# The Crisis of Generalization in Medical AI: Bridging the Gap in Pulmonary Nodule Malignancy Prediction via Tabular Foundation Models and Unsupervised Domain Adaptation

## 1. Introduction

The clinical management of solitary pulmonary nodules (SPNs) represents one of the most prevalent and high-stakes diagnostic challenges in modern pulmonology and thoracic radiology. With the widespread adoption of low-dose computed tomography (LDCT) for lung cancer screening, the incidental detection of SPNs has surged, creating a massive triage burden for healthcare systems globally. The fundamental clinical question—distinguishing benign granulomas or hamartomas from early-stage malignant adenocarcinomas—requires a delicate balance between sensitivity, to prevent missed diagnoses of potentially curable cancers, and specificity, to avoid unnecessary invasive procedures such as biopsies or resections on benign lesions. For decades, this decision-making process has been augmented by clinical risk prediction models, statistical tools designed to estimate the probability of malignancy based on radiographic and demographic variables. However, a pervasive and systemic failure mode has emerged as these tools have migrated from research environments to diverse clinical settings: the failure of generalization.

The central thesis of this report posits that the traditional paradigm of static risk modeling—whether based on logistic regression or conventional deep learning—is fundamentally ill-equipped to handle the distributional heterogeneity inherent in multi-institutional healthcare data. We argue that the future of robust clinical decision support lies in the convergence of two emerging computational methodologies: Tabular Foundation Models (TFMs), which leverage large-scale pre-training to handle small-sample sparsity, and Unsupervised Domain Adaptation (UDA), which provides the mathematical machinery to align divergent feature distributions across hospital systems.

Current standards of care, such as the Mayo Clinic model, the Veterans Affairs (VA) model, and the Brock (PanCan) University model, have served as the bedrock of nodule stratification for years. Yet, extensive empirical evidence accumulated between 2013 and 2025 demonstrates that these models exhibit significant performance degradation when applied outside their derivation cohorts. This phenomenon is not merely a statistical nuance but a clinical hazard; a model that predicts malignancy with an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.85 in a North American veteran population may plummet to an AUC of 0.60 when applied to an East Asian non-smoking cohort.1 Such precipitous drops render these tools unsafe for deployment, leading to either over-investigation of benign nodules or, more catastrophically, the false reassurance of patients with malignancy.

The friction resisting effective cross-hospital deployment is tripartite. First, the "Small Sample Problem" plagues medical tabular data. Unlike imaging datasets which may contain millions of pixels, structured clinical cohorts often consist of only a few hundred confirmed cases, insufficient for training high-capacity deep neural networks from scratch without severe overfitting. Second, "Distributional Shift" occurs as patient demographics, disease prevalence, and scanner protocols vary across sites, fundamentally altering the statistical relationships between features and outcomes. Third, "Feature Heterogeneity" presents a uniquely tabular challenge; unlike images that share a fixed pixel grid, clinical databases are often fragmented, with different institutions recording disjoint sets of biomarkers or clinical variables.3

In response, the field is witnessing a paradigm shift toward Tabular Foundation Models, exemplified by architectures like the Tabular Prior-Data Fitted Network (TabPFN). By pre-training on millions of synthetic datasets generated from structural causal models, these transformers learn to approximate Bayesian inference, enabling them to generalize from small datasets with remarkable efficiency.5 However, while TFMs solve the sample scarcity issue, they remain vulnerable to domain shift. This report extensively analyzes the PANDA (Pretrained Adaptation Network with Domain Alignment) framework as a pivotal case study. PANDA demonstrates how coupling a foundation model with unsupervised alignment techniques like Transfer Component Analysis (TCA) and Cross-Domain Recursive Feature Elimination (RFE) can stabilize performance across institutions, turning a fragile predictive engine into a robust clinical tool.3

This comprehensive analysis will dissect the anatomy of generalization failure in pulmonary nodule prediction, explore the theoretical and practical mechanics of tabular foundation models, and evaluate the role of unsupervised domain adaptation in creating the next generation of adaptive medical AI. Through this lens, we aim to chart a path away from brittle, site-specific models toward dynamic systems capable of navigating the complex, heterogeneous landscape of global healthcare.

## 2. The Landscape of Pulmonary Nodule Malignancy Prediction (2013–2025)

The evolution of malignancy prediction models tracks the broader trajectory of medical statistics and artificial intelligence, moving from simple linear models to complex non-linear estimators. Despite increasing sophistication, the challenge of external validity remains the Achilles' heel of the field.

### 2.1 The Decay of Classical Clinical Risk Models

Clinical risk calculators are the most widely validated tools in current practice. They are typically derived using multivariable logistic regression on specific cohorts, identifying independent predictors of malignancy. While robust in their training environments, their rigidity becomes a liability in external settings.

#### 2.1.1 The Mayo Clinic Model: A Study in Demographic Mismatch

The Mayo Clinic model, developed in 1997, remains one of the most cited risk calculators. It estimates malignancy probability based on age, smoking history, history of extrathoracic cancer, nodule diameter, spiculation, and upper lobe location.

* **Derivation Context:** The model was derived from a cohort of roughly 600 patients at the Mayo Clinic, a tertiary referral center in the United States. The population was predominantly white, with a high prevalence of smoking-related malignancies and a significant baseline rate of granulomatous disease (e.g., histoplasmosis).
* **External Validation Performance:** Contemporary validation studies reveal severe limitations. In Asian populations, where lung cancer epidemiology differs significantly (e.g., higher rates of EGFR-driven adenocarcinomas in non-smokers), the Mayo model's reliance on smoking history as a heavy weighting factor leads to systematic underestimation of risk. Research indicates that in Chinese validation cohorts, the Mayo model’s AUC frequently drops to the range of **0.58 to 0.68**, a level of performance that is marginally better than chance and clinically actionable only with extreme caution.3
* **Mechanism of Failure:** The failure is rooted in "Concept Shift" ($P(Y|X)$ varies). The conditional probability of malignancy given a "non-smoking" status is fundamentally different in East Asian females compared to the Mayo derivation cohort. Furthermore, the prevalence of benign mimics differs; regions with high tuberculosis burdens present benign nodules that morphologically mimic malignancy (e.g., spiculation), confounding the specific morphological features the Mayo model relies upon.8

#### 2.1.2 The Brock (PanCan) Model: Screening vs. Clinical Cohorts

The Brock model, derived from the Pan-Canadian Early Detection of Lung Cancer Study (McWilliams et al., 2013), was designed specifically for screening populations—individuals at high risk who are undergoing surveillance. It incorporates nodule count, spiculation, and specific location parameters.

* **Performance Variability:** While the Brock model generally outperforms the Mayo model in screening settings (AUC ~0.90 in derivation), its transfer to clinical cohorts (patients presenting with symptoms or incidental findings) is fraught with instability. A 2024 systematic review and meta-analysis encompassing 52 studies and over 85,000 patients reported a pooled AUC of **0.796**. However, this aggregate figure obscures significant subgroup failures. In Asian populations, the pooled AUC dropped to **0.741**, and for subsolid nodules—a critical subtype often representing slow-growing adenocarcinomas—the AUC fell to **0.747**.2
* **Nodule Size and Type Sensitivity:** The model shows marked performance degradation for larger nodules (>15mm) and solitary nodules compared to multiple nodules. In direct head-to-head comparisons on Chinese clinical data, the Brock model has recorded AUCs as low as **0.612**, statistically indistinguishable from the failing Mayo model in the same setting.1 This suggests that the "screening" prior (high pre-test probability of early-stage disease) does not translate well to "clinical" populations where the spectrum of disease is broader and includes more advanced or inflammatory conditions.

#### 2.1.3 The PKUPH Model: Local Optimization and Regional Limits

Recognizing the failures of Western models, researchers developed the Peking University People’s Hospital (PKUPH) model, tailored specifically for the Chinese demographic.

* **Internal vs. External Validity:** While the PKUPH model demonstrates superior internal validity (AUC > 0.85) within its development institution, it is not immune to the generalization paradox. External validation studies within China show a drop in performance, albeit less severe than Western models. In comparative analyses, the PKUPH model achieved an external AUC of **0.670** in some cohorts, and **0.636** in others.1
* **Implication:** The fact that a regionally optimized model still struggles to generalize across hospitals within the same country highlights that "nationality" is a crude proxy for domain. True domain shift is driven by more granular factors: scanner calibration, radiologist measurement variability (inter-observer variability in measuring diameter or spiculation), and local disease endemicity.3

### 2.2 The False Dawn of Deep Learning and Radiomics

The promise of Artificial Intelligence, specifically Convolutional Neural Networks (CNNs) and Radiomics (high-throughput feature extraction), was to bypass the limitations of handcrafted features by learning data-driven representations.

#### 2.2.1 Radiomics: The Reproducibility Crisis

Radiomics involves extracting hundreds or thousands of quantitative features (e.g., Gray Level Co-occurrence Matrix textures, wavelet transforms) from segmented nodule images.

* **Scanner Dependency:** Radiomic features are notoriously sensitive to image acquisition parameters. Variations in slice thickness (1mm vs. 5mm), reconstruction kernels (sharp vs. soft), and tube current (mA) act as non-biological noise that distorts feature values. A texture feature indicating "heterogeneity" on a sharp kernel scan may simply reflect image noise rather than tumor necrosis.
* **The Need for Harmonization:** Studies utilizing the Lung-RADS protocol have shown that without statistical harmonization techniques like ComBat (which adjusts for batch effects), radiomics models can lose **0.10–0.15 AUC points** when applied to external data.9 The model essentially learns to identify the "scanner" rather than the "cancer," leading to high internal accuracy but catastrophic external failure.

#### 2.2.2 Deep Learning (CNNs): Black Box Fragility

Deep learning models, such as 3D-CNNs (e.g., C3D, Lung-Net), ingest volumetric CT data directly.

* **Internal Dominance:** In large, standardized datasets like the LIDC-IDRI or NLST, DL models consistently achieve AUCs exceeding **0.92–0.96**, significantly outperforming clinical risk calculators and often matching expert radiologists.10
* **External Vulnerability:** Despite this, DL models are prone to "shortcut learning." They may latch onto confounding artifacts—such as a specific text marker on a scan or the field-of-view padding—that correlate with malignancy in the training set but not in the real world. External validation studies have reported significant drops; for instance, a DL model that achieved 93% sensitivity in development dropped by **14.5%** in sensitivity when tested on a diverse external clinical dataset.10 This fragility is exacerbated by the "Small Sample" nature of many proprietary hospital datasets, where deep models memorize the training set rather than learning robust biological features.

### 2.3 Performance Synthesis: The Generalization Gap

The following table summarizes the typical degradation observed in peer-reviewed literature when models are moved from development to external validation.

| **Model Family** | **Development AUC (Typical)** | **External AUC (Typical Range)** | **Primary Mechanisms of Failure** |
| --- | --- | --- | --- |
| **Mayo Clinic** | 0.80 – 0.85 | 0.58 – 0.68 | **Concept Shift:** Mismatch in biological risk factors (smoking vs. genetic) and benign disease prevalence (fungal vs. TB). |
| **Brock (PanCan)** | 0.90 – 0.94 | 0.74 – 0.79 | **Population Shift:** Screening cohort priors do not match clinical/incidental findings; sensitivity to nodule size. |
| **PKUPH** | 0.85 – 0.88 | 0.63 – 0.67 | **Feature Noise:** Sensitivity to measurement variability and local practice patterns even within same region. |
| **Deep Learning** | 0.92 – 0.96 | 0.80 – 0.86 | **Covariate Shift:** Extreme sensitivity to CT scanner physics, reconstruction kernels, and acquisition protocols. |

## 3. The Taxonomy of Distribution Shift in Structured Medical Data

To engineer solutions for generalization, we must rigorously define the problem. In the context of structured medical data (tabular EHR and radiologic measurements), distribution shift manifests differently than in unstructured perceptual tasks like computer vision.

### 3.1 Covariate Shift vs. Concept Shift

The joint distribution of features $X$ and labels $Y$, denoted as $P(X, Y)$, can change in two primary ways between a source domain ($S$) and a target domain ($T$).

#### 3.1.1 Covariate Shift

Covariate shift occurs when the marginal distribution of the input features changes ($P\_S(X) \neq P\_T(X)$), but the conditional distribution of the labels remains constant ($P\_S(Y|X) = P\_T(Y|X)$).

* **Clinical Example:** Consider nodule size. A specialized cancer center (Source) might see a distribution of nodule diameters skewed toward larger, late-stage lesions ($mean = 20mm$). A community hospital (Target) might see smaller, incidental nodules ($mean = 8mm$). If a model is trained primarily on large nodules, it may be poorly calibrated for the smaller nodules in the target domain, even if the biological relationship between size and malignancy holds.
* **Impact on Foundation Models:** Tabular Foundation Models like TabPFN rely on "priors" formed during in-context learning. If the query samples from the target domain fall into regions of the feature space that are sparsely populated in the source context, the model's uncertainty estimates and predictions become unreliable.3

#### 3.1.2 Concept Shift

Concept shift occurs when the relationship between features and labels changes ($P\_S(Y|X) \neq P\_T(Y|X)$). This is arguably the most pernicious form of shift in medicine.

* **Clinical Example:** The "Spiculation" feature. In North American cohorts (Source), spiculation is a strong predictor of malignancy (invasion). In regions with endemic tuberculosis (Target), benign tuberculomas often present with spiculation due to fibrosis. Thus, the probability of malignancy given spiculation, $P(Y=Malignant | Spiculation=High)$, is significantly lower in the Target domain. A model that learns the Source relationship will systematically over-predict malignancy in the Target domain.

### 3.2 The Unique Challenge of Feature Heterogeneity

Unlike computer vision, where images universally share a pixel grid structure (e.g., $224 \times 224 \times 3$), tabular clinical data is characterized by **Feature Heterogeneity**.

* **The "Missing Column" Problem:** It is rare for two hospitals to collect identical sets of variables. Hospital A might use a specific tumor marker panel (CEA, NSE, CYFRA21-1), while Hospital B relies solely on radiologic measurements and basic blood counts.
* **Mathematical Implication:** Let $\mathcal{F}\_S$ be the feature space of the source and $\mathcal{F}\_T$ be the feature space of the target. In most transfer learning scenarios, we assume $\mathcal{F}\_S = \mathcal{F}\_T$. In cross-institutional healthcare, typically $\mathcal{F}\_S \cap \mathcal{F}\_T \subset \mathcal{F}\_S \cup \mathcal{F}\_T$. This forces a reduction in dimensionality to the intersection of features, potentially discarding high-value predictors available only in the source.
* **The PANDA Solution:** This heterogeneity necessitates advanced strategies like Cross-Domain Recursive Feature Elimination (RFE) to not only select *predictive* features but to select *stable* features that exist and function similarly across domains.3

### 3.3 Theoretical Bounds on Transferability

Statistical learning theory provides bounds on the error in the target domain based on the error in the source domain and the divergence between them. A classic bound (Ben-David et al.) states:

$$\epsilon\_T(h) \leq \epsilon\_S(h) + \frac{1}{2} d\_{\mathcal{H}\Delta\mathcal{H}}(D\_S, D\_T) + \lambda$$

Where:

* $\epsilon\_T(h)$ is the error of hypothesis $h$ on the target.
* $\epsilon\_S(h)$ is the error on the source.
* $d\_{\mathcal{H}\Delta\mathcal{H}}(D\_S, D\_T)$ is the $\mathcal{H}$-divergence between the source and target distributions.
* $\lambda$ is the error of the ideal joint hypothesis.

This inequality dictates that minimizing source error alone (standard supervised learning) is insufficient. One must also minimize the divergence term $d\_{\mathcal{H}\Delta\mathcal{H}}$. This is the theoretical justification for **Domain Adaptation**—algorithms designed explicitly to reduce the distance between $D\_S$ and $D\_T$ in a latent feature space.11 Without this alignment, even a "perfect" source model is mathematically guaranteed to have a high upper bound on target error if the domains are divergent.

## 4. The Rise of Tabular Foundation Models: Solving Sample Scarcity

While Gradient-Boosted Decision Trees (GBDTs) like XGBoost and LightGBM have long dominated tabular data competitions, they struggle in the "small data" regime typical of medical research ($N < 1000$). Deep learning alternatives like TabNet and FT-Transformer attempted to bridge this gap but often succumb to overfitting or require extensive hyperparameter tuning that is prone to data leakage. The years 2023–2025 have seen the emergence of Tabular Foundation Models (TFMs), offering a radically different approach.

### 4.1 TabPFN: The Mechanics of In-Context Learning

The Tabular Prior-Data Fitted Network (TabPFN) represents a breakthrough in small-sample tabular learning. Unlike traditional models that update weights via gradient descent to fit a specific training dataset, TabPFN is a transformer pre-trained to perform inference as a forward pass.

#### 4.1.1 Synthetic Priors and Meta-Learning

TabPFN is trained offline on millions of **synthetic datasets** generated from structural causal models (SCMs) and other diverse functional forms (Bayesian neural networks, Gaussian processes).

* **The Prior:** The model effectively learns a "prior" over the space of possible tabular functions. It learns how features typically interact, how classes are separated, and how to handle outliers, not from real-world data, but from the abstract mathematical structure of the synthetic tasks.5
* **In-Context Inference:** At inference time, the user provides a context set (the small training dataset $D\_{train} = \{(x\_i, y\_i)\}$) and a query set ($x\_{query}$). The transformer processes this sequence—$[ (x\_1, y\_1),..., (x\_n, y\_n), x\_{query} ]$—using self-attention mechanisms. It effectively "looks up" the relationship between the query and the context samples to predict $y\_{query}$. This mimics a nearest-neighbor or kernel density estimation process but is learned and highly non-linear.

#### 4.1.2 Advantages for Clinical Data

1. **Small Sample Efficiency:** TabPFN excels in the regime of $N < 10,000$ samples. Since most single-institution nodule cohorts range from 200 to 2,000 patients, TabPFN operates in its optimal zone, often outperforming tuned XGBoost models which require more data to stabilize their tree splits.13
2. **No Hyperparameter Tuning:** TabPFN is a "zero-shot" or "one-pass" model. It does not require the tuning of learning rates, tree depths, or regularization terms. This is a massive safety advantage in clinical research, where improper cross-validation during tuning is a major source of reproducibility failure (the "optimism bias").14

### 4.2 Limitations and Evolving Architectures

Despite their promise, first-generation TFMs face significant hurdles in the context of cross-domain deployment.

#### 4.2.1 The Scalability Bottleneck

The original TabPFN relies on full attention, which scales quadratically $O(N^2)$ with the number of context samples. This limits its application to datasets smaller than roughly 10,000 rows. While sufficient for many specific nodule cohorts, it prevents the model from ingesting large-scale national registries or federated datasets.

* **TabPFN-2.5:** Recent iterations (2025) have introduced approximations and optimizations to scale this limit up to 50,000 samples, broadening the utility but not fully solving the "Big Data" problem of EHR mining.14

#### 4.2.2 The Closed-World Assumption

TabPFN fundamentally assumes that the query sample comes from the same distribution as the context samples. It has no internal mechanism to detect or correct for distribution shift. If provided with a context from Hospital A and a query from Hospital B, the self-attention mechanism will compute similarity scores based on the source distribution's geometry. If covariate shift has distorted the metric space (e.g., features are scaled differently due to scanner calibration), the attention weights will be incorrect, leading to poor predictions. This necessitates the use of external **Domain Adaptation** wrappers.17

#### 4.2.3 Drift-Resilient Architectures

Recognizing the fragility to temporal and domain shift, researchers have begun developing **Drift-Resilient TabPFN** variants. These models are pre-trained on synthetic data that explicitly incorporates "drift" parameters in the generating SCMs. By learning to recognize non-stationary distributions during pre-training, these models can theoretically adapt to shifts at inference time, though their application to complex medical shifts is still in nascent validation stages.19

## 5. Unsupervised Domain Adaptation (UDA) Methodologies

To deploy foundation models safely across hospitals, we must bridge the gap between the theoretical bounds of transfer learning and the practical reality of feature mismatch. Unsupervised Domain Adaptation (UDA) provides the algorithmic toolkit to achieve this alignment without requiring labels in the target domain.

### 5.1 Transfer Component Analysis (TCA)

TCA is a kernel-based dimensionality reduction method that is central to frameworks like PANDA. It addresses the assumption that the data in source and target domains lies on related but distinct manifolds.

* **Mechanism:** TCA seeks a projection matrix $W$ that maps input features $X$ into a lower-dimensional Reproducing Kernel Hilbert Space (RKHS). In this latent space, two objectives are optimized simultaneously:
  1. **Variance Preservation:** The intrinsic structure (variance) of the data should be preserved to maintain discriminatory power.
  2. **Distribution Alignment:** The Maximum Mean Discrepancy (MMD) between the projected source data $P(W^T X\_S)$ and projected target data $P(W^T X\_T)$ is minimized.
* Mathematical Formulation:  
    
  $$\min\_W \text{tr}(W^T K L K W) + \mu \text{tr}(W^T W)$$  
    
  Here, $K$ is the kernel matrix, $L$ is the MMD matrix (encoding the difference in means between domains), and $\mu$ is a regularization parameter.
* **Clinical Interpretation:** TCA effectively finds a "common language" for the features. It might learn that a "20mm" nodule on Scanner A is statistically equivalent to an "18mm" nodule on Scanner B, aligning them in the latent space so the classifier treats them identically.3

### 5.2 Cross-Domain Recursive Feature Elimination (RFE)

While TCA aligns distributions, it assumes a fixed set of features. Feature Heterogeneity requires a selection step.

* **Process:** Cross-Domain RFE involves training a model on the intersection of available features and iteratively removing those that contribute to domain instability. In the PANDA study, this process reduced an initial set of 63 features down to a core set of 8 highly stable biomarkers.
* **Why it matters:** Many features in EHRs are "noisy proxies." For example, "Hospital Duration" might correlate with severity in one hospital but with administrative inefficiency in another. RFE filters out these unstable proxies, leaving only the features with robust biological causality (e.g., Spiculation, Diameter, Age).3

### 5.3 Federation and Privacy: The Next Frontier

Traditional UDA requires pooling data to compute alignment matrices (like the Kernel matrix in TCA). In healthcare, privacy laws (HIPAA, GDPR) often prevent this.

* **FedFusion and FedDAFL:** Emerging frameworks like FedFusion extend DA to Federated Learning. They allow for "Diversity-Aware Encoding" where local sites train personalized encoders to handle their specific feature heterogeneity, while a central server aggregates a shared classifier. This allows for domain adaptation without raw data ever leaving the hospital firewalls.22

## 6. Case Study: The PANDA Framework

The PANDA (Pretrained Adaptation Network with Domain Alignment) framework 3 serves as the definitive proof-of-concept for integrating TFMs and UDA in pulmonary nodule analysis.

### 6.1 Experimental Design

* **Cohorts:** The study utilized a training cohort (Cohort A) of 295 patients from one Chinese institution and an external validation cohort (Cohort B) of 190 patients from a separate institution. This $N \approx 500$ total sample size is classic "small data" territory where deep learning typically fails.
* **The Pipeline:**
  1. **Preprocessing:** Multi-branch pipeline including feature rotation and quantile transformation to normalize inputs.
  2. **Feature Selection:** Cross-Domain RFE reduced the feature space to 8 stable variables.
  3. **Foundation Model:** A TabPFN backbone was used to generate embeddings.
  4. **Alignment:** TCA was applied to the TabPFN embeddings to align Cohort A and Cohort B in the latent space.
  5. **Ensemble:** Predictions were aggregated across multiple random seeds and preprocessing branches.

### 6.2 Results and Analysis

* **Baseline Collapse:** The Mayo Clinic model, when applied to Cohort B, achieved an AUC of **0.584**, indicating near-random performance. The PKUPH model achieved **0.636**.
* **PANDA Performance:** The full PANDA framework achieved an external AUC of **0.705**. While numerically this may appear modest compared to internal validation scores of >0.80, it represents a massive relative improvement over the clinical baselines.
* **Sensitivity (Recall):** Most critically, the TCA-adapted PANDA model achieved a Recall of **94.4%**. In a screening context, sensitivity is paramount; a tool that misses only 5% of cancers (PANDA) is clinically viable, whereas a tool that misses significantly more (unadapted models) is dangerous.
* **The Role of TCA:** An ablation study showed that PANDA without TCA achieved an AUC of 0.698. The addition of TCA pushed this to 0.705 and significantly improved calibration. This confirms that while the foundation model provides a strong baseline, explicit domain alignment is necessary to squeeze out the maximum generalization performance and ensure safety.3

### 6.3 Why PANDA Works: The Smoothness Hypothesis

Theoretical analysis suggests a synergistic effect between the Foundation Model and the UDA step. Foundation models, due to their pre-training on diverse functions, tend to learn "smooth" representations (Lipschitz continuous). Smooth representations are easier to align. A small shift in the input space results in a small, predictable shift in the TabPFN latent space, which linear methods like TCA can easily correct. In contrast, deep neural networks trained from scratch on small data often learn "jagged" decision boundaries that are chaotic and difficult to align.3

## 7. Benchmarking Robustness: TableShift and Beyond

To move beyond single-study validations, the field is increasingly adopting standardized benchmarks to quantify robustness.

### 7.1 TableShift: A Standard for Tabular Shift

TableShift is a benchmark suite introduced to evaluate tabular models under rigorous distribution shift conditions. It includes 15 binary classification tasks across healthcare, finance, and policy.

* **Key Finding:** The benchmark reveals a strong linear correlation between In-Distribution (ID) and Out-Of-Distribution (OOD) accuracy. This implies that "better" models (stronger priors, better architectures) are generally more robust. However, it also highlights that specialized robustness methods often trade off some ID accuracy to reduce the generalization gap.24

### 7.2 The BRFSS Diabetes Race-Shift Task

One of the most clinically relevant tasks in TableShift is the "Diabetes Prediction" task using the BRFSS dataset, where the shift is defined by patient race (e.g., Train on White, Test on Black/Hispanic).

* **Performance Drop:** Standard models show a baseline performance drop of approximately **4.5%** accuracy when shifted across racial groups.25
* **Implications:** This mirrors the cross-continental failure of the Mayo model. It confirms that models relying on socio-biological correlations (like BMI-to-Diabetes risk) are fragile when those correlations vary by subpopulation. It underscores the need for "fairness-aware" domain adaptation that ensures performance parity across demographic groups, a critical ethical requirement for medical AI deployment.

## 8. Conclusion and Future Outlook

The era of the "single-hospital, single-model" study is ending. The overwhelming evidence from the past decade confirms that static clinical risk models and naive deep learning approaches are fundamentally fragile in the face of the complex, heterogeneous reality of global healthcare. The external validation drops of the Mayo, Brock, and PKUPH models—often exceeding 0.20 AUC points—are not anomalies but expected consequences of Concept and Covariate Shift.

The integration of Tabular Foundation Models and Unsupervised Domain Adaptation offers a robust path forward. Frameworks like PANDA demonstrate that we can leverage the "universal priors" of models like TabPFN to solve the small-sample problem, while simultaneously employing rigorous alignment techniques like TCA to solve the distribution shift problem. This dual approach respects the unique constraints of medical data: its sparsity, its heterogeneity, and its high stakes.

**Actionable Recommendations for Future Research:**

1. **Adoption of UDA:** Clinical AI studies should effectively be required to demonstrate performance not just on a held-out test set, but on an external cohort *after* applying unsupervised domain adaptation. "Zero-shot" transfer should be replaced by "Adaptive Transfer" as the gold standard.
2. **Invest in Federated DA:** Privacy remains a barrier to pooling data for alignment. Federated Domain Adaptation methods (FedFusion, FedDAFL) that align distributions without sharing patient records are the critical infrastructure needed to scale these solutions.
3. **Multimodal Foundation Models:** The future lies in models that can jointly align tabular data and imaging data. Aligning the joint distribution $P(Image, Clinical)$ will require new classes of multimodal foundation models that can perform in-context learning across data types.
4. **Dynamic Regulation:** Regulatory bodies (FDA, EMA) must evolve frameworks to validate "adaptive" algorithms—software that is not fixed at the factory, but which includes a validated "calibration procedure" (like TCA) to safely adapt to the local population of the deploying hospital.

By embracing these adaptive paradigms, we can move closer to the goal of precision oncology: predictive tools that are as reliable in a community clinic in rural Asia as they are in a tertiary academic center in North America.

#### Works cited

1. Evaluation of models for predicting the probability of malignancy in patients with pulmonary nodules - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7048676/>
2. Pulmonary nodule malignancy probability: a meta-analysis of the Brock model - BINASSS, accessed November 25, 2025, <https://www.binasss.sa.cr/mar25/17.pdf>
3. AI\_Healthcare\_Analytics\_2025.pdf
4. Towards Data-Centric AI: A Comprehensive Survey of Traditional, Reinforcement, and Generative Approaches for Tabular Data Transformation - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2501.10555v1>
5. New AI model TabPFN enables faster and more accurate predictions on small tabular data sets - Universität Freiburg, accessed November 25, 2025, <https://uni-freiburg.de/en/new-ai-model-tabpfn-enables-faster-and-more-accurate-predictions-on-small-tabular-data-sets/>
6. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second - Table Representation Learning Workshop, accessed November 25, 2025, <https://table-representation-learning.github.io/assets/papers/tabpfn_a_transformer_that_solv.pdf>
7. Establishment and validation of multiclassification prediction models for pulmonary nodules based on machine learning - PMC - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11089274/>
8. Comparison of Veterans Affairs, Mayo, Brock classification models and radiologist diagnosis for classifying the malignancy of pulmonary nodules in Chinese clinical population - Translational Lung Cancer Research, accessed November 25, 2025, <https://tlcr.amegroups.org/article/view/32487/html>
9. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7708421/>
10. Deep Learning Models for Predicting Malignancy Risk in CT-Detected Pulmonary Nodules: A Systematic Review and Meta-analysis - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11427562/>
11. Unsupervised Domain Adaptation for Extra Features in the Target Domain Using Optimal Transport - arXiv, accessed November 25, 2025, <https://arxiv.org/pdf/2209.04594>
12. Domain adaptation in small-scale and heterogeneous biological datasets - PMC - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11661433/>
13. Accurate predictions on small data with a tabular foundation model - PubMed, accessed November 25, 2025, <https://pubmed.ncbi.nlm.nih.gov/39780007/>
14. TabPFN-2.5 Model Report - Prior Labs, accessed November 25, 2025, <https://priorlabs.ai/technical-reports/tabpfn-2-5-model-report>
15. Turning Tabular Foundation Models into Graph Foundation Models - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2508.20906v1>
16. (PDF) TabPFN-2.5: Advancing the State of the Art in Tabular Foundation Models, accessed November 25, 2025, <https://www.researchgate.net/publication/397555905_TabPFN-25_Advancing_the_State_of_the_Art_in_Tabular_Foundation_Models>
17. A Closer Look at TabPFN v2: Strength, Limitation, and Extension - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2502.17361v1>
18. Realistic Evaluation of TabPFN v2 in Open Environments - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2505.16226v1>
19. Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data - OpenReview, accessed November 25, 2025, <https://openreview.net/pdf/cd4d201c84349c24868d6907ee8950a9d89af245.pdf>
20. automl/Drift-Resilient\_TabPFN: Drift-Resilient TabPFN is a method using In-Context Learning via a Prior-Data Fitted Network, to address temporal distribution shifts in tabular data, outperforming existing methods in terms of performance and calibration. - GitHub, accessed November 25, 2025, <https://github.com/automl/Drift-Resilient_TabPFN>
21. Unsupervised Domain Adaptation for Mitigating Sensor Variability and Interspecies Heterogeneity in Animal Activity Recognition - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10603736/>
22. FedFusion: Federated Learning with Diversity- and Cluster-Aware Encoders for Robust Adaptation under Label Scarcity - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2509.19220v1>
23. Federated learning for medical imaging radiology - PMC - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10546441/>
24. Benchmarking Distribution Shift in Tabular Data with TableShift, accessed November 25, 2025, <https://proceedings.neurips.cc/paper_files/paper/2023/file/a76a757ed479a1e6a5f8134bea492f83-Paper-Datasets_and_Benchmarks.pdf>
25. mlfoundations/tableshift: A benchmark for distribution shift in tabular data - GitHub, accessed November 25, 2025, <https://github.com/mlfoundations/tableshift>