceptibility, information gain and coverage constraints, rather than equating adaptation with a large language model. A candidate constrained policy is given in Equation (6).

$$j_{next} = arg\,max_{j \in E}\{ I_{ij} + \lambda\, e_{ij} - \mu\, r_{ij} \} \quad (6)$$

For learner i and candidate case j, I represents information about a prespecified reasoning skill, e the estimated vulnerability to the relevant failure mode, and r prior exposure. The eligible set E excludes assessment cases, caps repeated specialties and enforces minimum exposure to safe advice. Coefficients are policy settings requiring pilot evaluation, not fitted findings. A small pilot cannot reliably identify discrimination, difficulty and guessing parameters for an unrestricted three-parameter item-response model; transparent rules or a simpler calibrated model should precede such complexity [26]. Logging eligible cases, policy versions, selection probabilities and overrides would permit later auditing of who received which learning opportunities.

Failure modes would include premature closure, ignored contraindications, omitted counter-evidence, unsupported citation claims and calculation errors. Teaching examples would contain matched safe variants so that clinical context, rather than a recognizable error cue, determines the correct response. Illustrative prescribing examples should not be reused as universal clinical answers. In particular, a brief vignette cannot establish every patient's diagnosis or justify a uniform investigation or antimicrobial regimen. Qualified educators must validate each case locally, explain deliberately misleading content during debriefing and ensure that learners cannot export it into patient care. The adaptive policy itself was not executed or validated in the computational experiment.

3.6. Resource Accounting and Reproducibility

Resource comparison uses a stated service boundary: content development, validation, faculty debriefing, implementation, maintenance, inference and infrastructure availability. Wall-clock

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

learner time is distinguished from active server time, which depends on concurrency, batching and response latency. Full-system electricity is integrated over measured power, with idle energy allocated by the institution's actual sharing or scheduling policy. PUE is applied only where the measurement boundary excludes facility overhead [27,28]. Cloud energy is recorded as unavailable unless a provider supplies a compatible workload-specific allocation. An API invoice cannot substitute for an electricity meter.

$$E = PUE \int P_{sys}(t)\, dt$$

$$C_{annual} = \frac{K_{cap}}{L} + C_{staff} + C_{support} + p_E E + C_{API} \quad (7)$$

In Equation (7), power is in kW and time in hours; Kcap is capital expenditure and L is hardware lifetime in years. Annual cost separates capital amortization, staff, support, electricity and metered API charges. Carbon accounting additionally requires location- and time-relevant emission factors and an explicitly separate embodied-hardware boundary [29]. The illustrative planning document's token volumes, hypothetical prices and electrical assumptions were recomputed without representing them as current quotations or measurements. Cost and energy remain separate outcomes; lower expenditure does not establish lower emissions. Code, scenario parameters, replication summaries and figure sources are retained in the accompanying reproducibility package. The prospective trial should register its protocol and analysis plan, adapting SPIRIT-AI and CONSORT-AI reporting items as appropriate [30,31].

4. Empirical Results from Computational Experiments

4.1. Identifiability and Estimator Performance

The deterministic design-matrix check confirmed the structural problem. With columns for intercept, time and unfamiliar-system exposure, the diagonal design had rank two for three parameters. The fully crossed design had rank three, allowing separate additive time and system effects; its fourth cell also permits a time-by-system interaction. Three appropriately chosen cells suffice for familiar-system retention and delayed transfer alone. No regression estimator can recover separately identified time and system coefficients from the diagonal observations without external assumptions.

Across 5,000 replications per scenario, approximately 51 of 60 synthetic learners contributed complete pairs. Table 3 reports performance in percentage points. With no transfer penalty, the diagonal comparison estimated retention at −7.96 points and achieved 95.20% coverage of the true −8-point change. The crossed retention estimator produced −7.94 points, RMSE 1.93 points and 95.04% coverage. Under this special condition, both comparisons addressed effectively the same target. Similar numerical performance here should not be interpreted as evidence that system familiarity is immaterial in actual clinical assessments.

When the unfamiliar-system penalty was −5 points, the diagonal mean became −12.98 points. Interpreting that estimate as retention introduced −4.98 points of bias, increased RMSE to 5.34 points and reduced 95% interval coverage of retention to 30.42%. The crossed estimator remained near the specified retention change: −8.00 points, RMSE 1.91 points and 95.30% coverage. With a −10-point transfer penalty, diagonal bias approached −9.96 points and retention coverage fell to 0.22%; crossed retention coverage remained 94.98%. Figure 2 presents the divergence in mean estimates alongside coverage.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

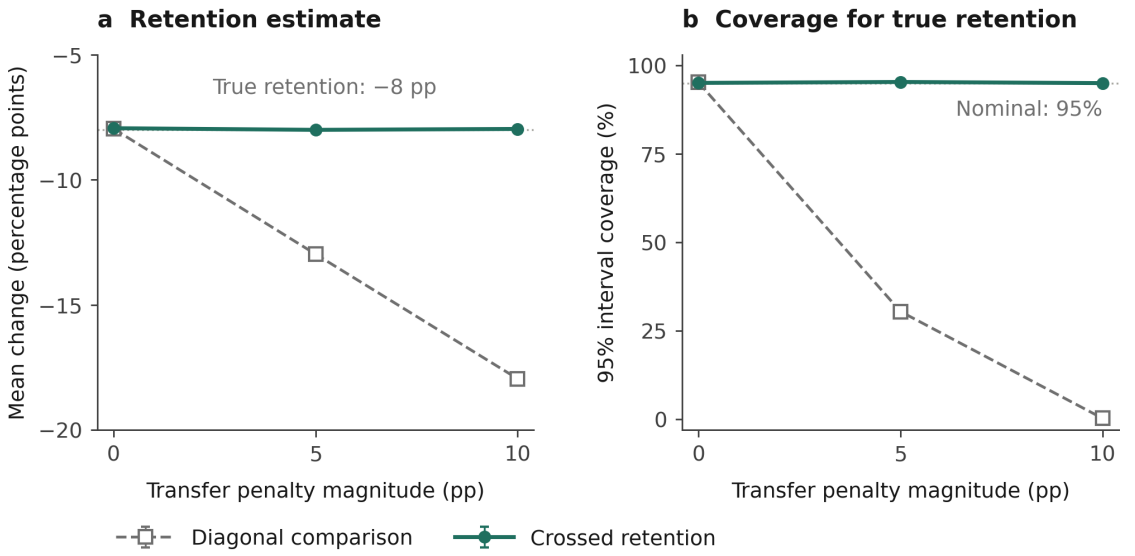

Figure 2. Computational comparison under a fixed −8-point familiar-system retention change and increasing unfamiliar-system penalty. Means and 95% interval coverage come from 5,000 replications per scenario, 60 enrolled learners, independent 15% attrition and 20 decisions per cell. The coverage target is retention. Monte Carlo uncertainty is reported in the text; plotted means are not observed clinical effects.

These results reflect an estimand mismatch, rather than a generic failure of paired confidence intervals. Evaluated against its own compound target, the diagonal estimator retained coverage of 95.20%, 95.22% and 95.28% across the three scenarios. Its mean interval width increased only from 7.68 to 8.15 points. Thus an apparently well-behaved statistical analysis could support a misleading educational interpretation if time-plus-system change were labelled forgetting. The simulation confirms the algebraic decomposition while quantifying its finite-sample implications.

The crossed delayed-transfer estimator recovered the stipulated transfer contrasts: −0.02, −4.97 and −9.99 points, with respective coverage of 94.92%, 94.92% and 95.30%. Monte Carlo standard errors for mean retention bias were approximately 0.03 percentage points. Coverage standard errors were about 0.30–0.31 points for the crossed retention estimates and 0.65 points for diagonal coverage in the −5-point scenario. The 0.22% diagonal retention coverage in the strongest-penalty scenario had a Monte Carlo standard error of 0.07 points. Simulation noise therefore does not explain the material differences between comparisons.

Table 3. Retention estimator performance across 5,000 replications per scenario.

Penalty (pp)	Comparison	Mean (pp)	Bias (pp)	RMSE (pp)	Coverage (%)
0	Diagonal as retention	-7.96	0.04	1.91	95.20
0	Crossed retention	-7.94	0.06	1.93	95.04
-5	Diagonal as retention	-12.98	-4.98	5.34	30.42
-5	Crossed retention	-8.00	0.00	1.91	95.30
-10	Diagonal as retention	-17.96	-9.96	10.16	0.22
-10	Crossed retention	-7.97	0.03	1.92	94.98

True familiar-system retention: −8 pp in every scenario. Coverage is for paired-learner 95% t intervals targeting retention. Fixed-bank inference; approximately 51 complete pairs per replication. pp: percentage points; RMSE: root mean squared error.

4.2. Why Override Alone can Misclassify Educational Success

A deterministic counterexample demonstrates the endpoint problem independently of the stochastic generator. Consider a selective evaluator with AOR 70% and AAR 90%, compared with an

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

always-reject strategy with AOR 100% and AAR 0%. Their balanced accuracies are 80% and 50%, respectively. Ranking curricula by override alone would select the always-reject strategy despite its complete failure to use safe assistance. These values are illustrative decision rules, not measured learner groups. This example concerns advice classification and does not establish correct replacement management.

As Figure 3 shows, their prevalence-weighted accuracies also depend on the advice mixture. The selective evaluator has A(w)=0.90−0.20w, whereas always-reject has A(w)=w. The curves cross at w=0.75. Below 75% unsafe advice, the selective evaluator is more accurate under equal error severity. A test composed entirely of unsafe recommendations would reverse that ranking. This threshold is specific to the chosen example and severity convention; it is not a proposed clinical operating threshold. Reporting conditional rates preserves the information needed to evaluate alternative deployment distributions and consequences.

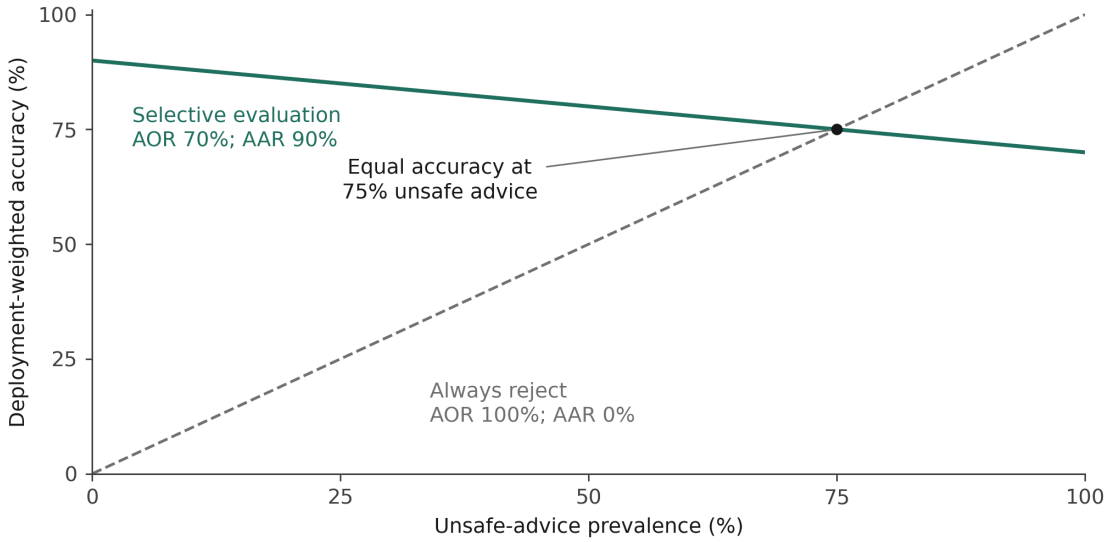

Figure 3. Deterministic decision-rule example under equal error severity. The selective evaluator has AOR 70% and AAR 90%; always-reject has AOR 100% and AAR 0%. Lines show prevalence-weighted accuracy, not observed participant outcomes. Their crossing at 75% unsafe advice is specific to these stipulated values. AOR: appropriate override rate; AAR: appropriate acceptance rate.

4.3. Resource Audit Under Illustrative Budget Assumptions

The illustrative annual workload comprises 36,000 calls, 90 million input tokens and 28.8 million output tokens. At the illustrative assumed prices of USD 3 and USD 15 per million input and output tokens, annual API expenditure is USD 702, rather than USD 8,200. With USD 5,000 annual administration, three-year cloud expenditure becomes USD 17,106. These figures exclude common educational-content costs and any additional provider charges; they are an arithmetic scenario, not a procurement quotation.

The assumed local schedule yields 6,048 kWh active energy and 3,116.4 kWh idle energy, totalling 9,164.4 kWh annually. At USD 0.15/kWh, electricity costs USD 1,374.66 annually. Adding USD 42,000 capital, USD 18,000 maintenance and USD 4,000 support yields USD 68,123.98 over three years (Table 4). The calculation remains conditional on the unsupported equivalence between learner-hours and active server-hours. The cloud estimate of 415.8 kWh follows only from an assumed token-energy coefficient, not provider telemetry. Consequently, neither the apparent energy ratio nor a claim of environmental superiority is an empirical finding of this study.

Table 4. Arithmetic audit under illustrative source assumptions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

Quantity	Source amount	Recomputed amount
Annual API charge (USD)	8,200.00	702.00
Annual local electricity (USD)	1,200.00	1,374.66
Three-year cloud cost (USD)	39,600.00	17,106.00
Three-year local cost with support (USD)	67,600.00	68,123.98
Annual local schedule energy (kWh)	9,164 (rounded)	9,164.4

Recomputed values apply the stated hypothetical inputs only. Cloud administration: USD 5,000/year. Local capital: USD 42,000; maintenance: USD 18,000/three years; support: USD 4,000. Common educational costs are excluded. Local energy is a schedule calculation; cloud energy is unmeasured.

5. Discussion

The study identifies a design requirement for evaluating durable clinical judgement: the assessment must observe the combinations needed to distinguish time from system change. The underlying decomposition is elementary, but its consequences are clinically material. A programme might appear to lose effectiveness because the follow-up model is harder to supervise. Conversely, a more helpful follow-up system might conceal deterioration in independent judgement. The Monte Carlo experiments quantify those errors under controlled conditions; they do not establish the size or direction of transfer effects in practice.

This changes how an adaptive simulation trial should be interpreted. The relevant educational claim concerns appropriate reliance across unfamiliar situations, while a simulator can also improve scores through familiarity with its error templates. Distinct training and assessment banks, counterbalanced form allocation, multiple cases per failure mode and genuinely different system interfaces are therefore central validity conditions. Transfer across an underlying model, a display format and a clinical specialty should be described separately. Calling any change of vendor a universal test of transfer would overstate what has been learned. Independent reasoning without AI should also be reassessed, because competent supervision of assistance may coexist with declining unaided performance.

The acceptance–override example exposes a second vulnerability. A curriculum that repeatedly rewards rejection can increase AOR while weakening discrimination between trustworthy and untrustworthy advice. Balanced assessment makes that trade-off visible, and deployment-weighted analysis shows why its importance changes with the frequency of erroneous advice. The appropriate response is not to choose a favourable examination mixture. Investigators should publish the two conditional rates, report consequences of wrong actions and examine several externally justified prevalence and severity scenarios. Escalation deserves similar scrutiny: indiscriminate referral may appear safe in a vignette while imposing delay and workload on a real clinical service.

The resource audit also changes the procurement question. Educational effectiveness, clinical-content fidelity, availability and privacy controls must first meet a prespecified standard. Among acceptable configurations, institutions can compare cost, energy and faculty workload without assuming that a larger model improves learning or that local hosting automatically reduces environmental burden. Static, panel-approved advice can isolate pedagogical mechanisms during an initial trial and may require little inference. A later live-model evaluation would answer a different implementation question, with output variability, model updates and provider outages included. Economic reporting should follow an explicit perspective and horizon, consistent with CHEERS principles, rather than presenting token charges as a complete business case [32].

For Oman and comparable teaching settings, the next empirical step is a multicentre pilot involving educators, trainees and patient representatives. The pilot should test case comprehensibility, bilingual presentation where needed, assessment burden, assessor agreement and the feasibility of delayed follow-up. Reference standards must reflect locally available investigations, medicines and escalation pathways. Junior learners may need different debriefing and consent

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

arrangements from experienced doctors. Outcomes should be analysed for unequal access to preparation, digital resources and language support, without treating small subgroup estimates as stable evidence. These are prospective implementation requirements; the present study supplies no Omani learner data.

Several limitations bound the findings. The Monte Carlo model assumes homogeneous items conditional on learner ability and makes attrition independent of performance. It does not capture case-by-system interactions, teaching contamination, adaptive-policy instability or changes in clinical experience over six months. Repeated unfamiliar-system assessment can itself produce learning, so maintenance on that system is not equivalent to delayed first exposure. The t-interval procedure is evaluated only under the stated generator and fixed bank. Severity weights, acceptable acceptance-rate loss and meaningful educational improvement remain unresolved. The budget audit corrects arithmetic within assumed inputs, while genuine energy superiority requires matched measurement. A prospective analysis should address missing outcomes through a justified model and sensitivity analyses, retaining randomized participants in the target estimand rather than defining intention-to-treat as complete-case analysis [33].

Finally, reporting frameworks support transparency without conferring clinical validity. DECIDE-AI concerns early live clinical evaluations, so only its relevant planning principles inform this simulation study [34]. The present evidence supports the crossed assessment and the need for conditional safety outcomes. It does not support claiming an effective adaptive tutor, a validated clinical competency instrument or a greener deployment architecture. Those claims require different empirical observations.

6. Conclusion

Durable supervision of clinical AI cannot be evaluated adequately when assessment time and AI system change together. This methodological study formalized the resulting confounding and tested a crossed assessment in reproducible computational experiments. Under the specified assumptions, within-system comparisons recovered the intended retention contrast, whereas diagonal comparisons increasingly misrepresented retention as transfer penalties increased. These findings establish properties of an evaluation design, rather than the effectiveness of adaptive instruction.

Two further requirements follow. Appropriate acceptance must accompany appropriate override, because improved rejection of unsafe advice can coexist with poorer use of safe advice. Resource claims require consistent accounting boundaries, corrected arithmetic and measured workload-specific energy; illustrative API prices and assumed token intensities cannot establish environmental superiority.

A prospective multicentre study should now combine blinded adjudication, matched safe and unsafe cases, counterbalanced system exposure, delayed assessment and transparent adaptive-policy logs. Trial size should be justified using learner and case variability, attrition and clinically meaningful effects. Until those data exist, adaptive fallible-AI simulation remains a testable educational strategy. The contribution here is a clearer and reproducible way to determine whether it develops judgement that persists, transfers and merits its resource demands.

Ethics and data availability: This computational study involved no human participants or patient records. All numerical assessment data were generated under stated assumptions. Analysis code, parameters and outputs are provided in the accompanying reproducibility package. A future human study requires prospective ethics review, informed consent and registration.

References

1. E. Goh, R. Gallo, J. Hom, et al., “Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial,” JAMA Network Open, vol. 7, no. 10, Art. no. e2440969, 2024, doi: 10.1001/jamanetworkopen.2024.40969.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

2. E. Goh, R. J. Gallo, E. Strong, et al., "GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial," Nature Medicine, vol. 31, no. 4, pp. 1233-1238, 2025, doi: 10.1038/s41591-024-03456-y.
3. I. A. Qazi, A. Ali, A. U. Khawaja, et al., "Large language model diagnostic assistance for physicians in a lower-middle-income country: a randomized controlled trial," Nature Health, vol. 1, pp. 198-205, 2026, doi: 10.1038/s44360-025-00007-8.
4. I. A. Qazi, A. Ali, A. U. Khawaja, et al., "Automation Bias in Large Language Model-Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy—A Randomized Clinical Trial," NEJM AI, vol. 3, no. 5, Art. no. AIoa2501001, 2026, doi: 10.1056/AIoa2501001.
5. S. Gaube, H. Suresh, M. Raue, et al., "Do as AI say: susceptibility in deployment of clinical decision-aids," npj Digital Medicine, vol. 4, no. 1, Art. no. 31, 2021, doi: 10.1038/s41746-021-00385-9.
6. A. Kiani, B. Uyumazturk, P. Rajpurkar, et al., "Impact of a deep learning assistant on the histopathologic classification of liver cancer," npj Digital Medicine, vol. 3, Art. no. 23, 2020, doi: 10.1038/s41746-020-0232-8.
7. P. Tschandl, C. Rinner, Z. Apalla, et al., "Human-computer collaboration for skin cancer recognition," Nature Medicine, vol. 26, pp. 1229-1234, 2020, doi: 10.1038/s41591-020-0942-0.
8. S. Jabbour, D. Fouhey, S. Shepard, et al., "Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study," JAMA, vol. 330, no. 23, pp. 2275-2284, 2023, doi: 10.1001/jama.2023.22295.
9. Y. S. Kıyak, Ö. Coşkun, I. İ. Budakoğlu, "'ChatGPT can make mistakes' warnings fail: A randomized controlled trial," Medical Education, vol. 60, no. 2, pp. 138-142, 2026, doi: 10.1111/medu.70056.
10. D. Teng, L. Tan, Q. Cao, et al., "Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students," npj Digital Medicine, vol. 9, Art. no. 356, 2026, doi: 10.1038/s41746-026-02547-z.
11. D. B. Wayne, J. Butter, V. J. Siddall, et al., "Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice," Journal of General Internal Medicine, vol. 21, no. 3, pp. 251-256, 2006, doi: 10.1111/j.1525-1497.2006.00341.x.
12. S. Mamede, T. van Gog, K. van den Berge, et al., "Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents," JAMA, vol. 304, no. 11, pp. 1198-1203, 2010, doi: 10.1001/jama.2010.1276.
13. D. P. Larsen, A. C. Butler, H. L. Roediger III, "Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial," Medical Education, vol. 43, no. 12, pp. 1174-1181, 2009, doi: 10.1111/j.1365-2923.2009.03518.x.
14. K. Fraser, A. Peets, I. Walker, et al., "The effect of simulator training on clinical skills acquisition, retention and transfer," Medical Education, vol. 43, no. 8, pp. 784-789, 2009, doi: 10.1111/j.1365-2923.2009.03412.x.
15. Z. Buçinca, M. B. Malaya, K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW1, pp. 188:1-188:21, 2021, doi: 10.1145/3449287.
16. T. Thesen and S. H. Park, "A generative AI teaching assistant for personalized learning in medical education," npj Digital Medicine, vol. 8, Art. no. 627, 2025, doi: 10.1038/s41746-025-02022-1.
17. S. B. Issenberg, W. C. McGaghie, E. R. Petrusa, D. L. Gordon, and R. J. Scalese, "Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review," Medical Teacher, vol. 27, no. 1, pp. 10–28, 2005, doi: 10.1080/01421590500046924.
18. D. A. Cook, R. Hatala, R. Brydges, et al., "Technology-enhanced simulation for health professions education: a systematic review and meta-analysis," JAMA, vol. 306, no. 9, pp. 978–988, 2011, doi: 10.1001/jama.2011.1234.
19. W. C. McGaghie, S. B. Issenberg, E. R. Petrusa, and R. J. Scalese, "A critical review of simulation-based medical education research: 2003–2009," Medical Education, vol. 44, no. 1, pp. 50–63, 2010, doi: 10.1111/j.1365-2923.2009.03547.x.
20. W. C. McGaghie, S. B. Issenberg, E. R. Cohen, J. H. Barsuk, and D. B. Wayne, "Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 5 November 2026 https://doi.org/10.20944/preprints202611.0587.v1

analytic comparative review of the evidence," Academic Medicine, vol. 86, no. 6, pp. 706–711, 2011, doi: 10.1097/ACM.0b013e318217e119.

21. T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," Statistics in Medicine, vol. 38, no. 11, pp. 2074–2102, 2019, doi: 10.1002/sim.8086.
22. C. M. Judd, J. Westfall, and D. A. Kenny, "Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem," Journal of Personality and Social Psychology, vol. 103, no. 1, pp. 54–69, 2012, doi: 10.1037/a0028347.
23. J. Westfall, D. A. Kenny, and C. M. Judd, "Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli," Journal of Experimental Psychology: General, vol. 143, no. 5, pp. 2020–2045, 2014, doi: 10.1037/xge0000014.
24. G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
25. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Medicine, vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.
26. S. E. Embretson and S. P. Reise, Item Response Theory for Psychologists. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2000.
27. L. Lannelongue, J. Grealey, and M. Inouye, "Green Algorithms: Quantifying the Carbon Footprint of Computation," Advanced Science, vol. 8, no. 12, art. 2100707, 2021, doi: 10.1002/advs.202100707.
28. P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning," Journal of Machine Learning Research, vol. 21, no. 248, pp. 1–43, 2020.
29. A. S. Luccioni, Y. Jernite, and E. Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?," in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 85–99, 2024, doi: 10.1145/3630106.3658542.
30. S. Cruz Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," Nature Medicine, vol. 26, no. 9, pp. 1351–1363, 2020, doi: 10.1038/s41591-020-1037-7.
31. X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nature Medicine, vol. 26, no. 9, pp. 1364–1374, 2020, doi: 10.1038/s41591-020-1034-x.
32. D. Husereau, M. Drummond, F. Augustovski, et al., "Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations," BMJ, vol. 376, art. e067975, 2022, doi: 10.1136/bmj-2021-067975.
33. I. R. White, N. J. Horton, J. Carpenter, and S. J. Pocock, "Strategy for intention to treat analysis in randomised trials with missing outcome data," BMJ, vol. 342, art. d40, 2011, doi: 10.1136/bmj.d40.
34. B. Vasey, M. Nagendran, B. Campbell, et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," Nature Medicine, vol. 28, no. 5, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.