# Related Work: Methodological Foundations and Advances in Cross-Institutional Medical AI

## 1. Tabular Foundation Models and the Evolution of Structured Data Learning

The analysis of tabular data remains the bedrock of medical informatics, encompassing electronic health records (EHRs), laboratory biomarkers, and structured clinical assessments. Unlike perceptual modalities such as computer vision or natural language processing (NLP), where deep learning architectures have established unequivocal dominance, the trajectory of learning algorithms for structured data has been characterized by a prolonged and contentious competition between classical ensemble methods and emerging neural architectures. This section provides an exhaustive examination of this evolution, tracing the lineage from gradient-boosted decision trees to the recent advent of tabular foundation models, with a specific focus on their applicability to small-sample clinical cohorts characterized by high dimensionality and distributional volatility.

### 1.1. The Classical Paradigm: Gradient-Boosted Decision Trees (GBDTs)

For over a decade, Gradient-Boosted Decision Trees (GBDTs) have functioned as the *de facto* standard for tabular learning tasks, particularly within the biomedical domain where data often lacks the spatial or temporal continuity exploited by Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). The hegemony of algorithms such as XGBoost 1, LightGBM, and CatBoost 2 is not merely empirical but grounded in specific inductive biases that align remarkably well with the idiosyncrasies of clinical data.

Biomedical datasets are frequently populated by features with multimodal distributions, discrete values, and non-linear dependencies that do not conform to the smooth manifold assumptions underpinning many deep learning approaches.3 GBDTs address this through a process of iterative refinement, where an ensemble of weak learners—typically shallow decision trees—is constructed sequentially. Each subsequent tree targets the residuals (errors) of the preceding ensemble, thereby progressively reducing bias. XGBoost, in particular, revolutionized this space by introducing a sparsity-aware split-finding algorithm, allowing it to handle missing values—a pervasive issue in EHRs due to inconsistent testing protocols—without requiring explicit imputation strategies that might introduce noise.1 Furthermore, GBDTs are invariant to monotonic feature transformations, rendering them robust to the varying scales of clinical measurements (e.g., the difference in magnitude between white blood cell count and serum creatinine) without the necessity for rigorous normalization.4

Despite their robustness, GBDTs encounter severe theoretical and practical limitations when applied to the "small-sample" regime ($N < 1000$) characteristic of specialized medical cohorts, such as those for pulmonary nodule malignancy prediction.5 The boosting process relies on large-scale data to stabilize variance; in data-scarce environments, GBDTs are prone to overfitting, often memorizing stochastic noise rather than learning generalizable decision boundaries.6 Moreover, the non-differentiable nature of decision trees precludes their integration into end-to-end differentiable pipelines. This rigidity prevents the seamless application of gradient-based domain adaptation techniques or the joint training of multimodal systems (e.g., combining CT images with tabular biomarkers), necessitating complex and often suboptimal multi-stage workflows.7 The inability to perform "fine-tuning" in the deep learning sense means that a GBDT trained on one hospital's data cannot be easily adapted to another's without retraining from scratch or employing cumbersome model distillation techniques.

### 1.2. The "Deep Tabular" Renaissance: Attention and Embeddings

Motivated by the limitations of GBDTs and the desire to unify tabular learning with the differentiable ecosystem of deep learning, researchers have aggressively pursued neural architectures specialized for structured data. This "Deep Tabular" movement has sought to replicate the success of Transformers in NLP by developing mechanisms to capture feature interactions and handle categorical variables within a neural framework.

**TabNet**, introduced by Arik and Pfister, represented a seminal attempt to bridge the gap by embedding the feature selection logic of decision trees into a deep learning architecture.2 Utilizing a sequential attention mechanism, TabNet employs learnable masks to select salient features at each decision step, offering instance-wise interpretability analogous to tree paths. While theoretically compelling, extensive benchmarking has revealed that TabNet often requires significant hyperparameter tuning and large datasets to converge, frequently underperforming tuned XGBoost baselines on the smaller, noisy datasets typical of clinical studies.8

Subsequent innovations focused on adapting the Transformer architecture itself. **TabTransformer** introduced the concept of contextual embeddings for categorical features, mapping discrete values into dense vectors that are then processed by self-attention layers to learn high-order interactions.2 While effective for categorical-heavy datasets, this approach often neglected the nuances of numerical features, which are critical in medical contexts (e.g., laboratory values, radiomic features). **FT-Transformer (Feature Tokenizer Transformer)** addressed this by treating *all* features, numerical and categorical, as tokens.11 By projecting numerical values into a high-dimensional embedding space, FT-Transformer allows the self-attention mechanism to learn interactions across the entire feature set. Benchmarks indicate that FT-Transformer is among the strongest tabular deep learning models, often matching GBDTs on medium-sized datasets, yet it remains computationally expensive and data-hungry.13

**SAINT (Self-Attention and Intersample Attention Transformer)** further expanded this paradigm by introducing intersample attention.15 Unlike standard Transformers that process samples independently, SAINT allows the model to attend to other data points within the same batch during training. This mechanism theoretically improves performance on small datasets by implicitly learning from the relationships between similar patients (a form of nearest-neighbor reasoning). However, the quadratic complexity of attention ($O(N^2)$) limits its scalability, and rigorous evaluations have shown that on "medium-sized" datasets (1,000–10,000 samples), tree-based models often retain state-of-the-art performance while requiring a fraction of the training time.3

### 1.3. The Paradigm Shift: Tabular Foundation Models and In-Context Learning

The most significant recent development—and the methodological cornerstone of the PANDA framework discussed in the primary text—is the emergence of **Tabular Foundation Models**. This class of models represents a fundamental departure from the standard supervised learning paradigm where weights are initialized randomly and optimized via gradient descent on a specific target dataset. Instead, tabular foundation models leverage **Prior-Data Fitted Networks (PFNs)** and **In-Context Learning (ICL)** to generalize across tasks without parameter updates.14

TabPFN (Tabular Prior-Data Fitted Network):

Developed by Hollmann et al., TabPFN is a Transformer-based architecture pre-trained offline on millions of synthetic datasets generated from a diverse Structural Causal Model (SCM) prior.7 This prior encompasses a vast space of functional forms (including linear models, neural networks, and decision trees) and causal structures, effectively teaching the network to approximate Bayesian inference (Posterior Predictive Distribution, PPD) for any tabular classification task.19

During inference, TabPFN operates as a meta-learner. The entire labeled training set (the "context") and the unlabeled test samples (the "query") are fed into the Transformer in a single forward pass. The model utilizes self-attention to identify relationships between the query and the context examples, generating predictions based on the learned priors.20 This mechanism offers profound advantages for medical AI:

1. **Small-Sample Efficiency:** Because "learning" occurs via attention rather than iterative weight updates, TabPFN is immune to the traditional overfitting that plagues deep networks on small datasets ($N < 1,000$). It effectively interpolates based on the synthetic priors it has internalized.19
2. **Hyperparameter-Free Inference:** The model requires no architectural search, learning rate tuning, or epoch selection, eliminating "tuning bias" and drastically reducing the expertise required for deployment.7
3. **Computational Speed:** Inference is nearly instantaneous (< 1 second), facilitating real-time clinical decision support.15

Advancements and Variants:

The limitations of the original TabPFN, primarily its restriction to small context sizes (~1,000 samples) and classification tasks, have been addressed in subsequent iterations. TabPFN v2 and TabPFN-2.5 have expanded the context window to handle up to 10,000 samples and introduced support for regression, broadening the scope of applicability.15 Furthermore, Drift-Resilient TabPFN explicitly models temporal distribution shifts by incorporating a secondary SCM that governs how parameters evolve over time, showing superior calibration on out-of-distribution (OOD) data compared to static baselines.24 Other research has explored Tabular Large Language Models (TabLLMs), which serialize tabular data into text prompts to leverage the reasoning capabilities of LLMs like GPT-4.27 While promising for semantic reasoning, TabLLMs are computationally prohibitive for high-throughput risk prediction and often struggle with the precise numerical reasoning required for biomarker analysis.28

### 1.4. Critical Limitations in Clinical Deployment: The Closed-World Assumption

Despite the theoretical elegance of foundation models, their direct application to cross-hospital clinical tasks is hindered by the **closed-world assumption**. TabPFN, in its native form, assumes that the query sample (target domain) is drawn from the same distribution as the context set (source domain), i.e., $P\_{train}(X,Y) \approx P\_{test}(X,Y)$.25 This assumption is routinely violated in medical deployments due to:

1. **Covariate Shift:** Differences in medical equipment (e.g., CT scanner kernels, assay calibration) cause feature distributions to drift. A radiomic feature measuring "texture entropy" may have a mean of 0.5 at Hospital A and 0.8 at Hospital B purely due to acquisition parameters. TabPFN, relying on metric similarity in the input space, may attend to the wrong context examples, leading to overconfident but erroneous predictions.22
2. **Feature Heterogeneity:** Foundation models typically require a fixed or perfectly aligned feature set. They do not inherently resolve the "feature mismatch" problem where Hospital A records biomarkers $\{x\_1, x\_2, x\_3\}$ and Hospital B records $\{x\_1, x\_3, x\_4\}$.29 While RFE (Recursive Feature Elimination) can identify shared features, the reduction of dimensionality often discards valuable domain-specific signals.
3. **High-Dimensionality Bottleneck:** The quadratic complexity of attention limits TabPFN to datasets with a moderate number of features (typically < 100–500). High-dimensional inputs, such as raw radiomics or genomics, must be aggressively compressed, necessitating robust feature selection strategies that are themselves stable across domains.20

The following table summarizes the comparative strengths and weaknesses of these approaches in the context of medical tabular data.

| **Model Class** | **Representative Algorithms** | **Strengths in Medical AI** | **Limitations in Cross-Hospital Tasks** |
| --- | --- | --- | --- |
| **Tree Ensembles** | XGBoost, LightGBM, Random Forest | Interpretability, robustness to outliers, handles discrete/missing data naturally. | Overfits small samples ($N<500$); non-differentiable (hard to adapt); no inherent transfer learning. |
| **Deep Tabular** | TabNet, FT-Transformer, SAINT | Differentiable (allows gradients); captures complex interactions; multimodal integration potential. | Data hungry; complex tuning; often fails to beat trees on tabular benchmarks; computationally heavy. |
| **Foundation Models** | TabPFN, TabLLM | **State-of-the-art on small samples**; no tuning required; probabilistic output (Bayesian approximation). | **Sensitive to distribution shift**; limited context length; assumes aligned feature spaces; "black box" priors. |

Consequently, while foundation models solve the *sample efficiency* problem, they leave the *domain adaptation* problem unresolved. This necessitates a hybrid approach that combines the strong priors of foundation models with explicit, unsupervised alignment mechanisms (like TCA) to correct for distributional discrepancies before the in-context learning phase occurs.

## 2. Domain Adaptation in Medical AI: Bridging the Distributional Gap

The deployment of machine learning models in clinical practice is frequently stymied by the "domain shift" phenomenon, where a model trained on data from one institution (source domain) fails to generalize to data from another (target domain). In medical imaging and informatics, this drop in performance is not merely a technical nuisance but a safety hazard, potentially leading to systematic misdiagnosis in specific patient subpopulations. Domain Adaptation (DA) and Domain Generalization (DG) have emerged as critical subfields addressing this challenge, though their application to structured clinical data remains less mature than in imaging.

### 2.1. Taxonomy of Medical Domain Shift

Domain shift in medicine is multifaceted, typically categorized into three distinct forms that often occur simultaneously:

* **Covariate Shift:** Defined as $P\_s(X) \neq P\_t(X)$ while $P\_s(Y|X) = P\_t(Y|X)$. The marginal distribution of input features changes, but the fundamental biological relationship between features and outcomes remains constant. For example, pixel intensities in CT scans may vary due to different reconstruction kernels, tube voltages (kVp), or scanner manufacturers (e.g., GE vs. Siemens), causing the feature space to "drift".32 In tabular data, this manifests as varying reference ranges for laboratory tests across hospitals.
* **Prior Probability Shift (Label Shift):** Defined as $P\_s(Y) \neq P\_t(Y)$. The prevalence of the disease varies between cohorts. A specialized cancer center (source) might have a malignancy rate of 60% in pulmonary nodule patients, while a community screening program (target) might have a rate of only 5%. Models trained on balanced or high-prevalence data will be severely miscalibrated when applied to low-prevalence target domains.16
* **Concept Shift:** Defined as $P\_s(Y|X) \neq P\_t(Y|X)$. The conditional distribution changes, meaning the same feature vector implies a different outcome probability. This is the most pernicious shift, often caused by varying definitions of pathology (e.g., changing guidelines for "malignancy" over time) or unobserved confounders (e.g., a biomarker indicating sepsis in one hospital might indicate a different condition in another due to different comorbidity profiles).34

In multi-center studies, these shifts are compounded by **measurement heterogeneity**. Radiomic features, for instance, are notoriously sensitive to voxel spacing and segmentation algorithms; a "texture" feature quantifying tumor heterogeneity on one scanner may have a completely different value range on another, even for the same physical tumor.36

### 2.2. Dominant Paradigms: From Adversarial Learning to Statistical Alignment

The vast majority of medical DA research has focused on unstructured data, particularly medical imaging, giving rise to several dominant methodological families.

Adversarial Adaptation:

Inspired by Generative Adversarial Networks (GANs), Domain-Adversarial Neural Networks (DANN) employ a feature extractor trained to confuse a domain discriminator. The objective is a min-max game: the feature extractor tries to minimize classification error while maximizing the discriminator's confusion, theoretically forcing the learning of "domain-invariant" features.38 While effective for tasks like MRI-to-CT segmentation adaptation, adversarial training is notoriously unstable on small datasets. In the small-sample regime characteristic of tabular medical data, adversarial methods often suffer from "mode collapse" or the loss of discriminative features, as the discriminator can easily overpower the generator.40

Style Transfer and Generative Methods:

In imaging, CycleGANs are used to translate images from "Source Style" to "Target Style" (e.g., making a T1-weighted MRI look like a T2-weighted MRI) before feeding them into the classifier.32 However, generative methods carry the risk of "hallucination"—introducing artificial artifacts or erasing small pathologies (like pulmonary nodules)—which poses significant safety risks in diagnostic settings.

Statistical Moment Matching:

A more stable and mathematically grounded approach involves minimizing the distance between the statistical moments of the source and target distributions in a latent space. Maximum Mean Discrepancy (MMD) minimizes the difference in means between the two domains in a Reproducing Kernel Hilbert Space (RKHS), effectively aligning the first-order statistics.40 Deep CORAL (Correlation Alignment) extends this to second-order statistics, aligning the covariance matrices of the feature distributions.42 These methods are computationally efficient and theoretically derived from minimizing the generalization error bound ($R\_t \leq R\_s + \text{divergence}$), making them particularly suitable for tabular data where preserving the statistical structure is paramount.

### 2.3. The Specific Challenges of Tabular Domain Adaptation

While imaging DA is a mature field, domain adaptation for **structured clinical data** (EHRs, tabular biomarkers) remains significantly underexplored and faces distinct challenges that render many imaging-based techniques ineffective.

1. **Feature Heterogeneity and Mismatch:** Unlike images, which always share a common pixel grid structure, tabular datasets across hospitals often differ in their schema. Hospital A might record "WBC Count" while Hospital B records "Leukocytes" (semantic mismatch), or Hospital B might simply lack the variable entirely. **Open-Set Heterogeneous Domain Adaptation (OSHeDA)** addresses scenarios where both feature and label spaces differ, but these methods often require complex matrix completion or graph matching that scales poorly and introduces noise.43
2. **Missingness Shift:** The pattern of missing data is often a strong predictor in EHRs (e.g., a lactate test is only ordered if the patient is suspected of sepsis, making the *presence* of the value informative). If the "missingness mechanism" changes between hospitals due to policy differences (e.g., Hospital B orders lactate for all ER admissions), standard imputation and adaptation fail. Recent work on **Domain Adaptation under Missingness Shift (DAMS)** has theoretically proven that such shifts constitute a violation of standard covariate shift assumptions, requiring specialized adjustment mechanisms.30
3. **Discrete and Mixed Distributions:** Tabular data contains a mix of continuous, ordinal, and categorical variables. Geometric alignment methods like Optimal Transport (Earth Mover's Distance), which rely on the continuity of the underlying manifold (typical in images), often fail or become computationally intractable on the discrete, disjoint manifolds of tabular data.38

### 2.4. Transfer Component Analysis (TCA): A Robust Solution for Small Samples

Among the plethora of DA methods, **Transfer Component Analysis (TCA)** remains a seminal technique that is uniquely well-suited for small-sample tabular data. Proposed by Pan et al., TCA avoids the complexity and instability of adversarial training by learning a transformation matrix $W$ that projects data into a Reproducing Kernel Hilbert Space (RKHS).45

The optimization objective of TCA is to minimize the Maximum Mean Discrepancy (MMD) between source and target distributions in this projected space, while simultaneously preserving the variance (intrinsic structure) of the data to prevent the collapse of discriminative information. Unlike deep neural network adaptation, TCA has a closed-form solution based on generalized eigendecomposition. This property makes it **deterministic** and highly **stable**, even when $N < 500$—a scenario where deep learning-based DA often fails to converge.47 Recent applications in EEG signal processing and biological data integration have confirmed that TCA effectively aligns marginal distributions without requiring massive target datasets, positioning it as a pragmatic, mathematically sound choice for enhancing the embeddings of foundation models like TabPFN.48

### 2.5. Emerging Trends: Federated and Partial Adaptation

Privacy constraints imposed by regulations such as HIPAA and GDPR often make pooling data from multiple hospitals legally impossible. **Federated Learning (FL)** allows models to train locally and share weights, but standard FL struggles when data is Non-IID (non-identically distributed) across clients.29 **Federated Domain Adaptation** has emerged to bridge this gap, employing techniques like **FedFusion** or prototype alignment to align feature spaces without exchanging row-level data. Additionally, **Partial Domain Adaptation (PDA)** addresses the realistic scenario where the source domain has a larger label space (e.g., diverse disease subtypes) than the target domain, requiring the model to identify and ignore "outlier" source classes to prevent negative transfer.51

## 3. Pulmonary Nodule Risk Prediction and Cross-Hospital Generalization

The classification of solitary pulmonary nodules (SPNs) detected on Computed Tomography (CT) represents one of the most critical bottlenecks in thoracic oncology. With the widespread adoption of Low-Dose CT (LDCT) screening for lung cancer, the detection rate of incidental nodules has surged. However, the vast majority of these nodules are benign (granulomas, hamartomas, scars). The clinical imperative is to accurately identify the minority of malignant nodules to enable early intervention while sparing patients with benign nodules from unnecessary invasive biopsies, radiation exposure, and severe anxiety.

### 3.1. The Rise and Fall of Classical Clinical Risk Models

For decades, risk stratification has relied on logistic regression models derived from specific, well-curated cohorts. These "classical" models established the baseline for nodule evaluation but have revealed significant fragility when deployed outside their derivation settings.

* **The Mayo Clinic Model:** Developed in 1997 on a cohort of 629 patients, this model utilizes six variables: age, smoking history, history of cancer, nodule diameter, spiculation, and upper lobe location.27 While it achieved an AUC of ~0.83 in its derivation set, external validation studies have repeatedly shown performance degradation. A critical failure mode is observed in Asian populations: the "upper lobe location" variable, a strong predictor of malignancy in the US (where cancer often occurs in the upper lobes), is confounded by tuberculosis, which also predilects the upper lobes. Consequently, the specificity of this variable collapses, causing AUCs to drop to 0.60–0.65 in Chinese and Indian cohorts.54
* **The Brock (PanCan) Model:** Derived from the Pan-Canadian Early Detection of Lung Cancer Study, this model incorporates more granular radiological features such as emphysema and nodule count.2 It is generally considered more robust than the Mayo model for screening populations. However, external validation on the massive National Lung Screening Trial (NLST) dataset revealed significant calibration drift. While discrimination remained relatively high (AUC ~0.90), the model systematically overestimated malignancy risk, necessitating recalibration of the intercept and coefficients to align predicted probabilities with observed rates.56
* **The Veterans Affairs (VA) Model:** Developed on a military veteran population (predominantly male, heavy smokers, high comorbidity burden), this model generalizes poorly to general screening populations (which include females and non-smokers) due to severe Prior Probability Shift and Covariate Shift.58

The core failure mode of these classical models is their **rigidity**. The coefficients are fixed, learned from a specific "source domain" (e.g., 1990s Minnesota). When the "target domain" shifts (e.g., 2020s Beijing), the fixed weights no longer accurately reflect the conditional probability $P(Y|X)$, leading to substantial errors in risk estimation.

### 3.2. Radiomics and the Reproducibility Crisis

To transcend the limited variables of clinical models, **radiomics** emerged as a high-throughput alternative, extracting hundreds or thousands of quantitative features (texture, shape, wavelet transforms, GLCM, GLRLM) from the segmented nodule volume.60 While early single-center studies reported impressive AUCs (> 0.90), radiomics has faced a severe **reproducibility crisis**.

Radiomic features are essentially mathematical descriptors of pixel intensity distributions. However, these intensities are highly sensitive to CT acquisition parameters. A study comparing features across different scanners found that most texture features (e.g., GLCM entropy) varied significantly based solely on the **scanner vendor** (GE vs. Siemens), **reconstruction kernel** (sharp vs. smooth), and **slice thickness**.36 Consequently, a model trained on "sharp kernel" images interprets a "smooth kernel" image as having different texture properties, often leading to false classifications. External validation studies have shown that radiomic models can lose 0.10–0.20 in AUC when applied to external datasets, rendering them unreliable for clinical transportability.62 While harmonization techniques like **ComBat** (adapted from genomics) can regress out scanner effects, they require batch information that is often unavailable during single-patient inference.62

### 3.3. Deep Learning: Performance vs. Generalizability

The current state-of-the-art in nodule classification is defined by Deep Convolutional Neural Networks (CNNs), such as 3D-ResNets and DenseNets, which learn features directly from raw voxel data, theoretically bypassing the need for handcrafted radiomics.64

* **State-of-the-Art Performance:** In controlled benchmarks like the LUNA16 challenge or the NLST dataset, CNNs consistently outperform both clinical models and radiomics, achieving sensitivities exceeding 90%.65
* **The "Shortcut Learning" Trap:** However, CNNs are notoriously prone to learning "shortcuts"—spurious correlations specific to the source domain. For instance, a model might learn to associate a specific noise pattern, reconstruction artifact, or even the presence of a ruler/marker in the image (unique to a specific hospital) with malignancy, rather than the biological morphology of the nodule.66
* **Cross-Center Degradation:** Empirical studies have demonstrated that deep learning models suffer significant performance drops when tested on external cohorts. One study noted a sensitivity drop from 96% (internal) to 81% (external) due to shifts in prevalence and imaging characteristics.67 While Transfer Learning (fine-tuning on the target domain) can mitigate this, it requires labeled target data, which contradicts the unsupervised adaptation requirement of many real-world clinical deployments where local labels are scarce or nonexistent.

### 3.4. The "Triple Knot": A Unified Challenge

The literature reveals a distinct gap in current methodologies. Classical models are interpretable but too rigid and prone to calibration drift. Radiomics offers detail but suffers from extreme sensitivity to acquisition parameters. Deep CNNs offer high performance but are data-hungry "black boxes" prone to shortcut learning and poor generalization. Furthermore, most approaches treat "images" and "clinical tables" as separate entities or simply concatenate them without addressing the fundamental distributional shifts in the tabular components (biomarkers, demographics).

There is a conspicuous lack of methods that specifically address the **tabular** nature of clinical biomarkers using modern **foundation models** (to handle small $N$) while explicitly correcting for **distribution shift** via unsupervised alignment (to handle cross-hospital differences). The integration of **TabPFN** (for small-sample robustness), **RFE** (to select domain-stable features), and **TCA** (to align distributions) addresses this specific "triple knot" of challenges—Small Sample Size, Feature Heterogeneity, and Distribution Shift—representing a novel synthesis in the field of medical AI.

## 4. Benchmarking Tabular Distribution Shift

To rigorously quantify the phenomenon of tabular distribution shift and evaluate the robustness of proposed solutions, the machine learning community has recently moved towards standardized benchmarks, acknowledging that ad-hoc evaluation on single train/test splits is insufficient to measure true generalization.

### 4.1. TableShift and Emerging Standards

The **TableShift** benchmark 69 represents the most comprehensive effort to date to systematize the evaluation of tabular domain shift. It curates 15 binary classification tasks across diverse domains including healthcare (e.g., ICU mortality prediction, hospital readmission), finance, and public policy. Crucially, TableShift does not rely on random splits; it explicitly constructs "shift" splits based on real-world variables such as time (training on past data, testing on future data), geography (training on one state, testing on another), or subpopulation (e.g., training on patients with private insurance, testing on Medicaid patients).

Key findings from the TableShift benchmark analysis 16 have profound implications for medical modeling:

1. **The "Linear Trend":** There is a strong linear correlation between In-Distribution (ID) and Out-of-Distribution (OOD) accuracy. Models that perform better on the source domain generally perform better on the target domain. This suggests that "robustness" is often a side effect of better representation learning rather than the result of specialized robustness objectives. However, this trend breaks down under severe shifts, where specialized adaptation is required.
2. **Failure of Specialized Domain Generalization Methods:** Sophisticated domain generalization techniques such as **GroupDRO** (Group Distributionally Robust Optimization) or **IRM** (Invariant Risk Minimization), which have shown promise in computer vision, often fail to outperform simple Empirical Risk Minimization (ERM) or strong GBDT baselines on tabular data. This highlights the unique nature of the tabular manifold and the difficulty of learning invariant features in discrete spaces.
3. **Label Shift Dominance:** A significant portion of the performance drop in tabular shift is driven by label shift (changes in $P(Y)$), which feature alignment alone cannot fix. This necessitates the use of methods that can adapt to prevalence changes, such as the class-balanced sampling strategies employed in PANDA.

### 4.2. The Wild-Time Benchmark

Complementing TableShift, the **Wild-Time** benchmark focuses specifically on **temporal distribution shifts**.72 It evaluates models on data streams that evolve over time, a scenario highly relevant to healthcare where patient populations, disease strains (e.g., COVID-19 variants), and clinical protocols change dynamically. Wild-Time reinforces the finding that standard models degrade over time and that methods explicitly modeling the *drift* (like Drift-Resilient TabPFN) are necessary to maintain performance in longitudinal deployments.

### 4.3. Implications for the Current Study

These benchmarks underscore that achieving robustness in tabular medical data is non-trivial and that "off-the-shelf" domain generalization methods from the vision domain are unlikely to succeed. This reinforces the need for domain-specific adaptations—such as the explicit **cross-domain feature selection (RFE)** and **statistical alignment (TCA)** proposed in the PANDA framework—rather than relying solely on the "magic" of complex deep learning architectures. The combination of *foundation model priors* (which provide a strong, sample-efficient baseline) with *statistical alignment* (which corrects the specific distributional shift) appears to be a more promising theoretical path than end-to-end robust optimization for the specific constraints of cross-hospital pulmonary nodule prediction.21

### Conclusion

The synthesis of recent literature indicates that while pulmonary nodule characterization has advanced through radiomics and deep learning, cross-hospital generalization remains an unsolved problem due to the "triple knot" of small sample sizes, distribution shifts, and feature heterogeneity. Tabular foundation models like TabPFN offer a breakthrough in handling small samples but lack inherent mechanisms for domain adaptation. Conversely, domain adaptation methods like TCA are theoretically sound but have rarely been applied to the latent spaces of foundation models in this specific clinical context. The integration of these distinct streams of research represents a logical and necessary step toward robust, deployable medical AI.

#### Works cited

1. Industry Perspective: Tree-Based Models vs Deep Learning for Tabular Data - Aidan Cooper, accessed November 25, 2025, <https://www.aidancooper.co.uk/tree-based-models-vs-deep-learning/>
2. [2410.12034] A Survey on Deep Tabular Learning - arXiv, accessed November 25, 2025, <https://arxiv.org/abs/2410.12034>
3. Why do tree-based models still outperform deep learning on typical tabular data?, accessed November 25, 2025, <https://papers.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf>
4. Why Tree-Based Models Beat Deep Learning on Tabular Data | by Devansh | Geek Culture, accessed November 25, 2025, <https://medium.com/geekculture/why-tree-based-models-beat-deep-learning-on-tabular-data-fcad692b1456>
5. AI\_Healthcare\_Analytics\_2025.pdf
6. A Comprehensive Benchmark of Machine and Deep Learning Across Diverse Tabular Datasets - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2408.14817v1>
7. 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>
8. maciekmalachowski/TabNet-vs-XGBoost - GitHub, accessed November 25, 2025, <https://github.com/maciekmalachowski/TabNet-vs-XGBoost>
9. A Comparative Study of TabNet and XGBoost for Tabular Data Classification, accessed November 25, 2025, <https://www.researchgate.net/publication/392509412_A_Comparative_Study_of_TabNet_and_XGBoost_for_Tabular_Data_Classification>
10. Tabular prior-data fitted network in real-world CT radiomics: benign vs. malignant renal tumor classification - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12591794/>
11. Basis Transformers for Multi-Task Tabular Regression - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2506.06926>
12. FTTransformer: Transformer Architecture for Tabular Datasets | by Arash Khoeini - Medium, accessed November 25, 2025, <https://arashk.medium.com/fttransformer-transformer-architecture-for-tabular-datasets-d4bfe591d6fb>
13. Tabular deep learning: a comparative study applied to multi-task genome-wide prediction, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11452967/>
14. A Comparative Study of Advanced Transformer Learning Frameworks for Water Potability Analysis Using Physicochemical Parameters - MDPI, accessed November 25, 2025, <https://www.mdpi.com/2076-3417/15/13/7262>
15. TabPFN-2.5 Model Report - Prior Labs, accessed November 25, 2025, <https://priorlabs.ai/technical-reports/tabpfn-2-5-model-report>
16. 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>
17. Tabular Data: Is Deep Learning all you need? - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2402.03970>
18. How the Rise of Tabular Foundation Models Is Reshaping Data Science, accessed November 25, 2025, <https://towardsdatascience.com/tabular-foundation-models/>
19. Tabular Foundation Model: TabPFN - Emergent Mind, accessed November 25, 2025, <https://www.emergentmind.com/topics/tabular-foundation-model-tabpfn>
20. The Limitations of TabPFN for High-Dimensional RNA-seq Analysis - bioRxiv, accessed November 25, 2025, <https://www.biorxiv.org/content/10.1101/2025.08.15.670537v1.full.pdf>
21. TabPFN Opens New Avenues for Small-Data Tabular Learning in Drug Discovery, accessed November 25, 2025, <https://chemrxiv.org/engage/chemrxiv/article-details/68d29b1cf2aff1677025b18f>
22. A Closer Look at TabPFN v2: Strength, Limitation, and Extension - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2502.17361v1>
23. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data - ICML 2025, accessed November 25, 2025, <https://icml.cc/virtual/2025/poster/46681>
24. [Quick Review] Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data - Liner, accessed November 25, 2025, <https://liner.com/review/driftresilient-tabpfn-incontext-learning-temporal-distribution-shifts-on-tabular-data>
25. Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data - OpenReview, accessed November 25, 2025, <https://openreview.net/pdf/cd4d201c84349c24868d6907ee8950a9d89af245.pdf>
26. [2411.10634] Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data - arXiv, accessed November 25, 2025, <https://arxiv.org/abs/2411.10634>
27. Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2501.03540v1>
28. Transformers Boost the Performance of Decision Trees on Tabular Data across Sample Sizes - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2502.02672v1>
29. Open-Set Heterogeneous Domain Adaptation: Theoretical Analysis and Algorithm - AAAI Publications, accessed November 25, 2025, <https://ojs.aaai.org/index.php/AAAI/article/view/34191/36346>
30. Domain Adaptation under Missingness Shift - Proceedings of Machine Learning Research, accessed November 25, 2025, <https://proceedings.mlr.press/v206/zhou23b/zhou23b.pdf>
31. TabPFN Unleashed: A Scalable and Effective Solution to Tabular Classification Problems, accessed November 25, 2025, <https://arxiv.org/html/2502.02527v1>
32. Domain Adaptation for Medical Image Analysis: A Survey - PMC - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9011180/>
33. Federated learning for medical image analysis: A survey - PMC - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10976951/>
34. Domain Generalization for Medical Image Analysis: A Survey - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2310.08598v2>
35. AdaDiag: Adversarial Domain Adaptation of Diagnostic Prediction with Clinical Event Sequences - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9580228/>
36. Reproducibility of Lung Nodule Radiomic Features: Multivariable and Univariable Investigations that Account for Interactions Between CT Acquisition and Reconstruction Parameters - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8273077/>
37. Narrative review of radiomics for classifying pulmonary nodules and potential impact on lung cancer screening - Current Challenges in Thoracic Surgery, accessed November 25, 2025, <https://ccts.amegroups.org/article/view/46726/html>
38. [PDF] Domain Adaptation for Medical Image Analysis: A Survey | Semantic Scholar, accessed November 25, 2025, <https://www.semanticscholar.org/paper/Domain-Adaptation-for-Medical-Image-Analysis%3A-A-Guan-Liu/7c92cced2d958e9f43af88fdd1d9cc2c281a6465>
39. Domain Adaptation for Medical Image Analysis: A Survey - arXiv, accessed November 25, 2025, <https://arxiv.org/pdf/2102.09508>
40. Domain Adaptation Techniques for Natural and Medical Image Classification - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2508.20537v1>
41. Towards Backwards-Compatible Data with Confounded Domain Adaptation - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2203.12720v3>
42. (PDF) Comparative Analysis of Tree-Based Models and Deep Learning Architectures for Tabular Data: Performance Disparities and Underlying Factors - ResearchGate, accessed November 25, 2025, <https://www.researchgate.net/publication/378541500_Comparative_Analysis_of_Tree-Based_Models_and_Deep_Learning_Architectures_for_Tabular_Data_Performance_Disparities_and_Underlying_Factors>
43. Open-Set Heterogeneous Domain Adaptation: Theoretical Analysis and Algorithm - PMC, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12286580/>
44. Domain Adaptation under MNAR Missingness - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2504.00322v1>
45. Domain Generalization based on Transfer Component Analysis - Ancalime.de, accessed November 25, 2025, <https://www.ancalime.de/download/publications/Grubinger2015_DomainGeneralizationBasedOnTransferComponentAnalysis.pdf>
46. DomainATM: Domain adaptation toolbox for medical data analysis - PMC, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9908850/>
47. Domain Adaptation Principal Component Analysis: Base Linear Method for Learning with Out-of-Distribution Data - MDPI, accessed November 25, 2025, <https://www.mdpi.com/1099-4300/25/1/33>
48. [2102.09508] Domain Adaptation for Medical Image Analysis: A Survey - arXiv, accessed November 25, 2025, <https://arxiv.org/abs/2102.09508>
49. Transfer learning for non-image data in clinical research: A scoping review, accessed November 25, 2025, <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000014>
50. Federated Learning for Medical Image Analysis: A Survey - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2306.05980v4>
51. Contrastive Learning Assisted-Alignment for Partial Domain Adaptation - IEEE Xplore, accessed November 25, 2025, <https://ieeexplore.ieee.org/iel7/5962385/6104215/09705553.pdf>
52. Multi-attention representation network partial domain adaptation for COVID-19 diagnosis - PMC - PubMed Central, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9222222/>
53. Solitary Pulmonary Nodule (SPN) Malignancy Risk Score (Mayo Clinic Model) - MDCalc, accessed November 25, 2025, <https://www.mdcalc.com/calc/4057/solitary-pulmonary-nodule-spn-malignancy-risk-score-mayo-clinic-model>
54. Comparison of Brock University, Mayo Clinic and Herder models for pretest probability of cancer in solid pulmonary nodules - PMC - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC9629988/>
55. Predicting lung cancer prior to surgical resection in patients with lung nodules - PMC - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4272613/>
56. External validation and recalibration of the Brock model to predict probability of cancer in pulmonary nodules using NLST data, accessed November 25, 2025, <https://qcb.ucla.edu/wp-content/uploads/sites/14/2020/04/PMID-30898897.pdf>
57. External validation and recalibration of the Brock model to predict ..., accessed November 25, 2025, <https://pubmed.ncbi.nlm.nih.gov/30898897/>
58. Artificial Intelligence in Healthcare: 2024 Year in Review - medRxiv, accessed November 25, 2025, <https://www.medrxiv.org/content/10.1101/2025.02.26.25322978v1.full-text>
59. Comprehensive Analysis of Clinical Logistic and Machine Learning-Based Models for the Evaluation of Pulmonary Nodules - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8980995/>
60. Deep Neural Networks and Tabular Data: A Survey - IEEE Xplore, accessed November 25, 2025, <https://ieeexplore.ieee.org/iel7/5962385/10547160/09998482.pdf>
61. Reproducibility of radiomic features of pulmonary nodules between low-dose CT and conventional-dose CT - PMC - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8923849/>
62. External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis - PubMed, accessed November 25, 2025, <https://pubmed.ncbi.nlm.nih.gov/32488865/>
63. Performance of deep learning model and radiomics model for preoperative prediction of spread through air spaces in the surgically resected lung adenocarcinoma: a two-center comparative study - Translational Lung Cancer Research, accessed November 25, 2025, <https://tlcr.amegroups.org/article/view/94729/html>
64. Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules | American Journal of Respiratory and Critical Care Medicine, accessed November 25, 2025, <https://www.atsjournals.org/doi/10.1164/rccm.201903-0505OC>
65. A Novel Deep Learning-Based (3D U-Net Model) Automated Pulmonary Nodule Detection Tool for CT Imaging - MDPI, accessed November 25, 2025, <https://www.mdpi.com/1718-7729/32/2/95>
66. Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification - NIH, accessed November 25, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8637230/>
67. Impact of localized fine tuning in the performance of segmentation and classification of lung nodules from computed tomography scans using deep learning - Frontiers, accessed November 25, 2025, <https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1140635/full>
68. Enhanced Malignancy Prediction of Small Lung Nodules in Different Populations Using Transfer Learning on Low-Dose Computed Tomography - MDPI, accessed November 25, 2025, <https://www.mdpi.com/2075-4418/15/12/1460>
69. TableShift · TableShift, accessed November 25, 2025, <https://tableshift.org/>
70. A Survey on Domain Generalization for Medical Image Analysis - arXiv, accessed November 25, 2025, <https://arxiv.org/html/2402.05035v1>
71. [2312.07577] Benchmarking Distribution Shift in Tabular Data with TableShift - arXiv, accessed November 25, 2025, <https://arxiv.org/abs/2312.07577>
72. Wild-Tab: A Benchmark For Out-Of-Distribution Generalization In Tabular Regression - arXiv, accessed November 25, 2025, <https://arxiv.org/abs/2312.01792>