

Clinical AI Trust Half-Life: Time-Decaying Prediction Assurance Under Data and Fairness Drift

Nithesh Naik^{ID*} ¹, Lee Desmond Chong En¹, and Harikrishnan A/L Ramiah²

¹Jesselton University College, Kota Kinabalu, Sabah, Malaysia

²Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia

Abstract

Clinical artificial intelligence (AI) models are commonly validated at a fixed point, although the conditions that support their predictions can change after deployment. This paper introduces Clinical AI Trust Half-Life, a theoretical measure of the time required for overall prediction assurance to decline to one-half of its initially validated level. Assurance is represented as a normalized composite of predictive performance, calibration, distribution stability, and subgroup fairness, while workflow change and monitoring delay influence the rate and practical consequences of decay. A reproducible simulation trains logistic regression on synthetic baseline data for two patient groups and then evaluates the unchanged model across thirteen periods with gradual covariate drift, increasing outcome prevalence, changing predictor–outcome relationships, and faster degradation in one subgroup. In the executed simulation, assurance decreased from 0.858 at baseline to 0.334 at period 12. The empirical half-life occurred at period 9, when assurance first fell below 0.429; an exponential approximation yielded 8.78 periods. The simulation demonstrates that aggregate discrimination can remain superficially usable while calibration and subgroup sensitivity deteriorate substantially. Trust half-life is proposed as a governance summary, not a substitute for component-level surveillance or clinical validation. It can support risk-adjusted monitoring intervals, early-warning thresholds, subgroup audits, revalidation, recalibration, and explicit assignment of maintenance responsibilities.

Keywords: Clinical Artificial Intelligence; Data Drift; Model Calibration; Fairness Drift; Post-Deployment Monitoring; Prediction Assurance; Trust Half-Life

1. Introduction

Clinical artificial intelligence (AI) can support diagnosis, triage, prognosis, and treatment decisions, but its clinical value depends on more than technical operation. A model may continue returning probabilities without software failure while the evidential basis for trusting those probabilities weakens. Translation studies have therefore emphasized representative evaluation, generalizability, workflow integration, and post-market surveillance rather than assuming that development-set accuracy guarantees sustained clinical benefit [1]. The central problem is temporal: validation describes performance under a bounded sample and period, whereas deployment occurs in a changing health system.

*Corresponding Author: Nithesh Naik (nitheshnayak@gmail.com)

Received: 29 June 2026; Revised: 01 August 2026; Accepted: 11 September 2026; Published: 30 September 2026

© 2026 The Author(s).

This is an open access article licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0).

Healthcare data are non-stationary. Patient mix, disease prevalence, coding systems, laboratory assays, imaging equipment, referral patterns, and treatment policies can change the joint distribution linking predictors and outcomes. Dataset-shift theory distinguishes changes in input distributions, outcome prevalence, and predictor–outcome relationships [2]. Longitudinal electronic-health-record analyses have documented deterioration after record-keeping changes and over calendar time [3], and a systematic review found that calibration deterioration was reported more often than discrimination deterioration in studies addressing temporal shift [4]. Experiments using intensive-care data further showed that domain-generalization and adaptation techniques did not uniformly preserve performance across later periods [5].

The deployment problem is also partially observable. Outcomes may mature slowly, labels may be costly, and aggregate metrics may conceal subgroup-specific harm. Reviews of medical data drift describe time-varying patient populations, acquisition processes, and clinical context as recurring causes of performance decay [6]. Model-maintenance scholarship consequently calls for explicit policies covering surveillance, updating, and ownership [7]. Empirical imaging studies show that performance monitoring alone may miss clinically obvious distribution changes [8], while postmarket simulations demonstrate that shift-detection sensitivity depends on shift type and sample size [9]. More recent clinical work has combined label-agnostic detection with transfer or continual learning to identify and remediate harmful shifts [10].

Predictive discrimination is only one dimension of assurance. A model can retain a useful area under the receiver operating characteristic curve (AUROC) while its probabilities become systematically too high or too low. Calibration is therefore essential when predicted risks guide clinical action [11], and modern machine-learning models often require post-processing or recalibration to align confidence with observed frequencies [12]. Fairness introduces another time-varying dimension. Healthcare fairness depends on the clinical objective, error distribution, and resource consequences [13]; underdiagnosis can disproportionately affect underserved groups [14]; and fairness criteria must be interpreted within data-generation and workflow contexts [15]. Importantly, longitudinal evidence now shows that fairness drift can arise even when population-level model maintenance is performed [16]. External validation remains indispensable, but it does not itself guarantee future transportability after deployment conditions change [17].

This paper develops Clinical AI Trust Half-Life as a time-dependent assurance measure. Its contributions are fourfold. First, it defines an overall assurance score combining normalized performance, calibration, distribution stability, and subgroup fairness. Second, it defines empirical and exponential trust half-life estimators and clarifies how workflow change and monitoring delay affect practical risk. Third, it demonstrates the measure in an executable synthetic simulation with unequal subgroup degradation. Fourth, it translates the measure into governance actions while emphasizing that the framework is theoretical and requires prospective validation before clinical use.

2. Related Work

Clinical AI evaluation guidance increasingly treats deployment as a sociotechnical intervention rather than a static prediction benchmark. DECIDE-AI recommends transparent reporting of early live clinical evaluation, including human factors, workflow integration, and safety observations [18]. CONSORT-AI similarly requires description of intended use, input and output handling, human–AI interaction, and error analysis in clinical trials [19]. These standards strengthen prospective evidence but do not provide a compact time-to-degradation measure for routine operational surveillance.

Post-deployment monitoring research has proposed recurring evaluation of discrimination, calibration, threshold metrics, input distributions, and data quality. However, feedback loops can bias observed labels: successful intervention after a high-risk alert may make a correct prediction appear wrong. Simulation research has therefore proposed adherence- and sampling-weighted monitoring to estimate counterfactual performance more faithfully [20]. Subgroup robustness studies also warn that optimizing average or worst-case objectives does not reliably improve every patient subgroup and may worsen calibration [21]. Global reviews identify underrepresentation of countries, specialties, and populations as upstream sources of poor transportability and inequity [22]. Explainable drift monitoring has additionally been used to identify changes in feature importance during the COVID-19 period [23].

Ordinary monitoring answers whether a particular metric has changed. Trust half-life asks a different governance question: how long does the combined assurance supporting current predictions remain above a prespecified fraction of its validated baseline? The measure integrates several surveillance streams into a temporal summary, permits direct empirical estimation, and supports application-specific action horizons. It does not replace the underlying metrics; rather, it provides an interpretable clock whose decline can trigger deeper investigation, warning, recalibration, retraining, restricted use, or suspension.

3. Clinical AI Trust Half-Life Model

Let $A(t)$ denote prediction assurance at deployment time t , with all components scaled to $[0,1]$ and larger values indicating stronger assurance. A simple decay representation is

$$A(t) = A_0 \exp[-\lambda(t)t], \quad 0 < A_0 \leq 1 \quad (1)$$

where A_0 is assurance under the initial validated deployment condition and $\lambda(t)$ is a potentially time-varying decay rate. The rate summarizes how quickly evidence supporting the model becomes stale because of data change, miscalibration, performance loss, fairness drift, workflow change, or delayed corrective action. Equation (1) is an approximation rather than a biological law; assurance may plateau, recover after maintenance, or decline abruptly after a protocol or information-system change.

For operational use, assurance is decomposed into four monitored dimensions:

$$A(t) = w_p P(t) + w_c C(t) + w_d D(t) + w_f F(t) \quad (2)$$

$$w_p + w_c + w_d + w_f = 1, \quad w_j \geq 0. \quad (3)$$

Here $P(t)$ is normalized predictive performance, $C(t)$ is calibration assurance, $D(t)$ is distribution-stability assurance, and $F(t)$ is fairness assurance. Weights should be prespecified by a multidisciplinary governance group and justified by intended use, error severity, label availability, and patient impact. For example, a screening model may emphasize sensitivity and subgroup false-negative disparities, whereas a probability-based treatment model may place greater weight on calibration. To avoid compensation, the weighted score should be accompanied by hard component floors; high distribution stability must not conceal unacceptable subgroup performance.

Several normalizations are possible. For binary discrimination measured by AUROC, a chance-corrected transformation is $P(t) = \text{clip}\{2[\text{AUROC}(t) - 0.5], 0, 1\}$. Calibration assurance can be defined as $C(t) = \text{clip}[1 - \text{ECE}(t)/\tau_c, 0, 1]$, where ECE is expected calibration error and τ_c is a prespecified unacceptable-error scale. Distribution stability may use standardized mean shifts, classifier-based two-sample tests, maximum mean discrepancy, or domain-specific data-quality indicators. In the simulation, $D(t) = \exp[-s(t)/\tau_d]$, where $s(t)$ is the average absolute standardized change in feature means. These choices are illustrative; clinical deployments require metrics aligned with their data modality and failure modes.

A subgroup-sensitive equal-opportunity assurance can be defined as

$$F(t) = 1 - |\text{TPR}^{g_1}(t) - \text{TPR}^{g_2}(t)| \quad (4)$$

where TPR is the true-positive rate at the operational threshold. The formulation equals one when group sensitivities match and approaches zero as the gap grows. More than two groups can be handled by the maximum pairwise gap, the difference between worst and best group performance, or a harm-weighted summary. Fairness assurance should also be reported with subgroup sample sizes and uncertainty because sparse events can make apparent drift unstable.

The empirical trust half-life is

$$T_{1/2} = \inf\{t : A(t) \leq A(0)/2\}. \quad (5)$$

With discrete monitoring windows, $T_{1/2}$ is the first observed period below half of baseline; interpolation may be used when justified. If decay is approximately exponential with a constant rate λ , then

$$T_{1/2} = \ln(2)/\lambda. \quad (6)$$

The empirical estimate need not follow perfect exponential decay. A fitted λ is useful for planning, but the observed threshold crossing remains the primary operational estimate. A more conservative application can replace the 50% criterion with an action threshold $\theta > A(0)/2$, particularly for high-risk uses. Monitoring and update delays do not merely postpone discovery; they lengthen the interval in which degraded predictions continue to affect care. They can be represented through a response lag $L(t)$ or incorporated into an effective decay rate $\lambda_{\text{eff}}(t) = \lambda_{\text{technical}}(t) + \lambda_{\text{latency}}(t)$. Workflow and population changes may likewise modify $\lambda(t)$, component weights, or safety thresholds rather than appearing as a fifth interchangeable score.

The baseline $A(0)$ should be estimated from a clearly versioned validation window that reflects the intended clinical setting, not automatically set to one. If baseline performance is already marginal, a relative half-life can otherwise create false reassurance. Governance records should therefore retain both the absolute component values and the relative ratio $A(t)/A(0)$. A deployment should also define minimum event counts, confidence-interval rules, and missing-data procedures before monitoring starts. When outcome labels are delayed, provisional distribution and workflow signals may update assurance, but the score should visibly distinguish observed components from imputed or stale components.

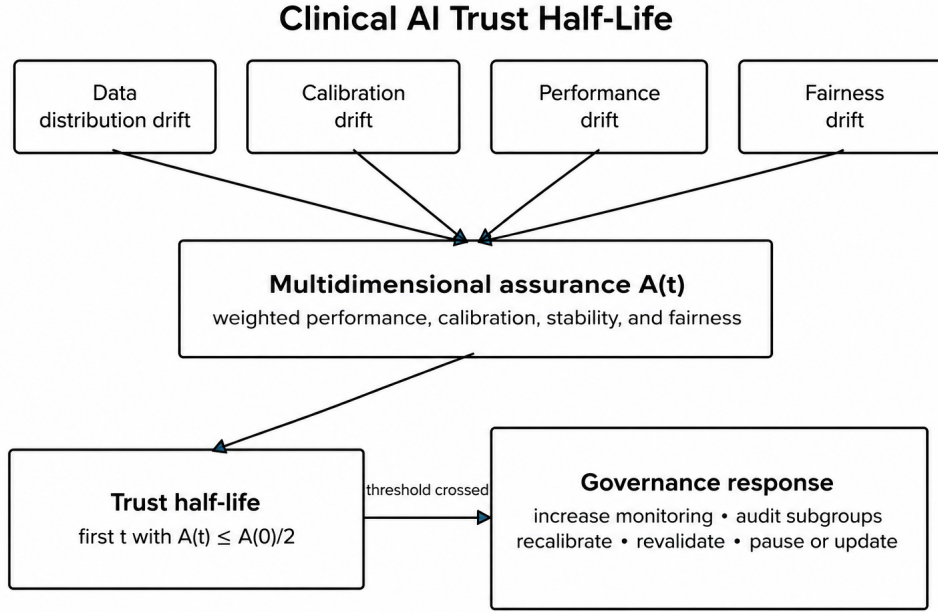


Figure 1: Clinical AI Trust Half-Life framework. Data, calibration, performance, and fairness drift reduce multi-dimensional prediction assurance. The first crossing of 50% of baseline defines the empirical half-life and prompts risk-adjusted governance action.

Table 1: Components of the Clinical AI Trust Half-Life framework.

Component	Indicator	Interpretation	Potential monitoring method
Predictive performance	AUROC, balanced accuracy, sensitivity	Ability to rank or classify current patients correctly	Outcome-linked rolling evaluation with confidence intervals
Calibration	ECE, Brier score, intercept/slope	Agreement between predicted risk and observed frequency	Reliability curves, delayed-label audits, recalibration tests
Distribution stability	Standardized shift, MMD, data quality	Similarity of current inputs to the validated domain	Feature dashboards, two-sample tests, missingness and device checks
Fairness	TPR gap, calibration gap, worst-group metric	Whether errors or risk estimates remain comparable across groups	Disaggregated audit with intersectional sample-size rules
Workflow/population context	Protocol, referral, prevalence, user changes	Change in how data are produced or predictions are acted upon	Change-control logs, clinical review, prevalence surveillance
Monitoring/update latency	Label delay, audit interval, remediation time	Duration of exposure before degradation is detected and corrected	Service-level targets, escalation timers, update records

4. Simulation Study

A reproducible simulation was executed to illustrate estimation of the proposed quantities; it is not evidence of clinical effectiveness. Four continuous features and a binary patient-group indicator were generated. A logistic-regression pipeline with standardization was trained on 8,000 baseline observations. The unchanged model was then evaluated on 4,000 new observations in each of thirteen periods. Common covariate drift progressively shifted three features, outcome prevalence increased through a rising intercept, and predictor–outcome coefficients changed over time. Group 1 experienced additional opposing feature shifts, time-varying interactions, and increasing label noise, producing faster subgroup degradation.

AUROC measured discrimination, while Brier score and ten-bin expected calibration error (ECE) measured probability quality. At the fixed 0.5 threshold, the true-positive rate (TPR) was calculated separately for each group. The component normalizations followed Section 3 with $\tau_c=0.25$ and $\tau_d=1.25$. Weights were $w_p=0.35$, $w_c=0.20$, $w_d=0.20$, and $w_f=0.25$. These weights are design choices for demonstration, not recommended clinical defaults.

The simulation deliberately combines several changes rather than isolating one mechanism. Consequently, the component trajectories should be interpreted as a worked monitoring scenario, not as causal estimates of how much each drift source contributes. Brier score is retained in the generated CSV as a proper scoring rule, while $C(t)$ uses ECE for transparent normalization. In a real audit, bootstrap or analytic intervals would be used to quantify uncertainty, and thresholds would be evaluated against clinically meaningful error costs rather than selected only for numerical convenience.

Table 2: Executed synthetic simulation results over time.

t	AUROC	P(t)	C(t)	D(t)	F(t)	A(t)
0	0.850	0.700	0.905	0.986	0.938	0.858
1	0.840	0.680	0.878	0.986	0.964	0.852
2	0.831	0.662	0.755	0.959	0.872	0.792
3	0.807	0.614	0.679	0.951	0.770	0.733
4	0.801	0.603	0.551	0.932	0.673	0.676
5	0.791	0.583	0.431	0.926	0.538	0.610
6	0.790	0.580	0.254	0.920	0.420	0.543
7	0.785	0.570	0.215	0.909	0.347	0.511
8	0.754	0.509	0.042	0.899	0.266	0.433
9	0.741	0.482	0.000	0.882	0.226	0.402
10	0.726	0.453	0.000	0.876	0.169	0.376
11	0.711	0.422	0.000	0.858	0.132	0.352
12	0.703	0.407	0.000	0.852	0.084	0.334

At baseline, AUROC was 0.850, ECE was 0.024, the fairness assurance was 0.938, and total assurance was 0.858. By period 12, AUROC had declined to 0.703 and ECE had increased to 0.349. The TPR for group 0 rose from 0.717 to 0.951 because prevalence and the fixed threshold increasingly favored positive classification, while the group 1 TPR fell from 0.779 to 0.036 under the subgroup-specific shift. Consequently, $F(t)$ declined to 0.084. This illustrates why a stable or improving metric in one group cannot be treated as overall safety.

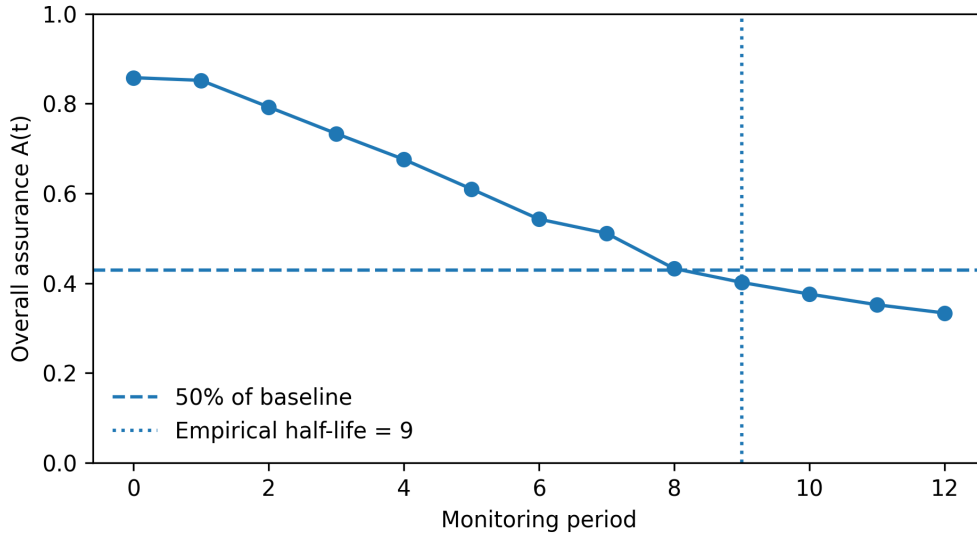


Figure 2: Overall assurance over thirteen monitoring periods. The dashed horizontal line is 50% of baseline assurance (0.429); the vertical marker identifies the first crossing at period 9.

The 50% threshold was 0.429. Assurance remained slightly above it at period 8 (0.433) and fell below it at period 9 (0.402), yielding an empirical trust half-life of nine monitoring periods. Fitting $\log[A(t)/A(0)]$ through the origin

produced $\lambda=0.079$ per period and an exponential half-life of 8.78 periods. The close values are specific to this simulated decay path and should not be generalized. Figure 2 shows the overall crossing; Figure 3 shows the widening subgroup TPR divergence.

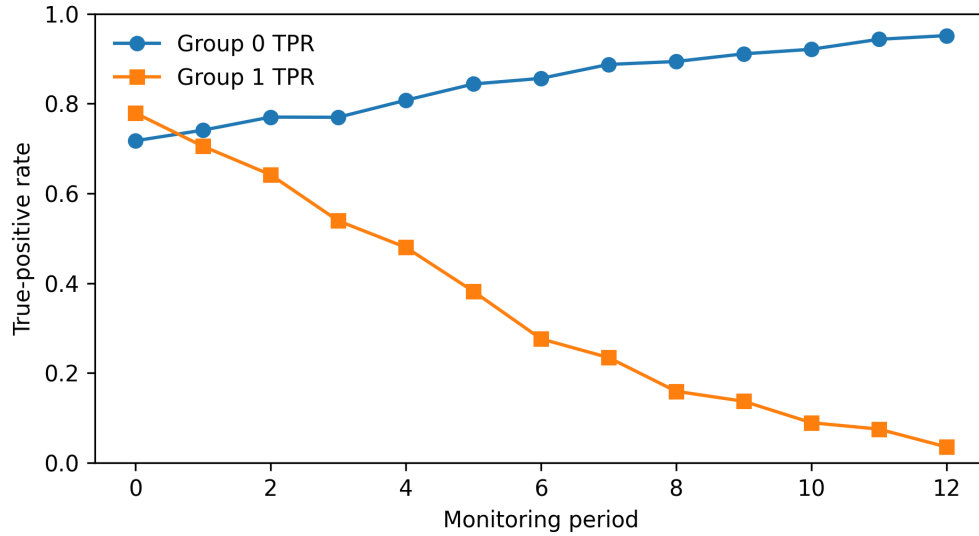


Figure 3: Subgroup true-positive-rate trajectories. Differential degradation becomes pronounced even while aggregate AUROC remains above 0.70 at the end of the simulation.

5. Clinical and Governance Implications

Trust half-life can convert surveillance evidence into a maintenance schedule. A model with an estimated half-life of twelve months should not be reviewed only annually: a rational audit interval would be a conservative fraction of the expected half-life, shortened when labels are delayed or drift accelerates. Distribution and data-quality indicators can provide leading warnings when outcomes are unavailable, while performance, calibration, and fairness should be updated as labels mature. Control charts or sequential tests can distinguish sustained degradation from random fluctuation, and uncertainty intervals should accompany every component.

Revalidation should be action-linked. Mild calibration decay with preserved ranking may justify intercept or slope recalibration. A change in predictor–outcome relationships, severe distribution shift, or subgroup-specific failure may require retraining, feature redesign, threshold revision, restricted use, or model withdrawal. A governance matrix should specify warning and critical thresholds for $A(t)$ and for each component, the accountable clinical and technical owners, required response times, validation data, approval authority, rollback procedures, and communication to users. Model documentation should record baseline definitions, weights, monitored subgroups, metric versions, audit windows, alerts, investigations, updates, and post-update performance.

Fairness auditing must be recurring rather than confined to development. At minimum, clinically relevant error rates and calibration should be disaggregated across prespecified demographic, geographic, and disease-severity groups. Intersectional analysis is desirable where sample sizes permit. If an update restores average performance but worsens a disadvantaged group, the aggregate assurance score may be unacceptable even before the 50% threshold. Hard fairness floors and worst-group reporting are therefore essential companions to the composite score.

Risk determines how the half-life should be used. In a high-risk application—such as escalation for possible sepsis, acute imaging triage, or withholding an intervention—the organization should adopt shorter monitoring windows, earlier action thresholds, stricter component floors, and rapid human review. Waiting until assurance halves may be unsafe; a threshold such as 80% of baseline may be more appropriate. A low-risk administrative prioritization tool may tolerate longer windows and a lower response urgency, provided that downstream access or resource allocation effects are still audited. Thus, trust half-life is not a universal expiry date but an application-specific measure embedded in clinical risk management.

Monitoring delay is itself governable. Label latency, infrequent audits, unassigned ownership, and slow vendor response create an unprotected interval during which a technically functioning model can continue to influence decisions after assurance has deteriorated. Institutions should therefore define service-level objectives for data availability, alert review, clinical adjudication, corrective validation, and deployment of approved updates. Human override rates and

clinician-reported anomalies should be monitored as complementary signals, because workflow changes or automation bias may alter realized safety even when statistical performance appears stable.

6. Limitations and Future Research

The framework has several limitations. First, its weights, transformations, and thresholds are normative and application dependent. The additive form permits compensation among components, which is why hard floors and component dashboards are required. Second, the simulation uses a single binary logistic model, two groups, fixed thresholds, immediate labels, and simple mean-shift detection. Real deployments involve censoring, missingness, recurrent outcomes, multimodal data, rare subgroups, changing interventions, and feedback loops. Third, fairness cannot be reduced to one metric; equal-opportunity difference may be inappropriate when treatment benefit, prevalence, or harm differs across groups. Fourth, estimated half-life is sensitive to audit window length, sampling variability, and missing or biased labels. It should therefore be reported with uncertainty and sensitivity analyses. Finally, the measure has not been prospectively validated and should not be interpreted as regulatory certification or proof of clinical safety.

Future work should evaluate the framework in multi-site prospective deployments and compare empirical half-life with patient-safety events, clinician overrides, and maintenance decisions. Bayesian or state-space models could estimate latent assurance with uncertainty and accommodate abrupt changes, recovery after updates, and irregular monitoring. Change-point detection could complement exponential fits. Research is also needed on optimizing weights under explicit clinical utility and equity constraints, defining statistically valid subgroup alarms, incorporating causal workflow indicators, handling intervention-induced label modification, and determining monitoring frequency through cost-sensitive decision analysis. A useful extension would estimate separate component half-lives and the worst-group half-life, preventing a reassuring aggregate value from masking rapid local decay.

7. Conclusion

Clinical AI can remain computationally operational while the justification for trusting its predictions decays. Clinical AI Trust Half-Life formalizes this distinction by defining the time at which multidimensional prediction assurance falls to one-half of its validated baseline. The proposed score combines performance, calibration, distribution stability, and subgroup fairness, while recognizing that workflow change and monitoring latency shape real-world risk. In the executed synthetic study, the empirical half-life was nine periods and the exponential approximation was 8.78 periods; these results illustrate calculation only and do not establish clinical validity. Used with component-level guardrails, uncertainty estimates, and risk-specific thresholds, trust half-life could support transparent monitoring schedules, escalation rules, revalidation, recalibration, subgroup auditing, and accountability. Prospective clinical evaluation is required before the measure can guide deployment decisions.

Author Contributions

Nithesh Naik (N.N): Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing – original draft, Project administration; Lee Desmond Cherng En (L.D.C.E): Methodology, Validation, Supervision, Writing – review & editing; Harikrishnan A/L Ramiah (H.R): Conceptualization, Methodology, Formal analysis, Validation, Visualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author, N.N., upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

Claude was used selectively during the research process to assist with experimental execution and computational analysis. The authors independently reviewed and verified the results and remain fully responsible for the accuracy, originality, and integrity of the work.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17:195. doi:10.1186/s12916-019-1426-2.
- [2] Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics.* 2020;21(2):345-352. doi:10.1093/biostatistics/kxz041.
- [3] Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *Proc Mach Learn Res.* 2019;106:381-405.
- [4] Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform.* 2021;12(4):808-815. doi:10.1055/s-0041-1735184.
- [5] Guo LL, Pfohl SR, Fries J, Johnson AEW, Posada J, Aftandilian C, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci Rep.* 2022;12:2726. doi:10.1038/s41598-022-06484-1.
- [6] Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol.* 2023;96(1150):20220878. doi:10.1259/bjr.20220878.
- [7] Davis SE, Walsh CG, Matheny ME. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. *Front Digit Health.* 2022;4:958284. doi:10.3389/fdgth.2022.958284.
- [8] Kore A, Abbasi Babil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat Commun.* 2024;15:1887. doi:10.1038/s41467-024-46142-w.
- [9] Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. *NPJ Digit Med.* 2024;7:120. doi:10.1038/s41746-024-01085-w.
- [10] Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open.* 2025;8(6):e2513685. doi:10.1001/jamanetworkopen.2025.13685.
- [11] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; STRATOS Topic Group. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230. doi:10.1186/s12916-019-1466-7.
- [12] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. *Proc Mach Learn Res.* 2017;70:1321-1330.
- [13] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* 2018;169(12):866-872. doi:10.7326/M18-1990.
- [14] Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0.
- [15] Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithm fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng.* 2023;7(6):719-742. doi:10.1038/s41551-023-01056-8.

- [16] Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *J Am Med Inform Assoc.* 2025;32(5):845-854. doi:10.1093/jamia/ocaf039.
- [17] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J.* 2021;14(1):49-58. doi:10.1093/ckj/sfaa188.
- [18] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al.; DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ.* 2022;377:e070904. doi:10.1136/bmj-2022-070904.
- [19] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health.* 2020;2(10):e537-e548. doi:10.1016/S2589-7500(20)30218-1.
- [20] Kim GYE, Corbin CK, Grolleau F, Baiocchi M, Chen JH. Monitoring strategies for continuous evaluation of deployed clinical prediction models. *J Biomed Inform.* 2025;168:104854. doi:10.1016/j.jbi.2025.104854.
- [21] Pfohl SR, Zhang H, Xu Y, Foryciarz A, Ghassemi M, Shah NH. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Sci Rep.* 2022;12:3254. doi:10.1038/s41598-022-07167-7.
- [22] Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al.; MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities-a global review. *PLOS Digit Health.* 2022;1(3):e0000022. doi:10.1371/journal.pdig.0000022.
- [23] Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TWV, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep.* 2021;11:23017. doi:10.1038/s41598-021-02481-y.