Beyond Single-Objective Accuracy: A Longitudinal Evaluation of Tabular Foundation Models under Dataset Shift

João Marcos Tomáz Silva Campos*,
Leonardo Chaves Dutra da Rocha†, Wagner Meira Junior† and Marcos André Gonçalves†
*Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
{joaomarcoscampos, mgoncalv, meira}@dcc.ufmg.br
†Universidade Federal de São João del-Rei (UFSJ), São João del-Rei, Brazil
lcrocha@ufsj.edu.br

***Abstract*—Evaluations of classification under dataset shift often rely on static benchmarks and single-objective metrics, obscuring how discrimination, calibration, and risk ranking may fail in different ways over time. We present a 19-year longitudinal benchmark of tabular models under real-world clinical shift, using a Brazilian renal registry with 9.6 million records from 1997–2015. We compare dynamically retrained tree-based ensembles with standard and drift-aware Tabular Foundation Models (TabPFN, TabICL) under uniform temporal protocols. Our results show that robustness is not monolithic: models with strong global survival rankings may suffer from calibration and same-year classification failures, while model rankings change across stable, covariate-shift, and structural/observation-process-shift regimes. We further introduce an exploratory Temporal Robustness Index and combine domain-aware SHAP with medication-record perturbations to diagnose model failures. The analysis reveals that standard models often rely on temporally unstable pharmaceutical proxies, whereas temporally conditioned models emphasize more stable physiological signals. Overall, our study provides a multi-objective and shift-aware auditing framework for evaluating tabular foundation models in dynamic clinical environments.**

***Index Terms*—Dataset Shift, Calibration, Temporal Robustness, Tabular Foundation Models, Clinical Forecasting**

I. Introduction

Evaluating machine learning models under dataset shift is a central challenge in high-stakes domains. Historically, evaluations of concept drift have relied on single-objective metrics, implicitly assuming that discriminative power, calibration, and risk ranking degrade uniformly [1]. In real-world longitudinal systems, this assumption rarely holds [2].

While tree-based ensembles [3], [4] have long dominated tabular prediction, recent Tabular Foundation Models (TFMs) such as TabPFN [5] and TabICL [6] introduce a new paradigm leveraging Transformers [7] and in-context learning. Their behavior under long-horizon temporal shift remains largely unexplored.

We investigate whether predictive discrimination, probabilistic calibration, and survival ranking lead to consistent conclusions about model quality under long-horizon dataset shift. To this end, we conduct a 19-year longitudinal evaluation on a national renal registry comprising 9.6 million records from 67,267 patients. Classical ensembles with dynamic sliding-window retraining are compared against standard TFMs and a temporally conditioned drift-aware variant under identical temporal windows. Models are assessed annually using discrimination, calibration, and IPCW-corrected survival ranking, complemented by shift-analysis and diagnostic techniques.

The results show that robustness is not a single property. Discrimination, calibration, and survival ranking diverge under temporal shift, and model rankings change across shift regimes. Standard TFMs perform strongly under stable conditions, whereas gradient boosting is more resilient under structural shift. Temporal conditioning further improves annual discrimination under drift. These findings show that aggregate accuracy can obscure clinically consequential failure modes and that model selection must account for both the evaluation objective and shift regime.

In sum, our contributions are:

- **A long-horizon benchmark for tabular models under shift.** We provide a long-horizon evaluation of classical ensembles, standard TFMs, and drift-aware TFMs under real-world dataset shift.
- **A multi-objective evaluation framework.** We jointly assess discrimination, calibration, and survival ranking, exposing failure modes hidden by single-metric evaluation.
- **Divergent objectives and ranking reversals under shift.** We show that evaluation objectives can lead to different conclusions about model quality and that model rankings reverse across shift regimes.
- **A shift-aware diagnostic methodology.** We combine multiple diagnostic analyses to explain when and why models fail.

II. Background and Related Work

A. Classification and Calibration under Dataset Shift

Most supervised learning methods, including tree-based ensembles such as XGBoost and LightGBM [3], [4], assume that training and test data are sampled from a stable distribution $P(X, y)$. In longitudinal clinical registries, this

assumption is routinely violated by dataset shift [1], [2], which may arise as covariate shift through changes in $P(X)$, label shift through changes in $P(y)$, or structural shift through changes in $P(y|X)$. Common mitigation strategies include sliding-window retraining, online adaptation, and periodic model updating [1], [8].

However, dynamic retraining does not by itself ensure robust evaluation. Existing benchmarks under dataset shift frequently summarize performance through a single discriminative metric, despite the fact that classification quality, probabilistic calibration, and longitudinal risk ordering may degrade differently under temporal shift [2]. Longitudinal benchmarks that jointly evaluate these objectives under real-world dataset shift remain comparatively rare. This limitation is particularly consequential in clinical settings, where a model may preserve relative risk ranking while producing unreliable probability estimates or unstable decision boundaries. Our work addresses this gap through a unified longitudinal protocol that jointly evaluates discrimination, calibration, and survival ranking across multiple shift regimes, enabling a more comprehensive assessment of model behavior under long-horizon dataset shift.

B. Tabular Foundation Models in Non-Stationary Environments

Tabular Foundation Models (TFMs) [9] extend the foundation-model paradigm to structured data by pretraining on large collections of synthetic tasks and performing prediction through in-context learning. TabPFN [5] approximates Bayesian posterior inference and achieves strong zero-shot performance, while TabICL [6] introduces attention mechanisms designed for larger tabular contexts. Drift-Resilient TabPFN [10] further incorporates explicit temporal conditioning during pretraining to improve adaptation across distributional regimes.

Despite rapid progress in static benchmarks, the behavior of TFMs under prolonged, real-world non-stationarity remains poorly understood. Prior studies have largely emphasized aggregate predictive performance, short-horizon perturbations, or synthetic drift, leaving open whether TFMs preserve calibration, discrimination, and risk ranking over years of evolving clinical and administrative practice. We provide one of the first large-scale longitudinal comparisons of standard and drift-aware TFMs against dynamically retrained classical ensembles under nearly two decades of observed shift.

C. Survival-Aware and Mechanistic Evaluation under Shift

Right-censoring complicates longitudinal evaluation because standard classification metrics may misrepresent patient-level risk trajectories [11]. Survival-aware measures such as the IPCW-weighted Concordance Index provide a principled alternative for assessing global risk ordering under censoring [11]. At the same time, post hoc attribution methods such as SHAP [12] can reveal whether predictions depend on stable physiological variables or temporally fragile proxies.

These tools are typically applied in isolation and rarely integrated into a common framework for distinguishing covariate from structural shift, comparing competing evaluation objectives, and testing whether attributed failure modes are plausible. Our study combines survival-aware evaluation, distributional diagnostics, domain-aware attribution, temporal robustness analysis, and controlled feature perturbation. This integration moves beyond documenting performance degradation by identifying *which* objectives fail, *under which* shift regimes, and *why* particular model families become vulnerable.

III. Methodology

A. Dataset and Problem Framing

Our study uses a longitudinal cohort derived from Brazil's national healthcare systems through deterministic and probabilistic record linkage [13]. The cohort comprises 67,267 patients and approximately 9.6 million clinical events recorded between 1997 and 2015. It covers more than 6,200 event types, including renal replacement therapies, chronic disease diagnoses, and medication dispensations, and exhibits pronounced temporal non-stationarity driven by changes in clinical protocols, administrative policies, and data-recording practices. The prediction target is all-cause mortality.

Although the data are longitudinal, our primary formulation is patient-year tabular prediction rather than sequence modeling. Preliminary experiments showed that the registry contains extreme sequence-length heterogeneity and that, under the severe reduction in recorded exposure observed in 2015, sequential architectures became highly sensitive to zero-padding patterns. These patterns reflect changes in the observation process rather than clinical state and may therefore act as spurious predictors. The patient-year representation mitigates this problem by encoding each annual history as a fixed-length vector of event frequencies, providing a more stable input space for comparing classical ensembles and Tabular Foundation Models under the same temporal conditions.

To satisfy TFM context constraints while retaining clinically relevant information, we select the 500 highest-ranked features according to LightGBM information gain computed exclusively on the 1997–2006 training period (Fig. 1). This procedure avoids using future test years during feature selection, although we acknowledge that model-based selection may favor predictors that are effective for tree-based learning. The resulting features are clinically coherent: hemodialysis, erythropoietin, and renal replacement therapies dominate, together with `c2MED`-prefixed medication records. These medication variables are highly predictive under relatively stable conditions, but, as later analyses indicate, are also sensitive to administrative and documentation changes.

All sampling is performed at the patient level. Because TFMs are constrained by finite context windows, they receive at most approximately 3,000 patient-year instances per class in each training window. Classical models use the complete balanced training set and therefore benefit from a larger effective sample size. We retain this asymmetry because it reflects the practical operating constraints of current TFMs; accordingly, comparisons should be interpreted as evaluations of deployable model configurations rather than as strictly

sample-matched capacity comparisons. A Fixed-K ablation further verifies that the main TFM trends are stable across alternative sampling configurations.

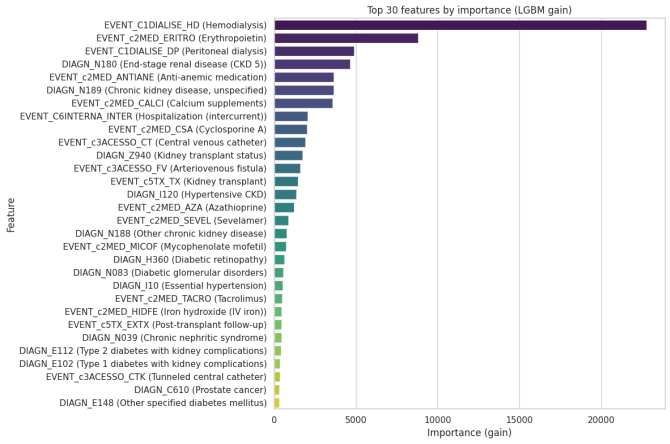

Fig. 1. Top 30 features by LightGBM information gain on the 1997–2006 training split. Core renal procedures dominate; `c2MED`-prefixed medication features emerge as high-importance predictors that are effective under relatively stable conditions but susceptible to administrative and documentation changes.

B. Evaluation Protocols

A common limitation of temporal-shift benchmarks is the use of a training set frozen at a single historical period. In practice, a model deployed in 2012 would be more plausibly updated with recent data than trained solely on records from the late 1990s. We therefore adopt two complementary temporal protocols, applied uniformly to all model families.

The **sliding-window protocol** constitutes our primary evaluation. For each test year $Y_{\text{test}} \in \{2007, \ldots, 2015\}$ and window size $W \in \{3, 5, 7, 10\}$, classical baselines are retrained from scratch using records from $[Y_{\text{test}} - W, Y_{\text{test}} - 1]$. TabPFN and TabICL receive the same temporal window as their in-context learning set, without gradient-based task-specific training. Drift-Resilient TabPFN additionally receives a temporal-domain tensor $d_i = \texttt{year\_to\_domain}[\text{year} * i]$, thereby enabling explicit conditioning on the temporal regime. We also evaluate an expanding-window variant trained on all available records from 1997 to $Y_{\text{test}} - 1$.

The **one-year-ahead stream protocol** serves as a robustness analysis of continuous deployment. For each $Y \in \{2000, \ldots, 2015\}$, every model is updated using all data available from 1997 to $Y - 1$ and then evaluated on year Y. The earliest years, 1998–1999, are excluded because the registry had not yet accumulated enough observations ($n < 50$) for balanced training. Agreement between this protocol and the main sliding- or expanding-window analyses provides evidence that the reported degradation patterns are not artifacts of a single update strategy.

Across both protocols, training data are balanced through patient-level sampling, with $N \leq 3{,}000$ instances per class for TFMs. Test sets preserve the naturally imbalanced mortality prevalence, which ranges from 4.2% to 26%, to approximate deployment conditions. Patients in a test pool never appear in the corresponding training window, preventing patient-level leakage across temporal partitions.

C. Multi-Objective Evaluation

Our primary evaluation treats robustness as a multi-objective property rather than as a single performance score. For each test year, we assess three related but non-equivalent objectives.

Discrimination is measured using Macro-F1, positive-class F1, and PR-AUC. These metrics evaluate annual mortality classification under severe class imbalance. The classification target is defined as the occurrence of any `DEATH` event within the same calendar year, and all post-death rows are removed to prevent leakage.

Calibration is assessed using Expected Calibration Error (ECE), which measures agreement between predicted probabilities and observed annual mortality frequencies. Because ECE is sensitive to sample size, prevalence, and binning, we interpret it as a comparative diagnostic of probability reliability across years and model families rather than as an absolute measure of calibration quality.

Global risk ranking is evaluated using the IPCW-adjusted Concordance Index [11] under a 2-year landmark survival horizon. The IPCW weights are computed as $W_i = \min(1/\hat{G}(T_i), \tau_{\max})$, with clipping at the 95th percentile to limit variance. This survival objective is intentionally broader than the annual classification target: it evaluates longitudinal ordering over a two-year horizon rather than same-year mortality decisions. The two tasks are therefore related but not identical, and divergence between them is interpreted as evidence of objective-specific robustness rather than as direct disagreement between interchangeable metrics.

Patient-level bootstrap confidence intervals for annual F1 estimates are computed using 1,000 resamples clustered by patient ID and are reported in the supplementary material. Year-over-year distributional change is quantified using Jensen–Shannon divergence over binary feature distributions.

D. Exploratory Shift Characterization and Failure Diagnostics

Beyond the primary metrics, we employ a set of exploratory diagnostics to characterize the observed shifts and investigate potential failure mechanisms. These analyses are intended to support interpretation rather than to define additional confirmatory endpoints.

The **Temporal Robustness Index** is defined as $\text{TRI}_y = (F1_y / F1_{\text{ref}})/(\text{JS}_y + \varepsilon)$, with $\varepsilon = 10^{-4}$. TRI contextualizes each model's performance retention relative to the magnitude of measured distributional change. Because the denominator can amplify differences when JS divergence is small, TRI is used exclusively as an intra-model diagnostic and not as a calibrated cross-model ranking statistic.

We further organize test years into an exploratory **drift-regime taxonomy**. Years with low JS divergence and relatively stable mortality prevalence are described as *Stable*. Years with

elevated feature divergence but limited prevalence change are characterized as *Type I Covariate Shift*. Years with simultaneous changes in feature sparsity, prevalence, censoring, and observed predictive relationships are characterized as *Type II Structural/Observation-Process Shift*. These labels summarize empirical patterns and domain evidence; they should not be interpreted as formal identification of the underlying causal distribution components. In particular, the 2015 regime reflects administrative truncation of the observation process, which changes the empirically available relationship between recorded features and outcomes without necessarily implying a change in the underlying biological mechanism.

To investigate model reliance on temporally stable or unstable predictors, **domain-aware SHAP** keeps temporal embeddings $E_{\mathrm{dom}}(d_i)$ fixed during permutation, so that attributions reflect feature contributions within the appropriate temporal context [12]. Finally, a **medication perturbation sweep** sets the 11 `c2MED` features to zero for increasing fractions $p \in \{0\%, \ldots, 100\%\}$ of 2015 test patients. This intervention does not establish a causal clinical effect of medication use; rather, it tests whether model predictions depend on medication-recording patterns that became unreliable under the 2015 administrative shift.

IV. Results and Dataset Shift Analysis

A. Divergence of Discrimination, Calibration, and Ranking

We evaluated all models year by year over the chronological test period (2007–2015) using the dynamic sliding-window protocol. Our central hypothesis is that robustness under long-horizon dataset shift is objective-dependent: discrimination, calibration, and global risk ranking may degrade differently and produce different model orderings. Table I supports this hypothesis across the evaluated architectures.

Predictive discrimination and survival ranking diverge substantially. TabICL achieves the strongest global risk ordering (IPCW C-index: 0.542), but this advantage does not transfer to same-year mortality classification in 2015, where its positive-class F1 falls to 0.109. In contrast, Drift-Resilient TabPFN achieves the best average annual F1 (0.316), PR-AUC (0.249), and 2015 F1 (0.212), although it does not lead the survival-ranking objective.

This divergence is practically relevant. Annual classification evaluates the same-year decision quality, whereas IPCW C-index evaluates patient ordering over a 2-year horizon. A model may therefore rank patients well over time while still providing unreliable probabilities or weak decision boundaries for immediate clinical action.

B. Characterizing Shift Regimes: Covariate (2012) and Structural/Observation-Process Shift (2015)

We contextualize degradation trajectories using JS divergence between binary feature distributions, interpreted alongside prevalence, sparsity, censoring, and domain evidence. Two distinct regimes emerge, but these labels summarize observed patterns rather than formally identifying changes in $P(X)$ or $P(y|X)$.

a) The 2012 Administrative Covariate Shift.: In 2012, all models show a sudden discriminative decline. Domain evidence links this period to SUS billing changes that temporarily altered medication-recording frequencies, including Sevelamer and Erythropoietin. JS divergence reaches approximately 0.042, while mortality prevalence remains comparatively stable, suggesting a Type I covariate shift affecting recorded inputs more than outcome frequency.

b) The 2015 Structural/Observation-Process Shift.: The 2015 cohort exhibits a more severe regime. Administrative truncation reduces exposure time, increases feature sparsity, and lowers recorded mortality prevalence from approximately 10% to 4.2%. These changes alter the empirical relationship between recorded features and observable outcomes without necessarily implying a biological change. We therefore describe 2015 as a Type II structural/observation-process shift. Under this regime, standard TFMs show relative Macro-F1 reductions above 10%, classical ensembles decline by approximately 9.4%, and Drift-Resilient TabPFN declines by approximately 3%.

Table II reports the Temporal Robustness Index. TRI is interpreted only within each model, since the ratio can become unstable when JS divergence is small. Under this interpretation, Drift-Resilient TabPFN retains the highest TRI in both 2012 and 2015, suggesting slower performance degradation relative to the observed distributional change.

c) Model Rankings Depend on Shift Regime.: Table III shows that aggregate rankings conceal regime-specific behavior. TabICL leads under stable conditions, all architectures perform similarly in the 2012 Type I regime, and XGBoost leads in the 2015 Type II regime. Although numerical differences should be read alongside supplementary uncertainty analyses, the pattern indicates that model suitability depends on the shift regime and cannot be inferred from aggregate performance alone.

C. Calibration under Long-Horizon Shift

Figure 2 shows reliability diagrams for the 2015 stress year under the fixed $W = 10$ window. Calibration deteriorates across all model families, but with different failure patterns.

Classical ensembles exhibit overconfidence in higher-probability bins, while standard TFMs are also substantially miscalibrated. TabPFN concentrates probability mass in high-confidence bins without corresponding empirical mortality, and TabICL departs from the diagonal across several bins. Because ECE is sensitive to prevalence, binning, and sample size, we interpret it comparatively rather than as an absolute calibration measure.

Together with Table I, these results highlight the central multi-objective finding: TabICL achieves the strongest IPCW C-index, yet its 2015 probability estimates remain poorly calibrated. Survival ranking and calibration therefore capture distinct deployment properties and should be audited jointly.

D. Survival Analysis and IPCW Correction

Table IV reports the 2-year landmark survival analysis for Drift-Resilient TabPFN. IPCW correction is estimable

TABLE I

SUMMARY OF LONGITUDINAL CLASSIFICATION AND SURVIVAL-RANKING PERFORMANCE (2007–2015). ANNUAL METRICS ARE MACRO-AVERAGED USING THE FIXED $W = 10$ WINDOW. LOWER ECE IS BETTER. IPCW C-INDEX USES A 2-YEAR LANDMARK HORIZON; 2014–2015 ARE EXCLUDED BECAUSE THE REQUIRED SURVIVAL ESTIMATES ARE NOT IDENTIFIABLE UNDER EXTREME CENSORING AND ABSENCE OF RECORDED EVENTS (SECTION IV-D). PATIENT-LEVEL BOOTSTRAP 95% CONFIDENCE INTERVALS FOR F1-MACRO ARE REPORTED IN THE SUPPLEMENTARY MATERIAL.

Model	**Annual Classification (avg. 2007–2015)**				**Survival Ranking**	
	Avg. $F1_{pos}$	Avg. PR-AUC	Avg. ECE	2015 $F1_{pos}$	C-index	IPCW C-index
XGBoost	0.238	0.164	0.180	0.164	0.527	0.445
LightGBM	0.247	0.173	0.177	0.178	0.536	0.425
Random Forest	0.197	0.135	**0.089**	0.109	0.516	0.377
TabPFN	0.221	0.155	0.215	0.103	0.521	0.420
TabICL	0.222	0.165	0.263	0.109	**0.586**	**0.542**
Drift-Resilient TabPFN	**0.316**	**0.249**	0.240	**0.212**	0.575	0.433

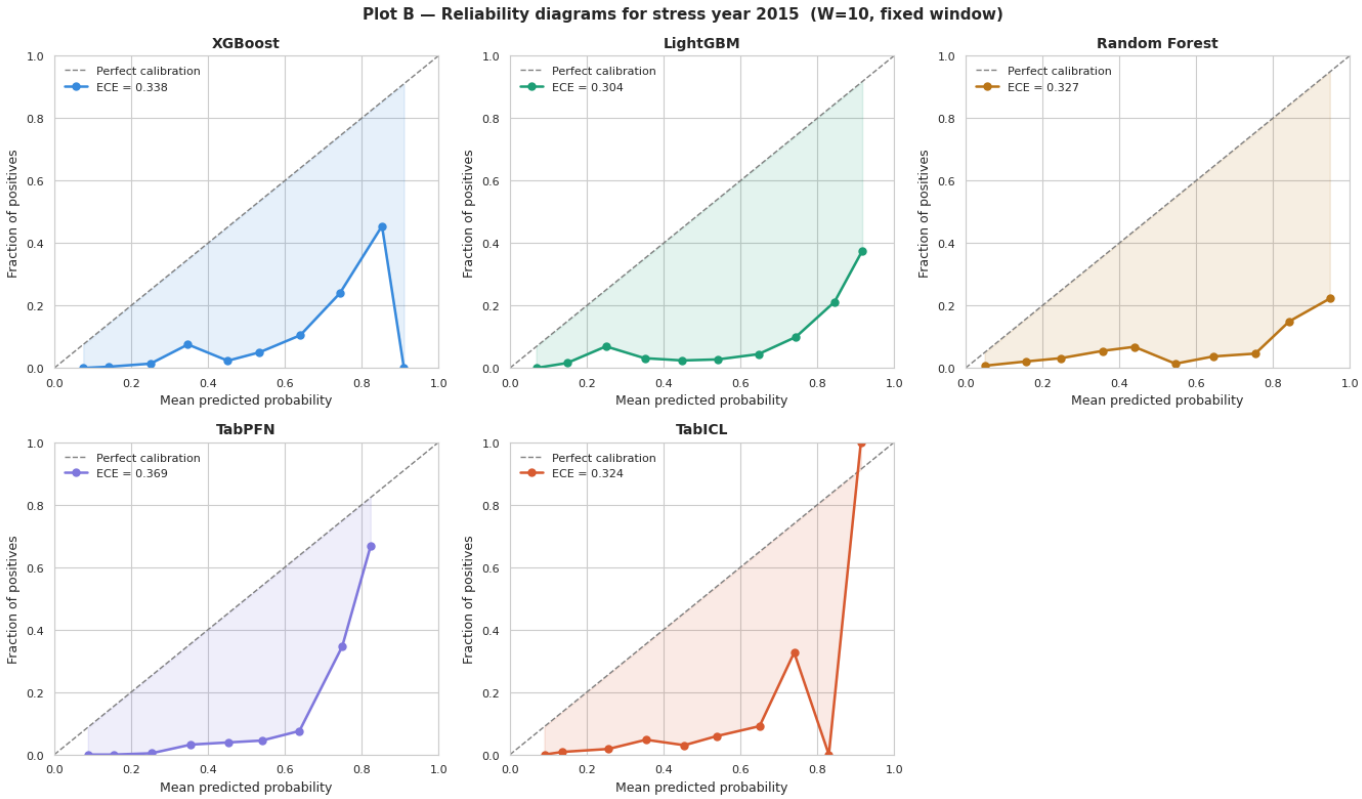

Fig. 2. Reliability diagrams for the 2015 stress year ($W = 10$, fixed window). All models exhibit calibration degradation, with standard TFMs showing pronounced departures from the diagonal despite competitive ranking performance. ECE values are annotated in each panel.

TABLE II

TEMPORAL ROBUSTNESS INDEX (TRI) ACROSS SHIFT REGIMES. HIGHER VALUES INDICATE GREATER RETENTION OF WITHIN-MODEL PERFORMANCE RELATIVE TO MEASURED DISTRIBUTIONAL CHANGE. TRI IS NOT USED AS A CALIBRATED CROSS-MODEL RANKING STATISTIC. $\varepsilon = 10^{-4}$.

Model	**Stable (2007)**	**Covariate Shift (2012)**	**Structural Shift (2015)**
XGBoost	**1262.1**	601.5	210.4
TabICL	**1262.1**	620.3	205.8
TabPFN	**1262.1**	743.7	294.6
Drift-Resilient TabPFN	1254.3	**770.3**	**410.2**

TABLE III

MEAN PR-AUC STRATIFIED BY EXPLORATORY DRIFT REGIME ACROSS ALL WINDOW CONFIGURATIONS ($W \in \{3, 5, 7, 10\}$, FIXED AND EXPANDING MODES). **STABLE**: LOW JS DIVERGENCE AND RELATIVELY STABLE PREVALENCE. **TYPE I**: 2012 ADMINISTRATIVE COVARIATE SHIFT. **TYPE II**: 2015 STRUCTURAL/OBSERVATION-PROCESS SHIFT.

Model	**Stable**	**Type I (Covariate)**	**Type II (Structural)**
XGBoost	0.198	0.208	**0.138**
LightGBM	0.179	0.184	0.101
Random Forest	0.159	0.169	0.084
TabPFN	0.196	**0.209**	0.131
TabICL	**0.209**	0.206	0.129

for 2007–2013, but not for 2014, when all observations are censored within the horizon, or for 2015, when no deaths are recorded after administrative truncation. These failures reflect limitations of the observed follow-up process and further illustrate the severity of the late-period observation-process shift.

TABLE IV
LANDMARK SURVIVAL ANALYSIS FOR DRIFT-RESILIENT TABPFN USING A 2-YEAR HORIZON. IPCW C-INDEX IS NOT ESTIMABLE FOR 2014 BECAUSE ALL OBSERVATIONS ARE CENSORED, OR FOR 2015 BECAUSE NO EVENTS ARE RECORDED WITHIN THE HORIZON.

Landmark	N	Events	Censored	C-index	IPCW C-index
2007	1,908	240	1,668	0.555	0.544
2008	2,409	284	2,125	0.550	0.565
2009	2,522	307	2,215	0.506	0.547
2010	2,619	311	2,308	0.503	0.511
2011	2,536	241	2,295	0.563	0.547
2012	2,553	256	2,297	0.594	0.597
2013	2,596	218	2,378	**0.627**	**0.628**
2014	2,615	71	2,544	0.592	– (all censored)
2015	2,310	0	2,310	–	– (no events)

Across estimable years, IPCW C-index ranges from 0.511 to 0.628. The mean absolute difference between corrected and uncorrected C-index is 0.017, indicating that censoring correction changes ranking estimates only modestly before the extreme late-period truncation. Censoring, therefore, contributes to instability but does not fully account for the divergence between annual classification and survival ranking observed over 2007–2013.

E. Diagnosing Proxy Reliance through Medication Perturbation

To complement the SHAP analysis, we perturb medication records in the 2015 stress year. For each model trained with the fixed $W = 10$ window, the 11 `c2MED` features are set to zero for increasing fractions of test patients. This intervention does not estimate the causal effect of medication exposure; rather, it evaluates whether predictions depend on medication-recording patterns that became unreliable after administrative truncation.

Figure 3 shows that all five standard models improve as medication features are progressively removed. XGBoost F1-pos rises from 0.149 to 0.181, LightGBM from 0.116 to 0.150, TabPFN from 0.141 to 0.201, and TabICL from 0.166 to 0.186. This monotonic trend suggests that medication-recording associations learned from earlier years became less reliable under the 2015 observation-process shift and that removing these features partially restores classification performance.

This observation is consistent with the SHAP analysis in Section V, which identifies medication variables among the most influential predictors for TabICL and standard TabPFN. Table V summarizes representative perturbation levels from the full sweep shown in Figure 3. Across all five standard models, positive-class F1 increases as medication records are progressively removed. In contrast, Drift-Resilient TabPFN exhibits a modest decline from 0.279 to 0.269 under the available 50% perturbation experiment. Although the Drift-Resilient evaluation was conducted separately and is therefore not directly comparable to the complete sweep, the opposite response direction is consistent with a qualitatively different dependence on medication-recording patterns.

TABLE V
MEDICATION-RECORD PERTURBATION AT THE 2015 STRESS YEAR. VALUES REPORT POSITIVE-CLASS F1. STANDARD MODELS WERE EVALUATED ACROSS THE FULL PERTURBATION SWEEP (FIGURE 3); REPRESENTATIVE PERTURBATION LEVELS ARE SHOWN HERE. DRIFT-RESILIENT TABPFN WAS EVALUATED SEPARATELY AT THE AVAILABLE 50% PERTURBATION LEVEL.

Model	0%	50%	100%
XGBoost	0.149	0.166	0.181
LightGBM	0.116	0.137	0.150
Random Forest	0.076	0.083	0.100
TabPFN	0.141	0.168	**0.201**
TabICL	0.166	**0.183**	0.186
Drift-Resilient TabPFN	**0.279**	0.269	*not evaluated*

F. Robustness Check: One-Year-Ahead Stream Protocol

We compare the one-year-ahead stream protocol with the $W = 10$ expanding-window evaluation to test whether conclusions depend on the update policy. Figure 4 shows that the two trajectories are closely aligned after 2007.

Three patterns are consistent across protocols: stream and expanding-window trajectories are nearly indistinguishable after 2007, performance peaks around 2008–2009 and declines toward 2015, and the aggregate results in Table VI do not materially change model ordering. These findings reduce the likelihood that the 2015 degradation is an artifact of a frozen training period or window specification.

TABLE VI
ONE-YEAR-AHEAD STREAM RESULTS AVERAGED OVER 2007–2015. VALUES ARE BROADLY CONSISTENT WITH THE EXPANDING-WINDOW ANALYSIS, INDICATING THAT CONTINUOUS ASSIMILATION OF HISTORICAL DATA DOES NOT MATERIALLY CHANGE THE MAIN DEGRADATION PATTERNS.

Model	Avg. F1-macro	Avg. F1-pos	Avg. PR-AUC	Avg. ECE
XGBoost	0.577	0.256	**0.202**	0.236
TabPFN	0.589	0.263	**0.202**	0.205
TabICL	**0.592**	**0.273**	0.195	0.207
LightGBM	0.565	0.241	0.175	**0.189**
Random Forest	0.550	0.225	0.159	0.203

G. Practical Implications: Drift Regime and Objective Shape Model Selection

The results suggest that model selection under temporal shift should jointly consider the anticipated drift regime and the deployment objective. Under stable conditions and the 2012 Type I regime, TFMs and gradient boosting methods achieve broadly comparable discrimination. Under the 2015 Type II structural/observation-process regime, sliding-window gradient boosting is comparatively resilient among standard models, while Drift-Resilient TabPFN achieves the strongest annual discrimination overall. For longitudinal risk ranking,

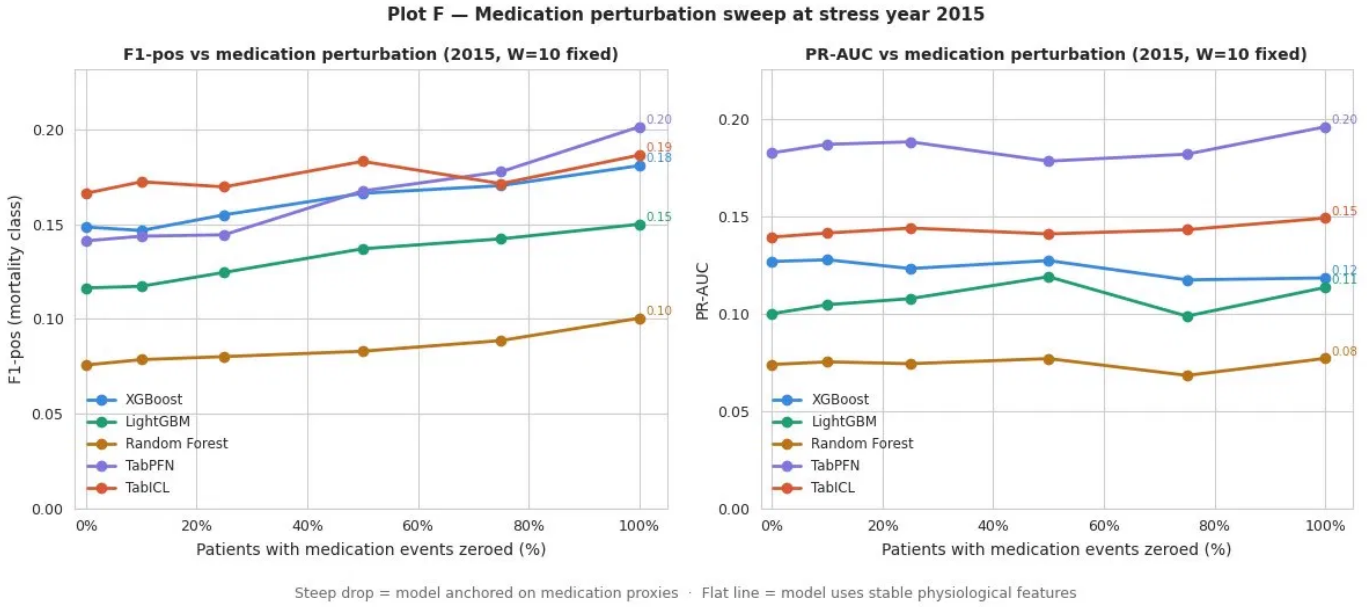

Fig. 3. Medication-record perturbation at the 2015 stress year. Left: positive-class F1 as a function of the fraction of patients with all `c2MED` features zeroed. Right: PR-AUC under the same perturbation. Standard models improve as medication information is removed, whereas the separately evaluated Drift-Resilient model declines, suggesting different dependence on medication-recording patterns.

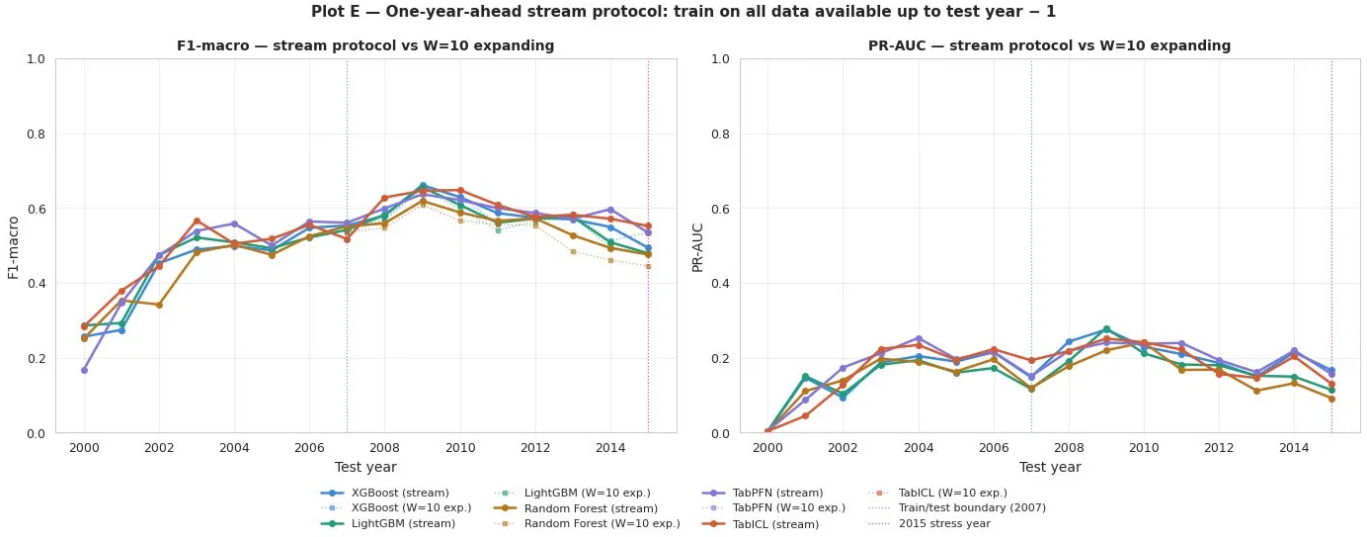

Fig. 4. One-year-ahead stream protocol (solid lines) versus $W = 10$ expanding window (dotted lines). Left: F1-macro; right: PR-AUC. The train/test boundary and the 2015 stress year are marked. Similar post-2007 trajectories indicate that the observed degradation is not specific to a single temporal update strategy.

TabICL provides the highest IPCW C-index, but this does not imply reliable calibration or same-year classification.

Taken together, these findings support the paper's central claim that robustness is not monolithic. Model superiority depends on the evaluated property, temporal regime, and operational decision. Multi-objective and shift-aware auditing is therefore necessary to avoid selecting models that appear robust under aggregate evaluation but fail on clinically relevant dimensions.

V. Why Models Fail: Domain-Aware Feature Attribution

To investigate the mechanisms behind the performance and calibration patterns observed under temporal shift, we apply a domain-aware XAI pipeline to audit model-level feature attributions. We extract global SHAP values across the longitudinal test period (2007–2015) and interpret them alongside the shift-regime and perturbation analyses reported in Section IV. These analyses are exploratory: they do not identify causal clinical effects, but help assess whether models rely on temporally sta-

ble physiological signals or on proxies that become unreliable as administrative and documentation practices change.

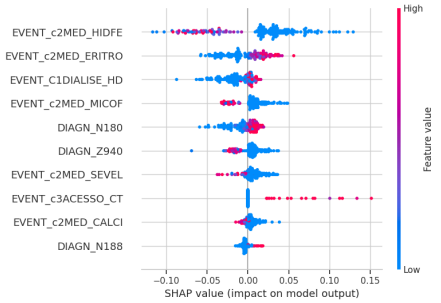

Fig. 5. TabICL assigns high importance to medication-record variables such as Iron and Sevelamer.

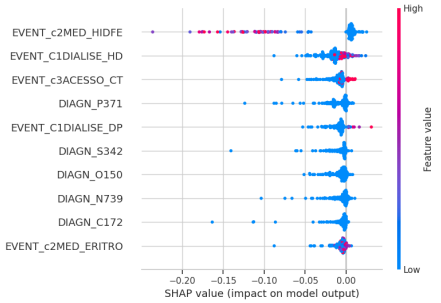

Fig. 6. Standard TabPFN distributes importance across a broader set of background variables.

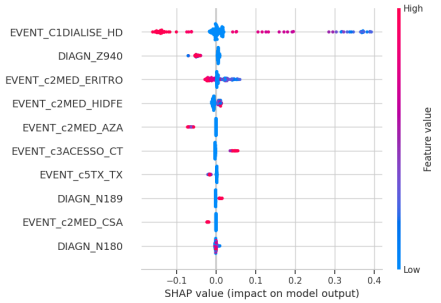

Fig. 7. Drift-Resilient TabPFN emphasizes core dialysis-related signals.

A comparative inspection of SHAP beeswarm plots reveals three attribution patterns.

a) TabICL and Standard TFMs.: TabICL assigns substantial importance to medication-related variables, notably Iron supplementation and Sevelamer. These features can be clinically meaningful, but they are also closely tied to prescription protocols, reimbursement rules, and documentation practices. As a result, they may behave as temporally fragile proxies when administrative recording changes. This interpretation is consistent with the 2015 calibration degradation and with the medication-record perturbation experiment, in which removing `c2MED` features improves the performance of standard models. Standard TabPFN shows a different pattern: importance is more diffusely spread across a long tail of background variables. We interpret this *signal diffusion* as a potential source of instability, since small temporal changes across many weak predictors may collectively affect decision boundaries under shift.

b) Drift-Resilient TabPFN.: In contrast, the temporally conditioned model places greater weight on core dialysis-related features, especially *EVENT_C1DIALISE_HD*. This pattern suggests that explicit temporal conditioning may reduce reliance on medication-record proxies and favor more persistent markers of disease severity. Clinically, hemodialysis has a dual interpretation: repeated sessions reflect ongoing treatment and care access, while dependence on renal replacement therapy also indicates severe baseline renal disease. By anchoring more strongly on such stable physiological signals, Drift-Resilient TabPFN provides a plausible explanation for its stronger annual discrimination under the 2015 stress regime.

A. Clinical Coherence of Learned Representations

The SHAP interaction analysis further suggests that the temporally conditioned model captures clinically plausible relationships. Higher dialysis frequency and erythropoietin administration appear as protective or stabilizing factors, consistent with treatment adherence and anemia management. Kidney transplantation is associated with extended survival, reflecting restoration of renal endocrine and metabolic functions. Conversely, dialysis catheter use is associated with a higher risk, consistent with complications related to central venous access, including infection and cardiovascular burden.

These patterns should be interpreted as evidence of clinical coherence rather than causal effect estimation. Their value lies in showing that the drift-aware model tends to emphasize variables whose meanings are more stable across time, whereas standard models place greater weight on medication-record patterns that become unreliable under the 2015 observation-process shift. The perturbation experiment in Section IV-E provides a complementary interventional stress test of this proxy-reliance hypothesis.

B. Predictive Uncertainty Under Shift

We also compute predictive entropy over the test period as an exploratory proxy for model uncertainty. For Drift-Resilient TabPFN, entropy increases toward 2015, suggesting that the model partially represents the late-period distributional change as increased uncertainty rather than only as overconfident error. This behavior is desirable in deployment settings because rising uncertainty may indicate cases or periods that require human review, recalibration, or model updates.

This uncertainty pattern is consistent with the broader diagnostic evidence: as the observation process becomes increasingly misaligned with historical training windows, the model retains comparatively stronger annual discrimination, but signals reduced confidence. We therefore interpret entropy not as a complete uncertainty quantification method, but as a deployment-relevant warning signal that can complement calibration, drift monitoring, and feature-attribution diagnostics under severe temporal shift.

VI. Conclusion and Future Work

As Tabular Foundation Models move toward real-world deployment, evaluating robustness under dataset shift becomes essential, especially in clinical settings where protocols, documentation practices, and patient populations evolve over time. We presented a 19-year longitudinal evaluation of classical ensembles, standard TFMs, and a drift-aware TFM on a national renal registry, using uniform temporal protocols and a multi-objective auditing framework spanning discrimination, calibration, and survival ranking.

Our results show that robustness is not monolithic. Models that perform well in global risk ranking may still fail in annual discrimination or probability calibration, and model rankings vary across stable, covariate-shift, and structural/observation-process-shift regimes. The analyses further suggest that standard models are vulnerable when they rely on medication-record proxies whose meaning changes over time, whereas temporal conditioning can encourage reliance on more stable physiological signals. These findings highlight the need to evaluate tabular models not only by aggregate accuracy, but also by how they behave across objectives, temporal regimes, and clinically meaningful failure modes.

Several limitations remain. The proposed drift taxonomy and TRI are exploratory and should be validated on additional longitudinal datasets. Our evidence comes from a single national registry and should be extended to multi-center health records and non-clinical domains. Future work should also integrate calibration-aware training, uncertainty-guided abstention, and adaptive updating strategies informed by shift-regime diagnostics. Overall, our study provides a practical framework for auditing foundation models under long-horizon dataset shift and for guiding safer deployment in dynamic tabular environments.

References

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.

[2] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, and P. W. Koh, "A unifying view on dataset shift in medical AI," *Nature Communications*, vol. 12, no. 1, p. 2776, 2021.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[6] J. Qu, D. Holzmüller, G. Varoquaux, and M. L. Morvan, "TabICL: A tabular foundation model for in-context learning on large data," *arXiv preprint arXiv:2502.05564*, 2025.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] A. Tsymbal, "The problem of concept drift: Definitions and related work," Department of Computer Science, Trinity College Dublin, Tech. Rep., 2004.

[9] J. Jiang, S. Liu, H. Cai, Q. Zhou, and H.-J. Ye, "Representation learning for tabular data: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[10] K. Helli, D. Schnurr, N. Hollmann, S. Müller, and F. Hutter, "Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 98 742–98 781.

[11] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] A. A. Guerra Junior, R. G. Pereira, E. I. Gurgel, M. Leal Cherchiglia, L. V. Dias, J. D. Ávila, N. Santos, A. Reis, F. de Assis Acurcio, and W. Meira Junior, "Building the national database of health centred on the individual: Administrative and epidemiological record linkage—brazil, 2000–2015," *International Journal of Population Data Science*, vol. 3, no. 1, p. 446, 2018.