Runtime Reliability Labeling for Safe Deployment of Oncology AI Under Distribution Shift

Per Magnus Swedenborg[1],

[1] DNAI Biotech Correspondence: per.swedenborg@dnai.bio | ORCID: 0009-0008-2750-9735

Abstract

Background: Clinical AI models frequently fail to generalize across institutions due to distribution shifts in patient demographics, assay protocols, and missing modalities. While techniques like Distributionally Robust Optimization (DRO) improve population-level stability, they do not provide the patient-level reliability assessments required for safe clinical decision-making. Clinicians need a traffic light systemknowing when to trust a prediction and when to abstainrather than a single aggregate metric.

Methods: We present a deployment framework integrated into the DNAI cancer digital twin platform. The system enforces reliability via **Runtime Reliability Labeling**, computing an **Information Sufficiency Score (ISS)** that aggregates prediction confidence (boundary proximity), latent-space Mahalanobis distance, manifold density, and data completeness. Crucially, to prevent artifacts from cross-platform encoding, OOD thresholds are derived from a **Ridge-projected reference manifold** of the internal validation set. We evaluated the framework on 5,070 external patients across CPTAC (pan-cancer), CGGA (glioma), and SCAN-B (breast) under a strict frozen protocol.

Results: The framework stratified patients into GREEN (reliable), YELLOW (caution), and RED (abstain) tiers. In the CPTAC cohort ($N = 1,031$), the GREEN tier (22.2% coverage, $N = 229$) achieved a C-index of **0.744** [95% CI: 0.6920.793], outperforming the unstratified baseline (0.686). GREEN patients survived significantly longer than RED (+310 days, $p = 8.6$ Œ 10^{7}). In CGGA ($N = 970$), the GREEN tier (61.4% coverage, $N = 596$) achieved a C-index of **0.675** [0.6430.706], with GREEN-vs-RED survival separation of +992 days ($p = 7.2$ Œ 10^{40}). In SCAN-B (RNA-only breast cancer), the system correctly enforced a **No-Green Veto** due to missing histology in a WSI-dependent cancer type, assigning 0% GREEN and 95% YELLOW, thereby preventing high-confidence errors.

Conclusion: Trust in clinical AI can be operationalized as a computable reliability label. By coupling selective prediction with rigorous OOD detection, this framework ensures models abstain when information is insufficient, prioritizing patient safety over broad coverage.

1. Introduction

The translation of machine learning models from *in silico* development to clinical practice is plagued by the deployment gap. Models that achieve state-of-the-art performance on curated benchmarks often degrade when exposed to the heterogeneity of real-world clinical data [1, 2]. In oncology, where treatment decisions rely on precise risk stratification, distribution shiftcaused by differences in scanners, staining protocols, or patient demographicsposes a critical safety risk.

Current approaches largely focus on **robust training**, employing techniques like Domain Adversarial Neural Networks (DANN) or Distributionally Robust Optimization (DRO) to force models to learn invariant features [3, 4]. While valuable, these methods optimize for *average* performance across a target population. They do not answer the specific question a clinician faces in the tumor board: *Can I trust the prediction for **this specific patient**?*

We propose that clinical AI requires a shift from robust models to **reliable deployment systems.** Reliability implies the ability to self-assess competence and abstain when evidence is insufficient or out-of-distribution (OOD) [7, 8]. This requires a Traffic Light protocol: **Green** (Standard of Care usage), **Yellow** (Human-in-the-loop review), and **Red** (Abstention).

In this work, we introduce a unified framework for **Runtime Reliability Labeling**. Built atop the DNAI cancer digital twin platform [5], this framework generates a dynamic reliability label for every patient. We introduce the **Information Sufficiency Score (ISS)**, a composite metric enabling granular abstention, and validate it on three external cohorts comprising over 5,000 patients. We demonstrate that reliability labeling improves clinical utility by filtering out patients for whom the models biological assumptions do not hold, yielding significant survival separation between triage tiers.

2. Related Work

2.1 Selective Prediction and Trust

Selective prediction allows models to abstain on low-confidence samples [8, 9]. While Conformal Prediction (CP) [6] offers distribution-free guarantees, standard CP methods for survival analysis often yield trivially wide intervals in high-censoring regimes [22]. Consequently, we adopt a Trust Score approach [21], utilizing distance to the training manifold and model uncertainty as proxies for reliability.

2.2 Regulatory Frameworks for AI Lifecycle

Recent regulatory guidance, such as the FDAs Predetermined Change Control Plan (PCCP) [25] and the EU AI Act [27], emphasizes the need for rigorous lifecycle management. The NIST AI Risk Management Framework (AI RMF) [28] specifically calls for measure and manage functions to handle system reliability. Our framework provides a technical implementation of these governance requirements, ensuring that the deployed software function matches the validated specification via cryptographic interlocks.

3. Methods

3.1 Host Model: DNAI Hypernet v3.2

The underlying predictive model is the **DNAI Hypernet v3.2 (Path A)**, a multi-modal cancer digital twin. * **Foundation:** A Variational Autoencoder (VAE v5.10) encodes RNA, DNA, CNV, and Methylation into a 328-dimensional biologically disentangled latent space (z_{full}). * **Inference:** The Hypernet conditions on z_{full} and Whole Slide Imaging (WSI) embeddings (UNI2-h) to predict tumor dynamics parameters (, ,) for a Neural ODE. * **Uncertainty:** While the model supports Monte Carlo (MC) Dropout, for runtime efficiency the ISS utilizes **calibrated boundary proximity** (C_{prox}) as a proxy for prediction confidence.

3.2 Addressing the Encoding Gap: The Reference Manifold

A critical challenge in external validation is the Encoding Gap. The VAE is trained on multi-omics (TCGA), but external cohorts (CPTAC, CGGA) often possess only RNA. To bridge this, we employ a **Ridge Projection** ($= 10$) to map standardized RNA expression to the 328-d VAE latent space ($R^2 = 0.638$).

Crucially, to ensure valid OOD detection, we do not compare Ridge-projected external patients against VAE-encoded training patients. Instead, we construct a **Reference Manifold** by projecting the *TCGA Validation Set* through the *same* Ridge regressor. The Mahalanobis distance covariance matrix ($_{ref}$) and mean ($_{ref}$) are computed on these Ridge-projected internal samples. This ensures that the OOD metric measures true biological shift, not the artifactual difference between VAE and Ridge encodings.

3.3 The TraceabilityEngine v1.2

The TraceabilityEngine wraps the host model, acting as a gatekeeper. It computes a **Reliability Label** for every inference request.

3.3.1 Information Sufficiency Score (ISS) The ISS is an additive scalar $[0, 1]$ aggregating four normalized reliability metrics.

$$\text{ISS} = w_1 C_{prox} + w_2 S_{ood} + w_3 C_{data} + w_4 N_{density}$$

Weights: $w = [0.35, 0.35, 0.20, 0.10]$, selected to balance prediction confidence and distributional distance.

1. C_{prox} **(Boundary Proximity / Confidence):** Derived from the calibrated survival probability p_{cal}. We define uncertainty $U = 2|p_{cal}\ 0.5|$ (where 0 is 0.5 probability, 1 is 0.0/1.0 probability). Thus, $C_{prox} = U$. High values indicate decisive predictions far from the decision boundary.
2. S_{ood} **(In-Distribution Score):** Based on Mahalanobis distance D_M in the Ridge-projected latent space (reduced to 46 PCs explaining 95% variance). We normalize by the 99th percentile of the Reference Manifold ($_{ood} = 10.31$).

$$S_{ood} = 1\ \text{clamp}(D_M/_{ood}, 0, 1)$$

3. C_{data} **(Completeness):** A discrete score reflecting modality presence: $C_{data} = 0.5 + 0.3\ \ I_{WSI} + 0.2\ \ I_{clinical}$.
4. $N_{density}$ **(Manifold Density):** Cosine similarity to the $k = 10$ nearest neighbors in the Reference Manifold (PC space), normalized to [0,1].

3.3.2 PolicyProfile Registry To resolve ambiguity in tier definitions, we define a **PolicyProfile Registry** (Table 1) that maps ISS scores to tiers based on the cancer types baseline performance. Separately, a **Modality Requirement** lookup enforces vetoes.

Table 1: PolicyProfile Registry

Policy Class	Definition (Val C-index)	ISS Threshold	Example Cancers
Permissive	$C > 0.65$	ISS 0.20	BRCA, LGG, LIHC, PAAD
Moderate	0.55 C 0.65	ISS 0.35	COAD, KIRC, UCEC, SARC
Aggressive	$C < 0.55$	ISS 0.50	GBM, LUAD, OV, BLCA

Veto Logic: * **No-Green Veto:** If a patient is missing a required modality (e.g., WSI for BRCA, KIRC, SKCM), the maximum tier is capped at **YELLOW**. * **Abstention Veto:** If ISS < ISS Threshold, tier is **RED** (below threshold) or **YELLOW** (near threshold but below GREEN eligibility).

3.3.3 Calibration Protocol We employ **Per-Horizon Isotonic Regression** [10]. Separate isotonic regressors are fit for T $\{1, 2, 3, 5\}$ years on the TCGA validation set. The ISS uses the 3-year horizon probability for C_{prox}.

3.4 Cryptographic Verification (SolverInterlock)

To prevent configuration drift, the inference configuration (weights hash, calibrator hash, policy ID) is signed using an **Ed25519** private key [14]. The **SolverInterlock** verifies this signature before GPU allocation.

3.5 Frozen Protocol Declaration

To ensure valid external evaluation, the following components were fixed using **only** TCGA data (Training/Validation splits) and were **never** updated using external cohorts (CPTAC/CGGA/SCAN-B): 1. **Ridge Regressor:** Weights fit on TCGA Training (RNA Latent). 2. **PCA Basis:** Fit on Ridge-projected TCGA Validation. 3. **Reference Statistics:** $_{ref}, _{ref}$ computed on Ridge-projected TCGA Validation. 4. **Isotonic Calibrators:** Fit on TCGA Validation predictions. 5. **ISS Weights & Thresholds:** Fixed a priori based on internal ablation.

3.6 Statistical Analysis

Discriminative performance was assessed using Harrells C-index [23]. Survival separation was tested using the log-rank test. 95% Confidence Intervals (CIs) were computed via bootstrapping ($B = 1,000$).

4. Experiments & Results

4.1 Reliability Labeling Stratifies Performance

We applied the TraceabilityEngine to stratify external patients. Results are summarized in Table 2.

Table 2: Performance Stratification by Reliability Tier

Cohort	Metric	Baseline (All)	**GREEN (Reliable)**	YELLOW (Caution)	RED (Abstain)
CPTAC (Pan-Cancer)	N (Coverage)	1,031 (100%)	**229 (22.2%)**	456 (44.2%)	346 (33.6%)
	C-index	0.686	**0.744** [0.6920.793]	0.565	0.709
	Median OS (d)		697	457	387
	GREEN-RED		**+310d** ($p = 8.6e^{7}$)		
CGGA (Glioma)	N (Coverage)	970 (100%)	**596 (61.4%)**	0 (0.0%)	374 (38.6%)
	C-index	0.708	**0.675** [0.6430.706]	N/A	0.571
	Median OS (d)		1,392	N/A	400
	GREEN-RED		**+992d** ($p = 7.2e^{40}$)		
SCAN-B (Breast)	N (Coverage)	3,069 (100%)	**0 (0.0%)**	2,915 (95.0%)	154 (5.0%)
	C-index	0.501	N/A	0.501	0.497

Key Findings: 1. **Discrimination Uplift:** In CPTAC, the GREEN tier improved discrimination by **+5.8 pp** over baseline (0.744 vs 0.686). 2. **Survival Separation:** GREEN-vs-RED survival separation was large and significant in both cohorts. The triage system identifies patients for whom the models predictions correlate with actual outcomes. 3. **Safety Veto:** In SCAN-B (RNA-only breast cancer), the policy correctly triggered the **No-Green Veto**. Because BRCA is a WSI-dependent cancer type, the system capped the maximum tier at YELLOW. Consequently, coverage was 0% GREEN, 95% YELLOW. This confirms the system can enforce safety constraints even when omics data is technically sufficient to generate a score.

4.2 The RED > YELLOW Paradox

In CPTAC, the RED tier achieved a higher C-index (0.709) than the YELLOW tier (0.565). Analysis of the ISS components reveals: * **RED Tier:** Dominated by high D_M (OOD distance). These patients often present with late-stage, high-burden tumors that are biologically distinct from the training set (OOD) but have very short survival times, making them easy to rank correctly despite model extrapolation. * **YELLOW Tier:** Composed of intermediate-risk patients where the decision boundary is subtle (C_{prox} 0). This tier represents the hard casesin-distribution but maximally uncertain. * **GREEN Tier:** Patients who are both in-distribution and biologically coherent.

4.3 CGGA Distribution

In CGGA (Glioma), we observed 0% YELLOW coverage. Under the Permissive policy (ISS threshold 0.20), patients are classified as GREEN if they meet the threshold and have no modality vetoes, or RED otherwise. Since LGG does not require WSI, there is no mechanism to assign YELLOW (which requires a modality veto). Consequently, the ISS scores produced a binary split: patients either met the threshold (GREEN, $N = 596$) or fell below it (RED, $N = 374$).

4.4 Component Analysis

The ISS combines four conceptually distinct dimensions. - **Confidence (C_{prox}):** Identifies patients near the decision boundary. - **OOD distance (S_{ood}):** Identifies patients outside the training manifold regardless of prediction confidence. - **Completeness (C_{data}):** Enforces hard gates for WSI-dependent cancers. - **Density ($N_{density}$):** Detects rare phenotypes.

5. Discussion

5.1 Operationalizing Trust

This study demonstrates that trust need not be a vague sentiment but can be a computable property. By defining a **Reference Manifold** that matches the deployment encoding (Ridge), we successfully identified reliable patients across diverse external cohorts. The overall reliable coverage across all evaluated cohorts is **16.3%** (825 GREEN patients out of 5,070 unique external patients). While modest, it represents a conservative lower boundthe system errs toward abstention.

5.2 Clinical Workflow Implications

The Traffic Light system supports a tiered clinical workflow: * **GREEN:** The model is competent. The risk score and mechanistic parameters (,) can be included in the clinical report. * **YELLOW:** The model is uncertain. The prediction should be flagged for Human Review. * **RED:** The model is incompetent. No score is displayed.

5.3 The False Density Trap

We observed that applying per-cohort affine calibration ($z_{cal} = z_{ext} +$) to reduce distribution shift improved GREEN certificate coverage (0% 25.5%) but **degraded pooled C-index** (5.0 pp). This confirms that biological signal is often encoded in the distribution shift itself. It is safer to detect and abstain on shift (via ISS) than to artificially smooth it away.

6. Limitations

1. **Ridge Bottleneck:** The reliance on Ridge projection for external RNA encoding establishes a performance lower bound.
2. **Retrospective Validation:** While we used a frozen protocol, prospective validation is required.
3. **WSI Dependency:** The strict veto for missing WSI in breast cancer limits the tools utility in resource-constrained settings.

7. Conclusion

We have demonstrated that **Runtime Reliability Labeling** effectively identifies when a clinical AI model is reliable under distribution shift. By rigorously defining the reference manifold and coupling selective prediction with cryptographic enforcement, the DNAI platform achieves a robust standard for deployment safety.

Data Availability

TCGA: [https://portal.gdc.cancer.gov/]. CPTAC: [https://proteomics.cancer.gov/data-portal]. CGGA: [http://www.cgga.org.cn/]. SCAN-B: NCBI GEO (GSE96058).

Code Availability

Core logic for TraceabilityEngine and ISS is available from the corresponding author upon reasonable request. Foundation model weights available upon request.

Author Contributions

P.M.S.: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing.

Competing Interests

P.M.S. is the lead architect of the DNAI platform.

Ethics Statement

This study utilized de-identified, publicly available retrospective data. IRB approval was not required.

8. References

1. Zech, J. R., et al. (2018). *PLoS Medicine*, 15(11), e1002683.
2. Obermeyer, Z., et al. (2019). *Science*, 366(6464), 447-453.
3. Sagawa, S., et al. (2020). *ICLR*.
4. Koh, P. W., et al. (2021). *ICML*.
5. Swedenborg, P. M. (2026). DNAI: A physics-constrained cancer digital twin platform. *Preprint*.
6. Angelopoulos, A. N., & Bates, S. (2021). *arXiv:2107.07511*.
7. El-Yaniv, R., & Wiener, Y. (2010). *JMLR*, 11, 1605-1641.
8. Geifman, Y., & El-Yaniv, R. (2017). *NeurIPS*.
9. Guo, C., et al. (2017). *ICML*.
10. Austin, P. C., et al. (2020). *Statistics in Medicine*.
11. Mitchell, M., et al. (2019). *FAT**.
12. Gebru, T., et al. (2021). *CACM*.
13. Gal, Y., & Ghahramani, Z. (2016). *ICML*.
14. Josefsson, S., & Liusvaara, I. (2017). *RFC 8032*.
15. Lee, K., et al. (2018). *NeurIPS*.
16. Ledoit, O., & Wolf, M. (2004). *J. Multivar. Anal.*.
17. Collins, G. S., et al. (2024). *BMJ*.
18. Wolff, R. F., et al. (2019). *Ann. Intern. Med.*.
19. Johnson, W. E., et al. (2007). *Biostatistics*.
20. Korsunsky, I., et al. (2019). *Nat. Methods*.
21. Jiang, H., et al. (2018). *NeurIPS*.
22. Candès, E., et al. (2023). *JRSS*.
23. Uno, H., et al. (2011). *Statistics in Medicine*.
24. Van Calster, B., et al. (2019). *Ann. Intern. Med.*.
25. FDA. (2023). *PCCP for AI/ML*.
26. FDA, Health Canada, MHRA. (2021). *GMLP*.
27. EU Commission. (2024). *AI Act*.
28. NIST. (2023). *AI RMF 1.0*.
29. Vickers, A. J., & Elkin, E. B. (2006). *Med Decis Making*.