Synthetic Data for Clinical Model Auditing

Yingxu Wang
University of California, Los Angeles
yingxuw@ucla.edu

Joshua Ward
University of California, Los Angeles
joshuawarducla.edu

Yue Zhang
University of Utah
yue.zhang@utah.edu

Robert E. Tillman
Optum AI Labs
rob.tillman@optum.com

Guang Cheng*
University of California, Los Angeles
guangcheng@ucla.edu

Abstract

Synthetic data offers compelling advantages, including data augmentation and privacy protection, yet its adoption in high-stakes clinical settings remains limited. At the same time, machine learning models deployed in clinical environments are vulnerable to endpoint drift, where changing patient populations, clinical practices, or data collection processes can silently degrade reliability. We propose a framework that uses a locked synthetic reference distribution to audit deployed clinical models. This enables label-free monitoring when ground-truth outcomes are delayed or unavailable, while preserving institutional privacy constraints. We argue that this provides a practical and lower-risk use of synthetic data for post-deployment model auditing in healthcare.

1 Introduction

In high stakes clinical settings, machine learning models are increasingly used for tasks ranging from predicting patient deterioration in intensive care units to estimating treatment efficacy in pharmaceutical trials. In these environments, decisions directly impact patient well-being, and models must undergo rigorous validation to ensure reliability and safety. As a result, practitioners and regulators remain cautious about relying on synthetic data, which may fail to fully capture real-world complexity or rare but clinically critical edge cases.

At the same time, models deployed in clinical settings face a fundamental challenge: the data distribution encountered in practice often differs from that seen during training. Shifts in patient populations, clinical protocols, and data collection processes can lead to silent degradation in model performance, where predictions remain confident but unreliable.

We propose a framework that uses synthetic data to audit deployed clinical models, focusing on detecting endpoint drift. Rather than training models, synthetic data is used to construct a privacy-preserving reference distribution against which incoming data can be continuously compared. Significant divergence from this reference signals that the model may be operating outside its validated regime. Such monitoring is critical for patient safety, as prior work shows that models often perform poorly on outlier patients [15].

This framework addresses three key challenges: enabling ongoing model auditing without access to sensitive training data; detecting distributional shifts without requiring ground-truth labels; supporting timely clinical decision-making by flagging when model outputs may no longer be reliable. Beyond its technical role, the use of synthetic data for model validation in clinical trials carries significant economic and societal benefits. It can reduce the cost and latency of validation pipelines, facilitate cross-institutional collaboration under strict privacy constraints, and help mitigate biases by enabling broader representation of under-served patient populations.

*Corresponding author

2 Motivation

2.1 The Data Scarcity Problem

Access to healthcare data is constrained by regulation. Under HIPAA's Privacy Rule (Title 45 CFR §164.502(b)), disclosures of protected health information must be limited to the minimum necessary for the intended purpose [13]. As a result, the datasets available for model validation are frequently small and fragmented, or non-representative. Tsegaye et al., (2025) found only 17/36 studies reported having enough data for the minimum sample size to even be calculated. Only 5/36 met the recommended minimum sample size required to estimate risk and minimize overfitting [12]. Meanwhile Osei-Bonsu (2025) found that 73% of clinical datasets used in AI training sampled exclusively Europe and North America, which only represent 22% of the global population [5, p. 2139].

The consequence of this data scarcity is a decrease in model performance when extrapolated to distributions different from the original training set. Even across hospitals within the same country, this performance degradation is substantial. Wong et al. (2021) found that a widely deployed proprietary sepsis prediction model achieved an AUROC of just 0.63 when externally validated at a different hospital system, far below the 0.76–0.83 range reported during internal development [16]. More broadly, a systematic review of AI diagnostic models in radiology found that 81% exhibited decreased accuracy on external datasets, with nearly a quarter experiencing an AUROC drop of 0.10 or greater [10, p. 8804].

2.2 The Regulatory Caution Problem

Research on synthetic data generation has accelerated rapidly in recent years, driven by advances in generative adversarial networks, variational autoencoders, and more recently large language models [8, 3]. Academic publications have demonstrated promising results across tabular electronic health records, medical imaging, and genomic data, with synthetic datasets shown to preserve statistical properties of the original data while reducing risk of inferring training data from the synthetic set [4, 14, 2]. Yet despite this progress, clinical deployment remains extremely limited. As of 2025, neither the FDA nor the EMA has approved any medical application based solely on synthetically generated data [6]. This disconnect between academic enthusiasm and practical adoption reflects a convergence of regulatory, institutional, and trust-related barriers.

In contrast, synthetic data has started to gain traction in other industries. For instance, in the autonomous vehicles sector, Waymo has shown augmenting training data with synthetic data improves model performance[19, p. 8]. In financial services, institutions such as JPMorgan Chase are actively researching generative models to produce realistic financial records that maintain statistical fidelity without exposing customer data [7]. These developments suggest that synthetic data could enable comparable progress in healthcare, provided that its use is governed by rigorous validation and appropriate safeguards.

2.3 Synthetic Data as a Complement

Synthetic data generation offers a promising path to address the constraints discussed in the previous section. More importantly, its practical deployment has recently become more user-friendly through the incorporation of user intent into the generation process; see Synthony [9].

A generative model trained on historical patient data can produce arbitrarily large datasets that approximate the statistical properties of the original population without exposing any individual patient's information. This has two immediate benefits for clinical model auditing. In terms of scale, there is no inherent limit to how much synthetic data can be produced. In terms of privacy, synthetic data can be shared across institutional boundaries and used by auditing teams without triggering the regulatory and ethical concerns associated with real patient data. Most importantly, the framework proposed in this paper does not seek to replace real clinical data in trial design or regulatory submission. Instead, we leverage synthetic data for a de-risking purpose: detecting endpoint drift in deployed models. By continuously comparing incoming data against a synthetic reference distribution, we add a supplementary layer of post-deployment monitoring that strengthens patient safety without interfering with established evidence standards.

3 Endpoint Drift in Clinical Settings

3.1 Defining Endpoint Drift

Endpoint drift occurs when the statistical relationship between a model's input features and its target outcome changes over time. In clinical settings, this can arise from multiple sources of variability.

- **Demographic shifts:** Changes in the patient population served by a clinical site, including age distribution, comorbidity profiles, or socioeconomic composition.
- **Evolving clinical practices:** Updates to treatment protocols, surgical techniques, or care pathways that alter the trajectory of patient outcomes.
- **Diagnostic criteria revisions:** Changes to how diseases are classified or staged, which can redefine the outcome a model was trained to predict.
- **Data collection changes:** Modifications to electronic health record systems, laboratory equipment calibration, or data entry workflows that alter feature distributions without any underlying change in patient health.
- **Treatment improvements:** Advances in pharmacotherapy, screening technologies, or supportive care that change how outcomes are recorded or labelled.

Over time, these factors can cause the statistical distribution of the target variable to deviate from what was observed during model development, leading to degraded predictive performance and potentially unsafe or misleading clinical recommendations.

3.2 The Ground-Truth Problem

What makes endpoint drift particularly dangerous in healthcare is that ground truth is often unavailable during the period when monitoring is most needed. During an active clinical trial, the true outcome (e.g., progression-free survival, treatment response) may not be observable for months or years. In urgent care settings, there is no time to wait for outcome data before deciding whether a model's predictions should be trusted.

This creates a fundamental monitoring challenge. Traditional approaches to detecting model degradation rely on comparing predictions to observed outcomes. When outcomes are delayed or absent, we must instead monitor the inputs to the model and assess whether the incoming data still resembles the data it was trained and validated on. If not, there is a risk that some underlying conditions or assumptions are no longer true, making the model's predictions unreliable and thus endangering the patients. This is the core motivation for our proposed framework.

4 Synthetic Data as Validation in Monitoring Endpoint Drift

Given a feature vector $X \in \mathbb{R}^d$ and a model $f^* : \mathbb{R}^d \to \mathbb{R}$, we define the model output as $S = f^*(X)$. Let $\phi(X) \in \mathbb{R}^p$ denote a fixed feature representation. We monitor the joint variable $W = (\phi(X), S)$, which captures both the input data and the corresponding model outputs. Our framework consists of three phases: synthetic reference generation, baseline locking, and continuous monitoring. Each phase is designed to be operationally simple while enabling statistically principled, label-free drift detection.

4.1 Synthetic Reference Generation

Prior to deployment or trial launch, a generative model is trained on historical patient data X from the target clinical context. The generative model learns to approximate the distribution of X, and therefore induces a distribution over $W = (\phi(X), f^*(X))$. Here, the f^* was learned to capture the relation between X and S that can be validated in real clinical settings of interest. The constructed synthetic reference distribution is denoted as Q_W. Monitoring the joint distribution of W rather than X alone improves sensitivity to shifts that are most relevant to the behavior of the model.

The choice of generative model is flexible. Depending on the complexity of the data and the fidelity requirements, options include variational autoencoders, generative adversarial networks, graphical models, or diffusion-based approaches. The key requirement is that the synthetic data is of sufficient quality to serve as a meaningful reference distribution [18].

4.2 Baseline Locking

Once the synthetic reference dataset has been generated and validated, it is **locked** prior to trial initiation or model deployment. This includes fixing the model f^*, the feature representation ϕ, the generative model specification, and all pre-processing steps.

The locked reference set serves as a fixed representation of the data environment at the time the model was considered fit for purpose. All subsequent comparisons are made against this frozen baseline, ensuring that drift detection reflects genuine changes in the incoming data rather than changes in the reference itself.

Because the reference set is synthetic, it addresses the aforementioned ground truth problem. Rather than validating model predictions, we instead ask whether incoming data still resembles the data under which the model was trained and validated. This provides an additional proxy, without the need for labels, for testing model reliability: if the joint distribution of incoming $(\phi(X), f^*(X))$ is close to the baseline, model performance is more likely to remain reliable as well.

4.3 Continuous Monitoring

During the trial or deployment period, incoming feature data induces a distribution P_W over $W = (\phi(X), f^*(X))$. At regular intervals, samples from P_W are compared against the locked synthetic reference distribution Q_W. *The key point here is to monitor the deviation of P_W from Q_W instead of only monitoring the fidelity of generated X.* This also echoes with the recent finding that the utility of synthetic data depends more on how well f^* is learned, but less on the fidelity of generated features X [17].

The monitoring process uses distributional drift metrics, including multivariate two-sample tests and conformity-based scores, to evaluate whether the deployment data can be statistically distinguished from the reference.

Specifically, to quantify the magnitude of drift, we adopt a contamination model in which P_W is expressed as a mixture of the reference distribution and an arbitrary alternative: $P_W = (1 - \pi)Q_W + \pi R_W$, where $\pi \in [0, 1]$ represents the fraction of observations that deviate from the reference regime. In practice, π can be estimated using conformal p-values and mixture modeling, yielding an interpretable measure of drift (e.g., a threshold such as $\pi \geq 0.25$).

Several classes of drift metric are applicable.

- **Classifier-based two-sample tests:** A discriminator model is trained to distinguish between synthetic reference data and incoming trial data. High classification accuracy indicates distributional divergence; see [11].
- **Propensity-based distinguishability metrics:** Propensity scores are estimated for each observation, quantifying the likelihood that a data point belongs to the trial set versus the reference set. Systematic differences in propensity scores signal drift.
- **Divergence measures:** Statistical divergence metrics such as the Kullback–Leibler divergence, Jensen–Shannon divergence, or maximum mean discrepancy can be applied to the feature distributions or to model output distributions.

When the distributions become increasingly distinguishable beyond a predefined threshold (based on the above drift metric), an alert is generated. This alert indicates that the trial data may be moving into regions of the feature space that were not well represented during model development, raising the possibility of degraded endpoint reliability.

This framework provides a rigorous and operationally actionable mechanism for detecting when the joint behavior of inputs and model outputs departs from the validated regime.

4.4 Decision Framework

A central motivation for this framework is to support rapid decision-making in clinical environments while mitigating risks for patients. In many deployment scenarios, such as urgent care and active trial monitoring, there is no opportunity to retrain models or wait for additional outcome data. Thus the question becomes not whether a model can be improved, but whether it can still be trusted right now.

The monitoring alerts feed into a simple decision framework. If drift is within acceptable bounds, the model continues to operate as deployed. If drift exceeds the threshold, clinical teams are notified that the model's outputs may no longer be reliable for the current patient population, and fallback procedures are invoked. The framework thus functions as a risk management tool, enabling clinicians and trial managers to identify when a model-based decision is potentially unreliable and signaling to switch to a safer alternative before patient outcomes are affected.

5 Advantages of Synthetic Data

While one might consider using the original training dataset as the reference baseline, it is fundamentally limited in this role. First, a real dataset is finite and often sparse in clinically important regions of the feature space, leading to inadequate coverage of rare or high-risk cases. This scarcity limits the ability to reliably characterize the full data environment against which future observations should be compared.

Second, reliance on real data introduces governance and operational constraints. Even when privacy is not the primary concern, access to raw datasets is frequently restricted, logged, or institution-specific. This limits the ability of independent monitoring teams, regulatory reviewers, or external auditors to directly reproduce analyses, creating a gap between reported results and verifiable evidence.

Lastly, synthetic data enables systematic exploration of regions of the feature space that are rare or underrepresented in the original dataset. This is particularly important for risk assessment: real-world datasets often lack sufficient coverage of edge cases, yet these are precisely the scenarios where model failures are most consequential. A synthetic reference can be used to probe these regions, providing insight into model behavior under conditions that may not yet have been observed but are nonetheless plausible.

Beyond monitoring, synthetic data can be used proactively to stress-test models before deployment. By generating synthetic datasets that represent plausible but challenging shifts in the feature distribution (e.g., an older patient population, a higher rate of comorbidities, or changes in lab value ranges), teams can evaluate model robustness under conditions that may arise in practice. This stress-testing paradigm uses synthetic data as a tool for pre-deployment risk assessment rather than post-deployment monitoring and complements the monitoring framework described above.

6 Conclusion

Deploying machine learning models in clinical settings requires ongoing vigilance. Models validated on historical data can silently degrade as the data environment changes, and the absence of real-time ground truth in many clinical contexts makes traditional monitoring approaches infeasible. This paper has proposed a framework that uses synthetic data to fill this gap, providing a privacy-preserving, distributable reference baseline against which incoming data can be continuously compared.

The framework is deliberately narrow in scope. In critical fields such as healthcare where mistakes can lead to severe consequences, it is understandable that the fidelity and reliability of synthetic data is subject to increased scrutiny. Our framework does not aim to disrupt these rigorous safety standards. There is no attempt to replace clinical trial data, train new models, or automate clinical decisions. Instead, the framework provides a practical tool for risk management: a way for clinical teams to know when a model's outputs may no longer be reliable. By monitoring inputs instead of waiting for outcomes, the framework enables proactive rather than reactive risk mitigation.

We believe this approach represents a responsible and low risk application of synthetic data in healthcare. The primary motivation is improved patient well-being through a more robust verification process for the models that increasingly inform clinical care.

References

[1] Pablo A. Apellániz, Juan Arroyo, and Juan A. Cuesta-Albertos. Divergence-based validation of synthetic data. *arXiv preprint arXiv:2405.07822*, 2024.

[2] Jessup Byun, Xiaofeng Lin, Josh Ward, and Guang Cheng. Risk in context: Benchmarking privacy leakage of foundation models in synthetic tabular data generation. *arXiv: 2507.17066*, 2025.

[3] Guang Cheng. A rising opportunity from synthetic data: Generative data science. 2026.

[4] Matteo Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *npj Digital Medicine*, 6:186, 2023.

[5] Johnson Gbenga Oyeniyi. From lab to clinic: Addressing bias and generalizability in ai diagnostic systems. *World Journal of Advanced Research and Reviews*, 28(3):2134–2179, 2025.

[6] Giuseppe Pasculli et al. Synthetic data in healthcare and drug development: Definitions, regulatory frameworks, issues. *CPT: Pharmacometrics & Systems Pharmacology*, 2025.

[7] Vamsi K. Potluru, Daniel Borrajo, Andrea Coletta, Niccolò Dalmasso, Yousef El-Laham, Elizabeth Fons, Mohsen Ghassemi, Sriram Gopalakrishnan, Vikesh Gosai, Eleonora Kreačić, Ganapathy Mani, Saheed Obitayo, Deepak Paramanand, Natraj Raman, Mikhail Solonin, Srijan Sood, Svitlana Vyetrenko, Haibei Zhu, Manuela Veloso, and Tucker Balch. Synthetic data applications in finance, 2024.

[8] Daniel Smolyak, Margrét V. Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. Large language models and synthetic health data: Progress and prospects. *JAMIA Open*, 7(4):ooae114, 2024.

[9] Hochan Son, Xiaofeng Lin, Jason Ni, and Guang Cheng. Synthony: A stress-aware, intent-conditioned agent for deep tabular generative models selection. *ICLR Workshop on Deep Generative Model in Machine Learning*, 2026.

[10] MU Suleman, M Mursaleen, U Khalil, A Saboor, M Bilal, SA Khan, MA Subhani, MA Hussnain, SN Tabassum, and M Tahir. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. *Annals of Medicine and Surgery*, 87(12):8803–8811, 2025.

[11] Lan Tao, Shirong Xu, Chi-Hua Wang, Namjoon Suh, and Guang Cheng. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv:2405.15337*, 2024.

[12] Biruk Tsegaye et al. Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: Methodological systematic review. *Journal of Clinical Epidemiology*, 180:111675, 2025.

[13] U.S. Department of Health and Human Services. Standards for privacy of individually identifiable health information. `https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.502`, 2013. 45 CFR §164.502(b), accessed April 2026.

[14] Josh Ward, Xiaofeng Lin, Chi-hua Wang, and Guang Cheng. Synth-mia: A testbed for auditing privacy leakage in tabular data synthesis. *arXiv: 2509.18014*, 2025.

[15] Chu Weng, Joshua Ward, Wesley Lin, Sherry Dong, Qi Liu, and Hanrui Zhang. Out-of-distribution detection as a risk-control strategy for medical classification machine learning models. *Clinical and Translational Science*, 2025.

[16] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krez, Jason Lin, Tamas Romber, Nishant Kenber, Celia Yber, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021.

[17] Shirong Xu, Wei Sun, and Guang Cheng. Utility theory of synthetic data generation. *arXiv:2305.10015*, 2023.

[18] Shirong Xu, Will Wei Sun, and Guang Cheng. Utility theory of synthetic data generation. *arXiv preprint arXiv:2305.10015*, 2023.

[19] Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, and Henrik Kretzschmar. Surfelgan: Synthesizing realistic sensor data for autonomous driving. *CoRR*, abs/2005.03844, 2020.

A Validating Synthetic Data Fidelity

The reliability of the monitoring framework depends on the quality of the synthetic reference data. If the synthetic dataset does not faithfully represent the original data distribution, drift detection will be unreliable, either failing to detect genuine drift or generating false alarms.

A.1 Distribution Matching

Evaluating synthetic data fidelity requires more than matching the distributions of individual columns. This is because preserving individual distributions does not necessarily mean the joint distributions are also preserved. Consequently, dependencies between columns may be lost. Such failures in multivariate fidelity would undermine the reliability of any drift metric that depends on the joint distribution.

A.2 Divergence-Based Validation

Recent work has proposed divergence-based validation methods that address this challenge. Apellániz et al. described an approach in which a discriminator model is trained to distinguish between real and synthetic observations. The discriminator's output can be used to estimate density ratios between the two distributions, enabling the computation of statistical divergences without explicitly estimating high-dimensional probability densities. This provides a scalable and practical method for assessing fidelity in complex multivariate datasets [1].

In our framework, this validation step serves as a quality gate. Before the synthetic reference set is locked, its fidelity is assessed using discriminator-based divergence metrics. If the synthetic data can be readily distinguished from the original training data, the generative model is refined or retrained until a satisfactory level of indistinguishability is achieved.

A.3 ML Utility as a Fidelity Criterion

An additional validation criterion is machine learning utility. A model trained on real data and evaluated on synthetic data should achieve comparable performance to one trained and evaluated completely on real data. If there is a significant performance degradation, then it follows that the synthetic data distribution likely deviates from the real distribution in ways that are relevant to the prediction task. Note that this option is included to show that there are other ways to test fidelity. It is unlikely to be relevant in our specific use case since real data is scarce.

B Privacy Protection

Ideally, authentic patient data would be used to verify model endpoints and detect distributional drift. In practice, however, such data is difficult to obtain and carries significant privacy risks. Datasets that are accessible are often fragmented and small, introducing sampling bias that can itself produce spurious drift signals. Synthetic data addresses both of these limitations simultaneously.

B.1 Training Data Inference

A fundamental question in any synthetic data application is how much of the original data can be inferred from the synthetic output by an adversary. This question has been formalized in the privacy literature through concepts such as membership inference resistance and attribute disclosure risk. The strength of the privacy guarantee depends on the generative model architecture, the training procedure, and any additional privacy mechanisms applied during data generation.

Differential privacy provides one formal framework for bounding disclosure risk. By introducing calibrated noise during the training of the generative model, differentially private synthetic data generators can provide mathematical guarantees that the inclusion or exclusion of any single patient's data in the training set has a bounded effect on the synthetic output. Several practical implementations of differentially private generative models have been developed for tabular health data.

B.2 Privacy in the Monitoring Context

The monitoring use case described in this paper has a favorable privacy profile for several reasons. The synthetic reference set is generated once and locked before deployment. It does not need to be regenerated as new patient data arrives. The monitoring process itself only requires computing aggregate statistical comparisons between the reference distribution and incoming data; it does not require linking synthetic records to real individuals. These properties limit the surface area for privacy attacks and make the framework amenable to conservative privacy guarantees.

C Connections to Transfer Learning and Out-of-Distribution Detection

The problem addressed in this paper intersects with two well-studied areas of the machine learning literature: transfer learning and out-of-distribution (OOD) detection.

C.1 Transfer Learning

Transfer learning concerns the application of a model trained in one domain or data distribution to a related but different one. In the clinical context, a model developed on historical data from one patient population is effectively being "transferred" to a new population whenever conditions change. The literature on domain adaptation, covariate shift, and dataset shift provides theoretical grounding for why model performance can degrade when the deployment distribution differs from the training distribution. Our framework operationalizes this insight by providing a practical mechanism to detect when such a shift has occurred.

C.2 Out-of-Distribution Detection

OOD detection methods aim to identify individual data points or batches that fall outside the training distribution of a model. These methods are complementary to our approach. While OOD detectors typically flag individual predictions as unreliable, our framework assesses whether the overall data environment has shifted, providing a population-level rather than instance-level view. The two approaches can be combined: the synthetic reference monitoring framework provides a macro-level drift signal, while OOD detection on individual predictions provides micro-level risk flagging.

D Risks and Limitations

It is important to acknowledge the limitations of using synthetic data since these are the main reasons why regulators have been slow to accept the use of synthetic data in healthcare.

D.1 Synthetic Data Fidelity Risk

The most immediate risk is that the synthetic reference data may not accurately capture the joint distribution of the original data. If the synthetic data has systematic biases, drift detection may be unreliable. This risk is mitigated by the validation procedures described in Section 5, but it cannot be eliminated entirely, particularly for complex, high-dimensional clinical datasets. In particular, one of the proposed fidelity measures of comparing ML utility is inappropriate for this use case.

D.2 Threshold Calibration

Choosing appropriate thresholds to flag drift alerts is a nontrivial problem. Thresholds that are too sensitive will produce frequent false alarms, eroding clinical trust in the monitoring system. Thresholds that are too lenient will fail to detect meaningful drift until model performance has already degraded. Calibration should be informed by historical data on the relationship between distributional shift and model performance degradation, ideally under the supervision of a subject matter expert.

D.3 Scope of Claims

It is important to be clear about what this framework does and does not do. We are not proposing to use synthetic data to replace clinical trial data, to train predictive models, or to make direct clinical decisions. The synthetic data serves a single, well-defined function: to provide a stable, privacy-preserving reference baseline for distributional comparison. The framework detects changes in model inputs; it does not diagnose the cause of those changes or predict their impact on model outputs. When drift is detected, further investigation by domain experts is required to determine the appropriate response.