Causal Pre-training Under the Fairness Lens: An Empirical Study of TabPFN

Qinyi Liu
University of Bergen
Centre for the Science of Learning & Technology (SLATE)
Bergen, Norway
qinyi.liu@uib.no

Mohammad Khalil
University of Bergen
Centre for the Science of Learning & Technology (SLATE)
Bergen, Norway
mohammad.khalil@uib.no

Naman Goel
Department of Computer Science
University of Oxford and Alan Turing Institute
Oxford, United Kingdom
naman.goel@cs.ox.ac.uk

Abstract

Foundation models for tabular data, such as the Tabular Prior-data Fitted Network (TabPFN), are pre-trained on a massive number of synthetic datasets generated by structural causal models (SCM). They leverage in-context learning to offer high predictive accuracy in real-world tasks. However, the fairness properties of these foundational models, which incorporate ideas from causal reasoning during pre-training, remain underexplored. In this work, we conduct a comprehensive empirical evaluation of TabPFN and its fine-tuned variants, assessing predictive performance, fairness, and robustness across varying dataset sizes and distributional shifts. Our results reveal that while TabPFN achieves stronger predictive accuracy compared to baselines and exhibits robustness to spurious correlations, improvements in fairness are moderate and inconsistent, particularly under missing-not-at-random (MNAR) covariate shifts. These findings suggest that the causal pre-training in TabPFN is helpful but insufficient for algorithmic fairness, highlighting implications for deploying TabPFN (and similar) models in practice and the need for further fairness interventions.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Social and professional topics** → *Algorithmic fairness.*

Keywords

Algorithmic Fairness, Tabular Data, Fairness, Bias Robustness, Machine Learning, Datasets

ACM Reference Format:
Qinyi Liu, Mohammad Khalil, and Naman Goel. 2026. Causal Pre-training Under the Fairness Lens: An Empirical Study of TabPFN. In *Proceedings of The Web Conference 2026 (WWW '26).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Algorithmic fairness is a critical concern in the deployment of machine learning (ML) models for high-stakes decision-making. This concern is often compounded by issues such as spurious correlations and distribution shifts in the data [18]. Foundation models for tabular data, such as Tabular Prior-data Fitted Network (TabPFN) [7], are pre-trained on millions of synthetic datasets based on structured causal models (SCM) and leverage in-context learning (ICL) to make accurate predictions on real datasets. This reduced reliance on real and biased data for prediction and the use of SCM generated synthetic data in pre-training motivates a systematic empirical evaluation of its fairness implications, since bias in machine learning models is partially attributed to different types of biases in the training datasets [13] (rigorous justification for the motivation is presented in Section 2.2).

However, despite impressive predictive accuracy and thus, growing interest in its widespread use in practical applications, the behavior of TabPFN with respect to fairness, especially under distributional shifts, has not been thoroughly studied.

We conduct a systematic empirical evaluation of TabPFN and its fine-tuned variants, which yields a comprehensive understanding of their predictive accuracy, fairness, and robustness under varying ICL dataset sizes and distribution shifts. We thus also explore the intrinsic fairness potential of SCM-based priors, revealing both strengths—such as robustness to spurious correlations—and limitations, particularly under missing-not-at-random (MNAR) covariate shifts. Our results offer practical insights for deploying these causally-informed foundation models in practice and highlighting scenarios where additional fairness intervention may be necessary.

2 Background and Related Work

2.1 Fairness and its Challenge in ML

Algorithmic fairness is a cornerstone of responsible ML and the broader WWW community, ensuring that predictions do not discriminate against protected groups [11]. Two prevalent notions of group fairness are *Demographic Parity (DP)* and *Equalized Odds (EO)* [4, 20]. DP requires predictions ($\hat{Y}$) to be independent of sensitive attributes (S): $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$. EO demands conditional independence given the true label (Y): $P(\hat{Y} = 1 \mid Y = y, S = 0) = P(\hat{Y} = 1 \mid Y = y, S = 1)$ for $y \in \{0, 1\}$. Except in trivial cases (e.g., equal base rates), different fairness and accuracy criteria cannot be satisfied simultaneously, this is the fundamental challenge in algorithmic fairness [3, 10].

Further, achieving fairness criteria is challenging in practice due to various types of biases in training datasets. Traditional ML models (e.g., logistic regression) achieve high accuracy by exploiting statistical correlations, but are vulnerable to two interrelated failures (among others), rooted in biased training data [13]:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WWW '26, Dubai
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/XXXXXXX.XXXXXXX

1. Spurious correlations arise from underspecified or imbalanced datasets [22]. When data are limited, multiple hypotheses can fit the evidence equally. Following Occam's Razor, models tend to prefer simpler hypotheses, which in biased datasets may correspond to non-causal shortcuts. For example, a model may use correlated proxies such as race or ZIP code to predict income, even in the absence of a causal link. Sampling noise and majority-group overrepresentation further reinforce such patterns, leading models to misattribute predictive power to sensitive attributes [13].
2. Covariate shift can originate from selection bias during data collection, where the distribution of features in the training set does not represent the true population. Formally, this occurs when $P_{\text{train}}(X) \neq P_{\text{test}}(X)$, often due to biased inclusion probabilities $P(\text{included} \mid X) \neq P(\text{included})$ [11]. As a result, models learn decision boundaries that rely on non-stable correlations specific to the training context rather than invariant causal relationships.

Both phenomena share similar origin: selection-biased data that fail to capture invariant causal mechanisms.

A promising direction for addressing the aforementioned challenges involves integrating causal reasoning into tabular models [1, 12]. Rather than solely relying on correlations found in potentially biased empirical data, these approaches aim to learn stable relationships that align with underlying causal mechanisms.

2.2 Pre-Trained Tabular Foundation Model

TabPFN is a recently proposed transformer-based foundation model for tabular supervised learning [5–7]. Unlike conventional classifiers trained from scratch on individual datasets, TabPFN is *pre-trained once* on millions of synthetic tabular tasks generated from SCMs. Synthetic tasks reflect different prediction tasks (diverse causal structures, noise levels, feature imbalances, etc). Rather than learning dataset-specific correlations, TabPFN approximates the posterior predictive distribution, effectively performing Bayesian inference over plausible causal mechanisms. The models use real task-specific data only at inference time, similar to ICL in LLMs [2, 8].

The TabPFN design yields several theoretical advantages that can be relevant to fairness and robustness challenges discussed in Section 2.1. First, the SCM-based prior grounds the model in stable causal dependencies, potentially mitigating reliance on spurious correlations tied to sensitive attributes in real datasets. Second, by meta-learning across a wide range of causal environments, TabPFN can potentially improve generalization under covariate and selection shifts. Last but not least, its Bayesian-style inference can potentially provide regularization against overfitting to biased or undersampled data, enhancing fairness in low-sample regimes.

Prior TabPFN studies have demonstrated state-of-the-art accuracy on a wide range of small- to medium-sized tabular benchmarks (smaller than 10,000 training samples), outperforming carefully tuned tree ensembles and AutoML systems [7]. While these results highlight strong predictive performance, a more detailed evaluation of TabPFN's fairness, particularly under spurious correlations and covariate shift is still lacking, extending beyond initial fairness explorations in tabular in-context learning (e.g., [8]).

FairPFN has recently been proposed to directly optimize causal fairness metrics (e.g. counterfactual fairness) during the model adaptation process [17]. However, its current formulation is restricted to binary sensitive attributes. In this study, we *instead* use TabPFN without such specific fairness adaptations, allowing us to study the group fairness behavior of the more widely used models (i.e base TabPFN and fine-tuned FT-TabPFN) under different stress-testing scenarios. This also allows us to study how far causal pre-training alone can improve group fairness, without further fairness interventions.

This paper, thus, addresses the following research questions (RQs):

RQ1: How do TabPFN and FT-TabPFN perform in terms of accuracy and fairness across different dataset sizes, compared with traditional models?

RQ2: How robustly do TabPFN and FT-TabPFN preserve both accuracy and fairness in the presence of covariate shift and spurious correlation?

3 Experiments

Datasets. We use four standard tabular fairness benchmark datasets: Heart (303 samples, 13 features; sensitive: age, sex) [15], Bank (5,000, 14; education, family) [14], Law (21k, 8; race, sex) [21], and Adult (48k, 14; race, sex) [16]. Features were standardized and categorical features one-hot encoded; all models share the same preprocessing.
Models. We evaluate five models on tabular classification tasks, comprising three classical baselines and two variants of TabPFN: (1) *Logistic Regression (LR)* with L2 regularization (C=0.1), (2) *Random Forest (RF)* with 50 trees of maximum depth 5 and minimum samples per split of 10, (3) *Multi-Layer Perceptron (MLP)* with one hidden layer of 50 units, L2 regularization (α=0.01), and up to 300 training iterations, (4) *TabPFN* in zero-shot mode with pretraining limits disabled (allowing training on datasets exceeding 10k samples), $n_estimators$=2, and inference using 5000 subsampled training examples as context, and (5) *FT-TabPFN*, fine-tuned for 10 epochs via Adam (lr=1×10^{-5}), meta-batch size 1, inner batch size 5000, and cross-entropy loss on batched in-context predictions, evaluated with the same 5000-sample context. Further implementation details can be found in code, available at https://github.com/ql909/An-Empirical-Study-of-TabPFN.

3.1 RQ1

To answer RQ1, we evaluate model performance using accuracy and two standard fairness metrics—EO and DP—implemented via FairLearn[1]. Although TabPFN is designed for small- to medium-sized tabular datasets (up to roughly 10k samples) [7], we also examine its behavior on larger datasets. For the Adult and Law datasets, we further evaluate TabPFN on subsets of 500, 1k, and 10k samples to analyze how fairness varies with dataset size.

3.2 RQ2

Spurious correlations To evaluate robustness against non-causal statistical dependencies, we follow the literature on spurious correlation stress-testing [19] and design a *pseudo-correlation perturbation* experiment. The key idea is to introduce an artificial feature Z_{spur} that is strongly correlated with the target label in the training set but has its correlation sign reversed in the test set, allowing

[1]FairLearn: A Toolkit for Assessing and Improving Fairness in AI (Bird et al., 2020; Version 0.7.0).

us to examine reliance on shortcut correlations versus invariant causal features.

Formally, we augment each dataset with a synthetic continuous variable $Z_{\text{spur}} \sim \mathcal{N}(\mu_y, 0.5)$, where $\mu_y = 1$ if $Y = 1$ and 0 otherwise in training data, and flipped to $\mu_y = -1$ if $Y = 1$ and 0 otherwise in the perturbed test data. Models capturing spurious dependencies should degrade in predictive performance and fairness upon inversion of the Z_{spur}–Y correlation. We measure this sensitivity via differences in accuracy and fairness between original and flipped test sets, plus the *Flip Consistency Rate* (proportion of unchanged predictions). All results are averaged over five independent runs with different random seeds for statistical reliability. Intuitively, causal-generalizing models should exhibit minimal degradation in utility and fairness upon flipping Z_{spur}.

Covariate shift We study a common form of covariate shift induced by *missing-not-at-random* (MNAR) selection bias, where the probability of a sample being included in the training set depends on observed covariates and the (unobserved) outcome [20]. Each biased subset contains $N = 500$ samples; we simulate systematic under-representation by removing 70% of instances from outcome- and attribute-dependent subgroups: women with high cholesterol who are less willing to seek medical attention (Heart), high-income middle-aged and older individuals reluctant to disclose finances (Bank), women with long working hours who are less willing to report income (Adult), and minority students with high LSAT scores who are less likely to be tracked (Law). To isolate bias effects beyond sample size, we construct a reduced-clean subset by randomly downsampling unbiased data to match size. Results are averaged over five independent runs with different random seeds. We evaluate whether TabPFN is less affected by these MNAR shifts, as its meta-learned prior may promote stable predictive patterns.

4 Results and Discussion

RQ1. Section 4 reports accuracy and fairness (DP and EO w.r.t. different sensitive attributes) on Bank and Heart benchmarks. Fig. 1 show results on Adult and Law benchmarks (DP, EO averaged over race and sex [2]) across training set sizes ranging from 500 to full. FT-TabPFN consistently achieves the highest or near-highest accuracy across all sizes and datasets, reaching 0.99 on both Bank and Heart. TabPFN closely follows in accuracy while demonstrating superior fairness, especially in small-data regimes. For e.g, at 500 samples on Adult, average DP for TabPFN and FT-TabPFN is significantly lower than other baselines. Thus, achieving better fairness and accuracy.

As training size increases beyond TabPFN's previously validated 10k-sample regime, both TabPFN and FT-TabPFN maintain strong accuracy. Fairness metrics, particularly EO, exhibit greater variability, showing moderate increases on Adult and slight fluctuations on Law. But the fairness performance remain at competitive levels.

RQ2. Table 2 and Table 3 summarize results under spurious correlations and MNAR covariate shift. As discussed in Section 3.2, a model that relies on causal mechanisms is expected to show minimal degradation when spurious attributes are flipped. Our results align with this intuition. Under pseudo-correlation perturbations,

[2]The same averaging is applied to the results under covariate shift in Table 3. Complete per-attribute results are available at https://github.com/ql909/An-Empirical-Study-of-TabPFN/tree/main.

Dataset	Model	Acc ↑	DP_1 ↓	EO_1 ↓	DP_2 ↓	EO_2 ↓
Bank	LR	0.95	**0.10**	0.58	**0.05**	0.12
	RF	**0.99**	**0.10**	0.14	0.08	0.09
	MLP	**0.99**	0.12	0.17	0.07	0.12
	TabPFN	**0.99**	**0.10**	**0.09**	0.06	0.14
	FT-TabPFN	**0.99**	0.11	0.10	0.08	**0.08**
Heart	LR	0.84	**0.23**	0.20	0.20	0.12
	RF	0.91	0.31	0.10	0.17	0.07
	MLP	0.90	0.28	0.04	0.22	0.15
	TabPFN	0.96	0.30	0.05	**0.14**	0.06
	FT-TabPFN	**0.99**	0.35	**0.03**	**0.14**	**0.03**

Table 1: Performance on Bank and Heart datasets. ↑ indicates the higher the better, ↓ indicates the lower the better. The same convention applies to the following tables/figures.

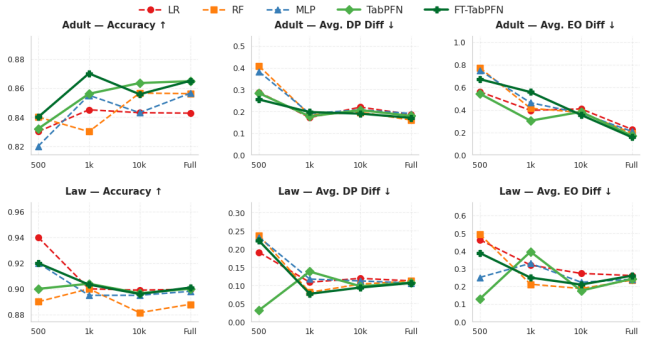

Figure 1: Performance of different models with varying dataset scale on Adult and Law datasets.

FT-TabPFN and TabPFN exhibit superior flip consistency, reaching 0.82–0.97 across datasets and leading the pack on Heart and Bank, indicating robust predictions less sensitive to the reversal of spurious attribute–label associations. RF follows as a strong second in consistency, while MLP and LR lag behind, highlighting their greater reliance on dataset-specific patterns. For Accuracy, FT-TabPFN and TabPFN have the highest Acc on most datasets (Bank 0.97, Heart/Adult 0.75-0.88). Fairness metrics (DP/EO) exhibit inconsistent patterns across models, with TabPFN/FT-TabPFN showing varying DP (Bank/Heart/Law/Adult 0.04-0.08, Adult 0.21-0.25) and EO (Heart/Law/Adult 0.09-0.12, Bank 0.22-0.27) performance.

Notably, the Law dataset exhibits anomalously low consistency (<0.62) and Acc (<0.56) across models post-flip, contrasting its high baseline Acc in Fig. 1 (especially small samples). This flip-induced collapse persists despite strong non-perturbed utility, stemming from extreme imbalance in sensitive attributes (e.g., race/sex) and label distributions that may foster proxy reliance; with anomalies noted in prior work as well for this dataset [9].

Table 3 reports model performance under MNAR-induced covariate shift. TabPFN and FT-TabPFN consistently achieve the highest accuracy on biased (Acc_b: 0.89–0.99) and clean (Acc_c: 0.86–0.99) data across datasets, outperforming baselines by 2–10%. However, fairness improvements are inconsistent: while they attain the lowest EO in Heart and Law, EO remains notably higher in Bank (up to 0.59) and Adult shows only marginal gains. DP is particularly

unstable, ranging from near-optimal in Law/Adult to the highest values in Heart/Bank (0.38–0.42).

Dataset	Model	Consistency↑	Acc↑	DP↓	EO↓
Heart	FT-TabPFN	**0.89**	**0.88**	0.21	**0.10**
	TabPFN	0.87	0.87	0.25	**0.10**
	RF	0.70	0.68	0.29	0.30
	MLP	0.56	0.54	0.10	0.12
	LR	0.51	0.49	**0.07**	0.11
Bank	FT-TabPFN	**0.97**	**0.97**	0.08	0.27
	TabPFN	**0.97**	**0.97**	0.08	0.22
	RF	0.96	0.94	0.07	0.44
	MLP	0.94	0.93	0.05	0.31
	LR	0.93	0.90	**0.02**	**0.08**
Law	FT-TabPFN	0.27	0.21	0.05	**0.11**
	TabPFN	0.27	0.21	**0.04**	0.12
	RF	**0.62**	**0.56**	0.16	0.18
	MLP	0.28	0.23	0.05	**0.11**
	LR	0.25	0.19	**0.04**	0.12
Adult	FT-TabPFN	0.82	0.76	0.06	0.10
	TabPFN	0.82	0.75	**0.05**	0.09
	RF	**0.85**	**0.77**	**0.05**	0.11
	MLP	0.81	0.74	**0.05**	**0.07**
	LR	0.82	0.74	**0.05**	0.08

Table 2: Accuracy (Acc), fairness (DP / EO, averaged over sensitive attributes) under spurious correlation.

Dataset	Model	Acc_b ↑	Acc_c ↑	DP_b ↓	DP_c ↓	EO_b ↓	EO_c ↓
	FT-TabPFN	0.96	0.97	0.42	0.41	0.11	0.06
	TabPFN	**0.97**	**0.98**	0.41	**0.40**	**0.08**	**0.04**
Heart	RF	0.93	0.95	0.41	**0.40**	0.12	0.07
	MLP	0.92	0.94	0.40	0.42	0.13	0.16
	LR	0.86	0.86	**0.38**	0.41	0.13	0.18
	FT-TabPFN	**0.99**	**0.99**	0.09	0.12	**0.15**	0.45
	TabPFN	**0.99**	**0.99**	0.10	0.12	0.18	0.49
Bank	RF	0.97	0.98	0.08	0.11	0.39	0.50
	MLP	0.97	0.98	0.11	0.12	0.58	0.59
	LR	0.93	0.94	**0.05**	**0.07**	0.33	**0.43**
	FT-TabPFN	**0.89**	**0.90**	0.15	0.19	0.33	0.49
	TabPFN	**0.89**	**0.90**	0.15	0.19	0.33	0.49
Law	RF	0.88	0.91	**0.12**	**0.12**	**0.23**	**0.33**
	MLP	0.88	0.91	0.16	0.16	0.35	0.46
	LR	**0.89**	**0.90**	**0.12**	0.14	0.27	0.41
	FT-TabPFN	0.81	0.86	0.23	0.16	0.34	0.29
	TabPFN	0.81	0.86	0.23	0.16	0.34	0.29
Adult	RF	0.81	0.86	**0.16**	**0.12**	0.33	0.29
	MLP	**0.82**	**0.88**	0.25	0.16	**0.32**	**0.25**
	LR	0.81	0.83	0.21	0.13	0.37	**0.25**

Table 3: Model performance under MNAR covariate shift. Acc_b, DP_b, EO_b refer to the biased model (trained on selected data). Acc_c, DP_c, EO_c refer to the clean/unbiased baseline.

5 Conclusion

Our empirical evaluation yields three key findings. First, the TabPFN family achieves superior predictive accuracy across dataset sizes and demonstrates robustness to spurious correlations, yet its fairness improvements are moderate and inconsistent, with group fairness metrics like EO fluctuating with scale. Second, and critically, we uncover a previously overlooked asymmetry: while causal pre-training effectively mitigates spurious statistical shortcuts, it provides limited to no fairness improvement under realistic MNAR selection biases, possibly because causal pre-training cannot compensate for the systematic under-representation of causal data from subpopulations, one of the main sources of unfairness in real-world settings. Third, this finding reveals a fundamental gap: causal pre-training emerges as a promising direction for bridging causal robustness and algorithmic fairness, yet it may fall short in entangled scenarios, such as imbalanced sensitive attributes. Future work could explore integrating causal imputation for missing attributes and developing dynamic fairness regularization.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities.* MIT press.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[4] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*.

[5] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. 2024. Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data. *arXiv preprint arXiv:2411.10634* (2024).

[6] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. [n. d.]. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations, 2023*.

[7] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature* (09 01 2025). doi:10.1038/s41586-024-08328-6

[8] Patrik Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. 2025. Towards Fair In-Context Learning with Tabular Foundation Models. *arXiv preprint arXiv:2505.09503* (2025). https://arxiv.org/pdf/2505.09503v3

[9] Woojin Kim and Hyeoncheol Kim. 2025. Counterfactual Fairness Evaluation of Machine Learning Models on Educational Datasets. arXiv preprint arXiv:2504.11504.

[10] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[11] Wouter M Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806* (2018). https://arxiv.org/pdf/1812.11806

[12] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 1–35. doi:10.1145/3457607

[14] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems* 62 (2014).

[15] UCI Machine Learning Repository. 1988. Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[16] UCI Machine Learning Repository. 1996. Adult (Census Income) Dataset. https://archive.ics.uci.edu/ml/datasets/adult

[17] Jake Robertson, Noah Hollmann, Samuel Müller, Noor Awad, and Frank Hutter. 2025. FAIRPFN: A Tabular Foundation Model for Causal Fairness. *arXiv preprint* (2025). https://arxiv.org/pdf/2506.07049

[18] Minglai Shao, Dong Li, Chen Zhao, Xintao Wu, Yujie Lin, and Qin Tian. 2024. Supervised algorithmic fairness in distribution shifts: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 909, 9 pages. doi:10.24963/ijcai.2024/909

[19] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests. In *International Conference on Learning Representations*.

[20] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 1 (2021), 101–119. doi:10.1007/s41060-021-00259-z

[21] Linda F. Wightman. 1998. *LSAC National Longitudinal Bar Passage Study.*

[22] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious Correlations in Machine Learning: A Survey. *arXiv preprint arXiv:2402.12715* (2024).