Bias Begins with Data: The *FairGround* Corpus for Robust and Reproducible Research on Algorithmic Fairness

Jan Simson JAN.SIMSON@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Alessandro Fabris ALESSANDRO.FABRIS@UNITS.IT
University of Trieste
Trieste, 33170, Italy

Cosima Fröhner C.FROEHNER@CAMPUS.LMU.DE
Department of Statistics, LMU Munich
Munich, 80539, Germany

Frauke Kreuter FRAUKE.KREUTER@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Christoph Kern CHRISTOPH.KERN@LMU.DE
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)
Munich, 80539, Germany

Abstract

As machine learning (ML) systems are increasingly adopted in high-stakes decision-making domains, ensuring fairness in their outputs has become a central challenge. At the core of fair ML research are the datasets used to investigate bias and develop mitigation strategies. Yet, much of the existing work relies on a narrow selection of datasets–often arbitrarily chosen, inconsistently processed, and lacking in diversity–undermining the generalizability and reproducibility of results.

To address these limitations, we present *FairGround*: a unified framework, data corpus, and Python package aimed at advancing reproducible research and critical data studies in fair ML classification. FairGround currently comprises 44 tabular datasets, each annotated with rich fairness-relevant metadata. Our accompanying Python package standardizes dataset loading, preprocessing, transformation, and splitting, streamlining experimental workflows. By providing a diverse and well-documented dataset corpus along with robust tooling, FairGround enables the development of fairer, more reliable, and more reproducible ML models. All resources are publicly available to support open and collaborative research.

Keywords: algorithmic fairness, dataset collections, dataset usage

1 Introduction

The field of algorithmic fairness has grown rapidly, reflecting the increasing recognition of fairness as a core concern in machine learning (Mehrabi et al., 2021; Pessach and Shmueli, 2023). Progress in this field is inevitably tied to data as the central ingredient to developing, testing and benchmarking more equitable algorithms. Given that these algorithms and fairness-enhancing techniques are often deployed in high-risk contexts [e.g., healthcare (Obermeyer et al., 2019; Barda et al., 2020), criminal justice (Angwin et al., 2016; Carton et al., 2016), jobseeker profiling (Kern et al., 2024; Achterhold et al., 2025)], systematic and transparent evaluations based on principled rather than ad-hoc selections of datasets are critical to understand which method works reliably under which conditions and which might not yet be ready for deployment.

Progress in Fair ML is challenged by (1) opacity in data practices and (2) critical limitations of the most prominent datasets currently used. A number of studies have shown that seemingly minor data processing and algorithmic design choices can significantly impact fairness outcomes, raising important questions about the robustness and generalizability of existing fairness interventions (Simson et al., 2024b; Friedler et al., 2019; Caton et al., 2022). Compounding these concerns, recent work has also highlighted reproducibility challenges that hinder consistent evaluation across settings (Cooper et al., 2024; Simson et al., 2024a). Furthermore, large-scale comparisons of fairness algorithms not only show strong sensitivity to data processing decisions, but also considerable performance differences between datasets, underlining the importance of the exact collection of data used for benchmarking and evaluation (Agrawal et al., 2021). Unfortunately, current studies commonly still focus on a narrow set of benchmark datasets–such as *Adult* (Kohavi, 1996), *COMPAS* (Angwin et al., 2016) and *German Credit* (Hofmann, 1994)–which suffer from known limitations, including contrived prediction tasks, noisy data, and severe coding mistakes (Ding et al., 2021; Bao et al., 2022; Grömping, 2019a). Taken together, these practices can lead to evaluations of fairness algorithms that are driven by methodological artifacts rather than representing reliable performance tests that justify the (non-)deployment of a given method in practice.

Addressing these limitations, this work introduces *FairGround*: a framework that emphasizes reproducible data processing pipelines, standardized evaluation protocols, and diverse collections of datasets tailored to specific needs (Figure 1). Our corpus contains 136 scenarios, i.e. combinations of 44 tabular datasets with available sensitive attributes. Each dataset comes with rich metadata (35 annotated and 27 computed meta-features), which allows for a principled selection of benchmarking collections and for failure testing of algorithms to identify data scenarios under which a proposed method struggles to perform. We further provide a data selection algorithm and associated collections of datasets that are small but diverse, i.e., present challenging scenarios with data that capture the variability present in the larger corpus. Our Python package facilitates transparent data practices in fair ML through reproducible and standardized, but customizable, processing pipelines. With FairGround, we contribute infrastructure that supports more robust and generalizable evaluation of fairness-aware machine learning methods.

2 Related work

2.1 Comparing fairness-enhancing algorithms

A number of prior studies have carried out systematic comparisons of fairness-enhancing algorithms across different settings (Friedler et al., 2019; Agrawal et al., 2021; Biswas and Rajan, 2020; Cruz and Hardt, 2024; Defrance et al., 2024; Han et al., 2023; Hort et al., 2021; Islam et al., 2021; L. Cardoso et al., 2019). While these comparative efforts have contributed valuable insights, they are often constrained by a narrow and inconsistently chosen set of benchmark datasets. In many cases, dataset selection is neither well-documented nor critically examined, resulting in evaluations that are difficult to reproduce and limited in scope.

The broader field continues to face fundamental challenges related to reproducibility and transparency in experimental design (Simson et al., 2024a; Cooper et al., 2024). One prominent issue is the lack of principled approaches to dataset processing and selection. Many existing works make ad hoc or arbitrary choices when selecting datasets (Ding et al., 2021; Bao et al., 2022; Grömping, 2019a), often relying on convenience or popularity rather than representativeness or relevance. These decisions can unintentionally bias results and restrict the generalizability of conclusions. A core concern here is that the datasets typically used in fairness evaluations do not adequately reflect the diversity and complexity of real-world deployment scenarios. The dominance of a small set of benchmark datasets has led to evaluations that cover only a limited subset of the problem space fairness algorithms are meant to address (Fabris et al., 2022).

Compounding this, there remains little clarity around the specific data conditions under which fairness methods are expected to succeed or fail. Without a systematic understanding of these contexts, practitioners are left with limited guidance on which algorithms to apply in practice (Richardson et al., 2021; Holstein et al., 2019), reducing the effectiveness and reliability of fairness interventions in real-world systems.

2.2 Fairness toolkits and data studies

Fairness datasets have been examined from both granular and comparative perspectives. Some works offer deep, dataset-specific critiques (Bandy and Vincent, 2021; Ding et al., 2021; Bao et al., 2022; Birhane et al., 2023), while others survey broader patterns across multiple datasets (Crawford and Paglen, 2021; Fabbrizzi et al., 2022; Fabris et al., 2022; Zhao et al., 2024). In parallel, fairness-focused toolkits such as AIF360 (Bellamy et al., 2018), Fairlearn (Weerts et al., 2023), and Aequitas (Jesus et al., 2024) implement popular algorithmic interventions and metrics, providing an accessible entry point for numerical comparisons—while including only a few illustrative datasets (Table A3). Despite overlapping goals, these two strands have remained largely disconnected. Toolkits often treat datasets as ancillary components and dataset-focused studies fail to produce machine-readable resources designed for seamless integration with software frameworks. Bridging critical data studies and fairness toolkits is essential for advancing the field, as meaningful integration can enable more rigorous, interpretable, and reproducible fairness evaluations–particularly by linking dataset properties to the behavior and impact of fairness interventions (Li et al., 2022; Favier et al., 2023).

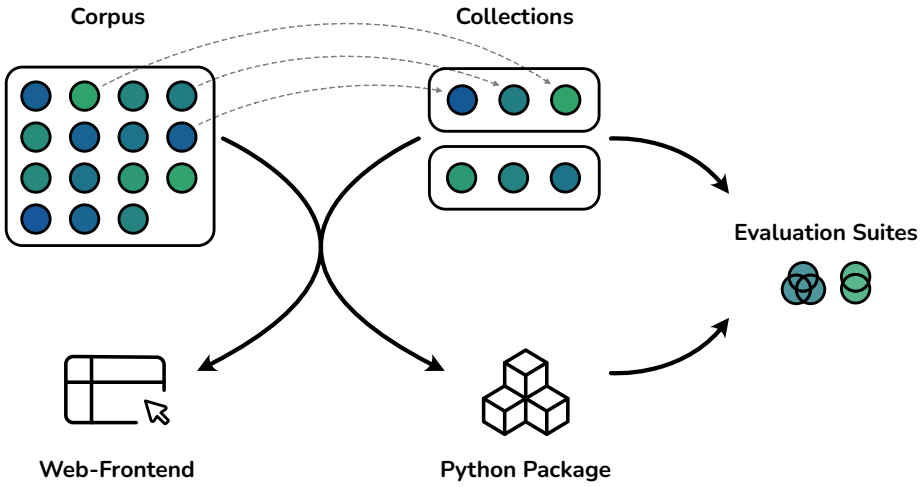

Figure 1: **The different components in the FairGround corpus.** We provide a comprehensive corpus of datasets and extract diverse collections of datasets via a selection algorithm. Both the corpus, collections and individual datasets are made accessible via a Python package and web-frontend. Collections paired with reproducible dataset loading and preparation allows for novel evaluation suites.

A recent article, most closely related to ours, lists several fairness resources and provides a tool for data fetching, but does not address integrated processing or annotation pipelines (Hirzel and Feffer, 2023). In this paper, we address this gap by introducing a benchmark suite that combines (1) a curated corpus of datasets accompanied by rich quantitative and qualitative annotations, (2) reproducible data fetching and processing pipelines, and (3) standardized collections and evaluation protocols. Our annotations provide a foundation for aligning datasets with fairness-aware methods in a consistent, reproducible, and extensible manner.

3 Framework

We introduce a unified framework of resources designed to support reproducible research and critical data studies in fair ML. While our current implementation focuses on tabular classification, which is prominent in fair ML research Mehrabi et al. (2021); Caton and Haas (2024), the underlying design is broadly applicable to other contexts.

3.1 Corpus

Building on and extending beyond prior surveys of datasets in fair ML research (Fabris et al., 2022; Le Quy et al., 2022), we compile a curated corpus of $N = 44$ tabular datasets. Each dataset is annotated with extensive fairness-relevant metadata, both quantitative and qualitative. While an additional 11 datasets were partially annotated, we excluded them from the final release due to issues such as dubious provenance or access restrictions (details in Section C.7).

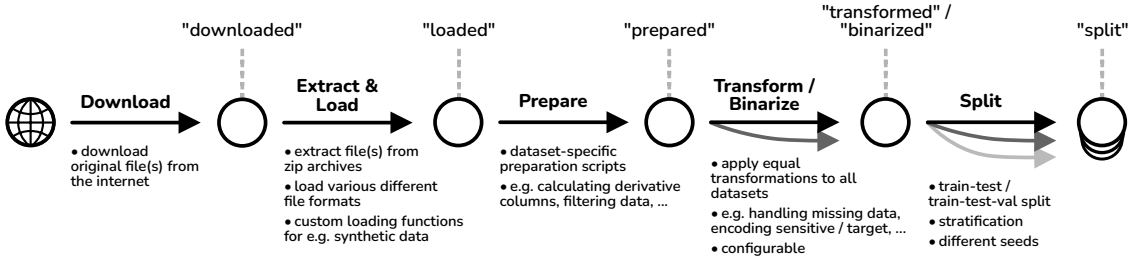

Figure 2: **The pipeline of steps involved when loading and processing a dataset in the package.** Datasets can be accessed / exported after each of the steps in the pipeline and most steps allow for configuration.

The corpus spans a wide range of dataset sizes, from 118 to over 3.2 million records, and 4 to 1,941 features. Most datasets originate from domains such as economics and law (each 23.4%), followed by finance (12.7%) and education (10.7%). Geographical representation is notably skewed: nearly 60% of datasets originate from the United States, with limited coverage from other regions (see Tables A and A2 for details). The dataset metadata can be explored interactively at: `https://reliable-ai.github.io/fairground/`.

Following prior work (Fabris et al., 2022; Le Quy et al., 2022), we annotate each dataset with contextual information (e.g., dataset name, domain), data-specific attributes (e.g., geography, time period), and technical metadata required for loading and preparing the data. Where multiple variants of a dataset exist, each version is treated as a distinct entry (cf. Section C.4). We also provide annotations relevant to fair ML tasks, including sensitive attribute selection, target variable definitions, and required preprocessing. While we do not claim our annotations are definitive, they serve as principled defaults that make implicit modeling decisions explicit, encouraging transparency in fair ML research (Simson et al., 2024a). Full details on our annotation procedure are provided in Sections C.4 and C.5.

In addition to manual annotations, we compute a range of metadata to support dataset selection, benchmarking, and critical analysis. This includes structural properties (e.g., missing values, feature types), statistical characteristics (e.g., bivariate correlations, sensitive AUC), and fairness-related properties (e.g., protected group prevalence, base rates, Gini-Simpson index) (Brzezinski et al., 2024; Mecati et al., 2022; Holland et al., 2020). These computed metadata features are detailed in Appendix C.6 and integrated into our Python tooling for streamlined access.

3.2 Infrastructure

To enable reproducible and scalable use of the corpus, we provide a Python package that operationalizes our framework. This package automates dataset acquisition, preprocessing, transformation, and splitting, applying the annotations to prepare datasets for downstream fair ML tasks (Figure 2). The package supports diverse data formats and includes default transformations, such as standard feature selection, handling of missing values and encoding

of sensitive attributes. We re-emphasize that defaults are not intended as universally correct, but rather as transparent baselines that can be fully customized. By surfacing and standardizing preprocessing decisions, the package encourages methodological rigor and reduces hidden variability in experimental pipelines (Simson et al., 2024b).

In particular, FairGround supports the following transformations to export data in a readily usable format. Users can choose to retain either the complete set of columns in a dataset or only the essential subset, which includes frequently-used features, sensitive attributes, and the target variable (default). To handle missing values, the framework supports three options: dropping the entire column, removing only rows with missing values, or imputing missing values using the median (default for numerical) or a placeholder value (default for categorical). The target variable can be binarized in several ways: based on an annotated preferable label, redefined to reflect a majority/minority split, or automatically selected between these options depending on metadata availability (default is based on the preferable label if provided). When multiple sensitive attributes exist, users can keep them separate (default) or combine them into a single binary attribute that captures their intersection (default in the binarized setting). Sensitive attribute values can be left unchanged or grouped into majority and minority categories (again, grouping is the default in binarized datasets). For categorical features, FairGround supports either leaving them as-is or converting them into binary indicators via dummy encoding (default). To control for high cardinality in categorical or text fields, the package applies an optional limit–by default, restricting each categorical or text column to a maximum of 200 unique values, with less frequent categories grouped together once this limit is exceeded.

The package also supports automatic metadata extraction (see Section 3.1). Importantly, we avoid redistributing raw data directly to respect licensing constraints and datasets are instead downloaded from their original sources and optionally cached locally.
The package is open source and available at: `https://github.com/reliable-ai/fairground`
Releases are archived on Zenodo: `https://doi.org/10.5281/zenodo.17288596`
Package installation: `pip install fairml-datasets`
Package documentation: `https://reliable-ai.github.io/fairground/docs/`
Code examples are provided in Appendix C.3.

In parallel, we release an interactive website that allows browsing the dataset corpus, metadata, and example usage. The site also offers sample code for specific datasets and is available at `https://reliable-ai.github.io/fairground/`.

3.3 Collections

To further support reproducible benchmarking and targeted experimentation, we define several curated dataset collections derived from the full corpus with an extensible algorithm. These include: two collections (small and large) optimized for diversity in algorithmic performance; three collections with permissive licenses; and three collections emphasizing geographic diversity (Tables A5–A7).

Combined with standardized data splits from our package, which are critical to fair ML reproducibility (Friedler et al., 2019), these collections provide ready-to-use evaluation suites for fair ML development.

4 Experiments

Leveraging the full FairGround dataset corpus, we conduct a series of experiments to systematically investigate the extent to which the choice of dataset influences the evaluation and observed performance of fairness-aware machine learning methods.

To reflect common practice in fairness research and enable broad coverage of methodological approaches, we evaluate a representative set of fairness-aware debiasing techniques spanning the three main intervention stages in the ML pipeline: *pre-processing*, *in-processing*, and *post-processing*. Specifically, we compare the following seven algorithms: *Learning Fair Representations* (pre) (Zemel et al., 2013), *Disparate Impact Remover* (pre) (Feldman et al., 2015), *Adversarial Debiasing* (in) (Zhang et al., 2018), *Meta-Algorithm* (in) (Celis et al., 2019), *Rich Subgroup Fairness / GerryFair* (in) (Kearns et al., 2018), *Grid Search Reduction* (in) (Agarwal et al., 2018), and *Group-Specific Thresholds* (post) (Hardt et al., 2016). We use logistic regression as a standard model for pre- and post-processing.

To satisfy the input constraints of all methods, datasets were converted to binarized numerical representations using the default transformation settings provided by our accompanying Python package (see Section 3.2). This ensures compatibility while preserving consistency across experiments.

Given that most fairness techniques are designed to optimize fairness with respect to a single sensitive attribute, we adopt a principled approach to define sensitive attribute configurations. For datasets containing fewer than four sensitive attributes, we evaluate all individual attributes and their pairwise intersections. For datasets with four or more sensitive attributes, we restrict evaluation to individual attributes to avoid combinatorial complexity. We refer to each combination of a dataset and its corresponding sensitive attribute selection as a unique ***scenario***.

We apply each of the seven processing methods and a baseline to each of the $n = 136$ datasets and sensitive attribute combinations (scenarios) across five separate seeds and train-test splits. This results in a total of $N = 5440$ different models that are trained and compared. For each model, we compute two commonly used measures of performance (*Balanced Accuracy*, Eq. 1; *F1 Score*, Eq. 2) and two measures of algorithmic fairness (*Equalized Odds Difference*, Eq. 3; *Demographic Parity Difference*, Eq. 4). The computational infrastructure (Section C.8) and software (Section C.9) used for experiments are described in the technical appendix.

4.1 Results

The experiments reveal substantial variation in both fairness and performance metrics across datasets and methods. F1 score, equalized odds difference, and demographic parity difference span the full $[0, 1]$ range, while balanced accuracy varies from approximately 0.2 to 1.0. To facilitate comparisons, we compute delta scores—metric differences relative to a logistic regression baseline without fairness interventions (Eq. 5). Figure 3 illustrates this calculation for one dataset, scenario, seed, and metric, with dashed lines indicating differences from the baseline.

The overall distribution of delta scores across all four metrics is shown in Figure B1. Importantly, fairness interventions often lead to minor deviations in scores, as highlighted by the large gray bar indicating an absolute change of ≤ 0.01, which correspond to scenarios

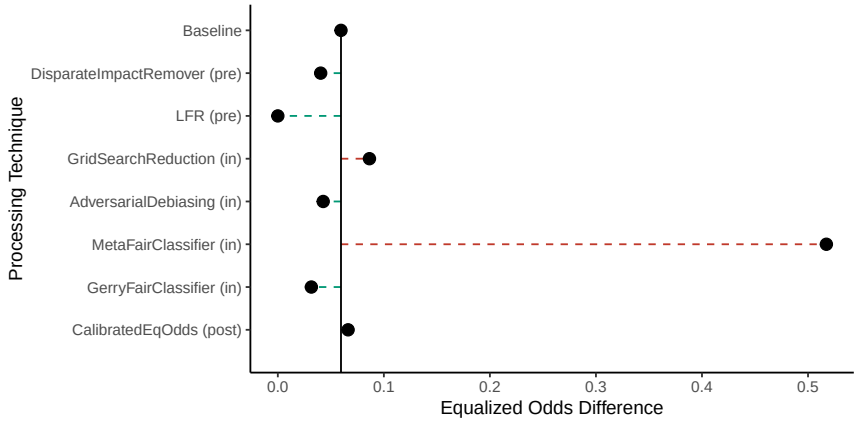

Figure 3: Scores from a single dataset (*Bank*), scenario (sensitive attribute: *Age*), seed (*80539*), and metric (*Equalized Odds Difference*) illustrating how delta scores with respect to baseline logistic regression are calculated. Delta scores correspond to dashed lines.

where popular fair ML methods are ineffective. A sizable portion produces meaningful differences, typically reflecting the well-known tradeoff between fairness and performance (Menon and Williamson, 2018; Islam et al., 2021): improvements in fairness often coincide with declines in predictive accuracy.

4.2 Rankings of Debiasing Techniques are not Stable

To reflect how practitioners might compare processing techniques in practice, we analyze the relative rankings of different methods. While some methods–such as *LFR*, *Grid Search Reduction*, and *Adversarial Debiasing*–tend to rank favorably, their positions vary considerably across scenarios, and no single method consistently outperforms the rest (Figure 4). High-performing methods often come with caveats. For instance, *LFR* occasionally fails due to convergence issues or label collapse during rebalancing, rendering it unusable in some cases. *Adversarial Debiasing* often presents sharp tradeoffs between fairness and predictive performance. These variations are influenced by the dataset and scenario characteristics.

4.3 Identifying Important Dataset Characteristics

To uncover which dataset properties affect method performance, we train simple machine learning models (random forests (Ho, 1995)) for each debiasing technique. These models use only computed metadata (Sections 3.1, C.6) to predict method effectiveness across individual scenarios. As shown in Figure B3, they capture substantial variance in observed outcomes. We analyze feature importance scores from these models to assess which dataset characteristics matter most. Figure 5 displays importances for predicting *Equalized Odds Difference*. A key trend is that the predictability of sensitive features from non-sensitive ones (`meta_sens_predictability_roc_auc`, top row) is influential across all methods. Base rate differences are critical for some techniques but negligible for others. These metadata-derived

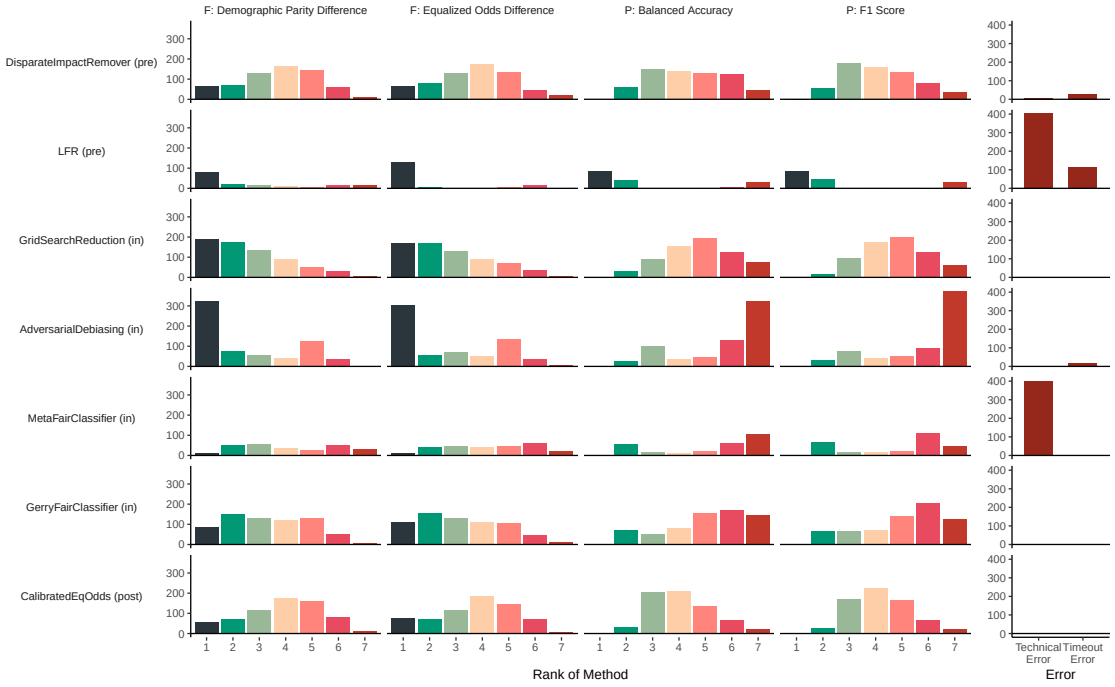

Figure 4: **Relative performance and efficacy of different fairness interventions is highly variable.** Relative ranking of different processing techniques across datasets and seeds (A), as well as prevalence of practical and timeout errors (B).

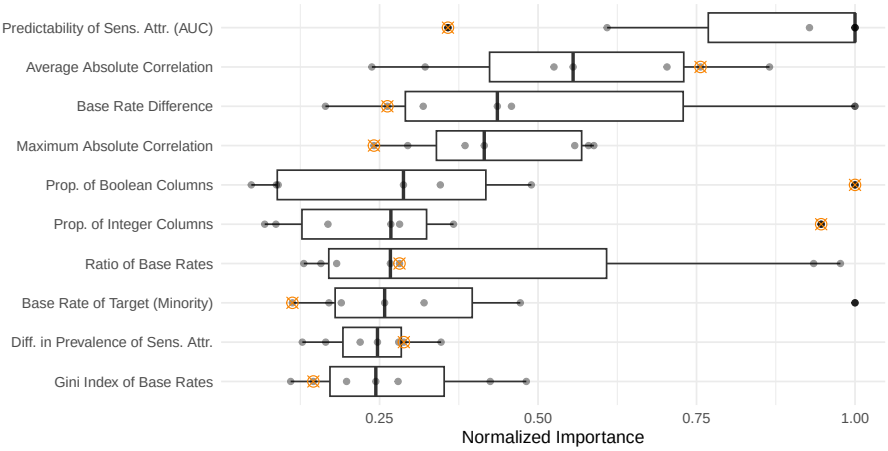

Figure 5: **The importance of different dataset characteristics can be highly variable between debiasing algorithms.** Normalized feature importance of the 10 most important computed metadata features to predict the difference in *Equalized Odds Difference* across all processing methods, ordered by average importance. Feature importance for *Adversarial Debiasing (in)* is highlighted in orange.

features help characterize the conditions under which fairness interventions are likely to succeed. Notably, *Adversarial Debiasing*, highlighted in orange, relies less on sensitive attribute predictability and more on structural features such as the proportion of boolean and integer columns. Relative importances for other metrics appear in Figure B2.

4.4 Developing Diverse Collections of Datasets

Evaluating fairness interventions across all possible datasets and scenarios is ideal but rarely feasible due to practical constraints like limited compute. To address this, we construct eight curated dataset collections, each optimized for a specific purpose. We use a principled algorithm to construct subsets of scenarios that exhibit diverse properties. We explicitly target predictive accuracy and fairness properties by building a collection of scenarios whose pairwise spearman correlations of delta scores (Eq. 5), across *Balanced Accuracy* (Eq. 1), *F1 Score* (Eq. 2), *Equalized Odds Difference* (Eq. 3) and *Demographic Parity Difference* (Eq. 4) are as low as possible. The underlying assumption is that datasets where debiasing techniques yield divergent fairness-performance tradeoffs make for more informative and challenging benchmarks. The algorithm greedily builds collections by adding the least correlated scenario while fulfilling optional secondary constraints, including selecting only a single scenario per dataset. The algorithm supports two different cutoff values, providing either a fixed number of k scenarios or a fixed upper bound for dataset correlation ($\bar{r}_{j\mathcal{C}} < \tau$) when added to the collection. The algorithm is described in detail in Section C.2. We use this selection process both to construct benchmark collections and to define default scenarios per dataset. We demonstrate how the *FairGround* corpus as well as its collections exhibit higher diversity in algorithm performance compared to other dataset collections (Table A4).

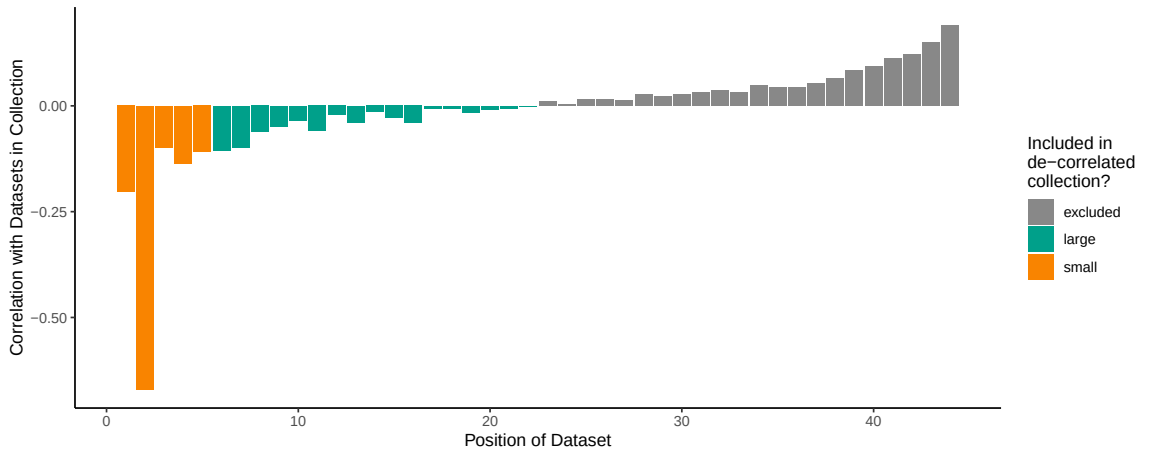

Figure 6: **A large number of negatively intercorrelated datasets is available for collection creation.** Average Spearman correlation of delta scores between the scenarios already in the collection and candidate scenarios at the time they are added to the collection. The very first scenario minimizes the average correlation with all other scenarios.

De-Correlated Datasets We construct two benchmark collections using the correlation-based algorithm with cutoffs $k = 5$ and $\tau < 0$, yielding sets of $n = 5$ and $n = 22$ scenarios, respectively (Table A5). A UMAP projection (McInnes et al., 2018) from the high-dimensional space of computed metadata (Figure 7) confirms that selected datasets span a wide range of characteristics.

Permissively Licensed Datasets To facilitate open sharing and reuse, we build three collections containing only datasets with permissive licenses. We construct these collections by only allowing datasets to be added to the collection which (1) have licensing information available and (2) are permissively licensed (e.g. Creative Commons, Apache, GNU licenses). One collection uses a fixed $k = 5$ cutoff, one uses a $\tau < 0$ threshold ($n = 16$), and one includes all permissively licensed datasets without filtering ($n = 32$). All three are listed in Table A6. We release these datasets in both prepared and binarized formats.

Geographically Diverse Datasets To address regional bias, we create three collections ensuring that no two datasets originate from the same country. We apply the selection algorithm with constraints and cutoffs of $k = 5$ and $\tau < 0$ ($n = 6$), as well as an unfiltered collection ($n = 10$) (Table A7). While this offers greater geographic diversity than is typical in ML fairness benchmarks, it remains insufficient. As prior work has emphasized (Septiandri et al., 2023; Mihalcea et al., 2025), future data efforts must expand beyond WEIRD contexts while carefully balancing this goal with ethical data practices.

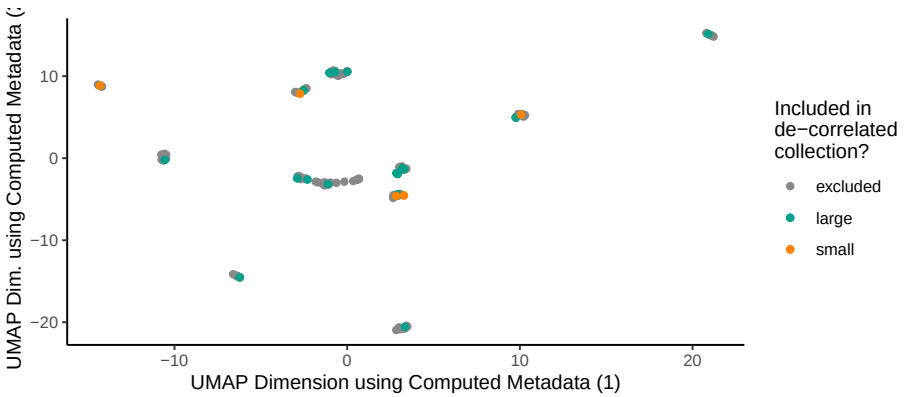

Figure 7: **Datasets in the de-correlated collections capture variability in the computed metadata features well.** Two dimensional mapping of datasets using UMAP on computed metadata features. Scenarios in the de-correlated collections are highlighted in different colors.

5 Limitations

While this work takes a substantial step toward improving reproducibility and empirical rigor in fair ML, it also operates within known constraints. Benchmarking, particularly in fairness research, can risk oversimplifying complex sociotechnical issues. Fairness cannot be fully captured by metrics or solved solely through optimization, and responsible development and evaluation of fair ML requires critical engagement with the broader context.

Our preprocessing and annotation decisions are not intended as universally optimal; their suitability depends on the specific dataset and use case. The experimental results presented here are illustrative rather than prescriptive—they demonstrate the kinds of analyses our corpus enables but are not meant to be definitive benchmarks.

Importantly, our dataset corpus is designed to be dynamic. Gaps in representation, especially with respect to geographic and demographic diversity, remain. We explicitly encourage community contributions of new datasets to help close these gaps (cf. Section 3.1). To support this, we provide a modular, versioned Python package that ensures transparency and reproducibility as the corpus evolves.

While our current focus is on tabular classification–a core setting in fair ML research–our framework is general. LLM evaluations, for example, may also benefit from the current tabular corpus through approaches such as folktexts (Cruz et al., 2025). In future work, we aim to extend our methodology and infrastructure to other data modalities, including text and image domains.

6 Discussion

We introduce *FairGround*, a comprehensive framework, dataset corpus, and Python package developed to address long-standing challenges in fair ML research. By curating a diverse collection of 44 tabular datasets, encompassing 136 scenarios, and providing fairness-relevant

metadata and reproducible preprocessing tools, FairGround enables transparent, rigorous, and extensible experimentation. The accompanying Python package supports reproducibility by exposing (and providing defaults for) key data processing decisions. We demonstrate its utility through a large-scale case study, illustrating how the framework facilitates robust comparative evaluations of debiasing techniques. Specifically, we show how our provided data collections better reflect the diverse performance of debiasing algorithms in comparison to collections currently used in fair ML research, while enabling new fairness analyses by connecting algorithm performance to dataset characteristics.

The significance of this work extends beyond its immediate technical contributions. By foregrounding the role of data infrastructure, FairGround highlights how dataset design, composition, and documentation fundamentally shape research trajectories and outcomes. These elements influence algorithmic behavior, reproducibility, and downstream system impact–making them critical to both scientific rigor and ethical responsibility.

Our framework is designed not only to support method development but also to position datasets as first-class research objects. It prompts researchers to interrogate representational biases, data provenance, and the implications of dataset selection–core concerns for equitable and socially responsible AI. In doing so, FairGround fosters deeper engagement with the sociotechnical dimensions of ML, encouraging reflection on how benchmarks reflect and reinforce power structures.

Additionally, FairGround lays essential groundwork for linking dataset characteristics to model fairness outcomes. This connection has important implications for anti-discrimination policy and regulation. For instance, under the EU AI Act, high-risk AI systems are subject to strict data governance requirements, including the obligation to assess datasets for bias and representational gaps (European Parliament, 2024). The metadata and fairness-relevant characteristics computed within FairGround can serve as a foundation for quantitative dataset documentation aligned with these legal mandates.

Broader Impact Statement

Our work aims to improve data practices in the field of algorithmic fairness, which in turn can lead to more robust and reproducible research, better and more ethical handling of datasets and increased transparency around dataset usage. By highlighting and quantifying the lack of geographic representation in popular datasets, we hope our work inspires the collection of novel and geographically diverse datasets. These positive changes have the possibility of affecting practices beyond research, ideally leading to the deployment of better and fairer algorithmic decision making and ML systems in production settings. Beyond the field of algorithmic fairness the *FairGround* framework provides a template for other fields to start developing dataset corpora and collections.

While this work encourages better data practices, there is a risk of it contributing to a benchmarking culture overly focused on quantitative and superficial notions of fairness, which we explicitly want to warn against. While it is important to use a diverse collection of datasets for evaluation, it is equally important, especially in applied contexts, to be aware of the sociotechnical context (ML) systems are developed and deployed in.

Acknowledgments and Disclosure of Funding

We would like to thank F. Weber and A. Szimmat for their help in the annotation process. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research and the Munich Center for Machine Learning (MCML).

This work is supported by BERD@NFDI and the Simons Institute for the Theory of Computing at the University of California, Berkeley.

References

Eva Achterhold, Monika Mühlböck, Nadia Steiber, and Christoph Kern. Fairness in algorithmic profiling: The AMAS case. *Minds and Machines*, 35(1):9, 2025.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Agency for Healthcare Research and Quality. Medical expenditure panel survey (meps), 8 2018. URL `https://www.ahrq.gov/data/meps.html`.

Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. Debiasing classifiers: is reality at variance with expectation?, 2021. URL `https://arxiv.org/abs/2011.02407`.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, Published on May 23, 2016.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 206–214, 2021.

Jack Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks, April 2022.

Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature communications*, 11(1):4439, 2020.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in neural information processing systems*, 36:21268–21284, 2023.

Sumon Biswas and Hridesh Rajan. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pages 642–653, New York, NY, USA, November 2020. Association for Computing Machinery. ISBN 978-1-4503-7043-1. doi: 10.1145/3368089.3409704.

Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinskyi. Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, 2024.

Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 67–76, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939698. URL https://doi.org/10.1145/2939672.2939698.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):166:1–166:38, 2024. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Simon Caton, Saiteja Malisetty, and Christian Haas. Impact of Imputation Strategies on Fairness in Machine Learning. *Journal of Artificial Intelligence Research*, 74, September 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13197.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Chicago Data Portal. Strategic subject list - historical, 2020. URL https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np.

Consumer Financial Protection Bureau. Home mortgage disclosure act (hmda) data, September 2022. URL https://www.consumerfinance.gov/data-research/hmda/.

A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and social prediction: The confounding role of variance in fair classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22004–22012, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i20.30203.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

Kate Crawford and Trevor Paglen. Excavating AI: the politics of images in machine learning training sets, 2021. URL `https://excavating.ai/`.

André F. Cruz and Moritz Hardt. Unprocessing Seven Years of Algorithmic Fairness, March 2024.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. ABCFair: an adaptable benchmark approach for comparing fairness methods. *Advances in Neural Information Processing Systems*, 37:40145–40163, December 2024.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. page 13, 2021. doi: 10.48550/ARXIV.2108.04884.

Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Martin Durant. fastparquet: A python interface to the parquet file format. https://pypi.org/project/fastparquet/.

European Parliament. Artificial intelligence act. `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf`, 2024.

Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, September 2022. ISSN 1573-756X. doi: 10.1007/s10618-022-00854-z.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. How to be fair? A study of label and selection bias. *Machine Learning*, 112(12):5081–5104, 2023. doi: 10.1007/S10994-023-06401-1. URL `https://doi.org/10.1007/s10994-023-06401-1`.

Jake Fawkes, Nic Fishman, Mel Andrews, and Zachary Lipton. The fragility of fairness: Causal sensitivity analysis for fair machine learning. *Advances in Neural Information Processing Systems*, 37:137105–137134, 2024.

Elaine Fehrman, Vincent Egan, and Evgeny Mirkes. Drug consumption (quantified). UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5TC7S.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338. Association for Computing Machinery, January 2019. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287589.

Ulrike Grömping. South german credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, 4:2019, 2019a.

Ulrike Grömping. South german credit. UCI Machine Learning Repository, 2019b. DOI: https://doi.org/10.24432/C5X89F.

H. Guvenir, Burak Acar, Haldun Muderrisoglu, and R. Quinlan. Arrhythmia. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5BS32.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *The Twelfth International Conference on Learning Representations*, October 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. URL `https://arxiv.org/abs/1610.02413`.

Martin Hirzel and Michael Feffer. A suite of fairness datasets for tabular classification. *arXiv preprint arXiv:2308.00133*, 2023.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282. IEEE, 1995.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and

Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 600. ACM, 2019. doi: 10.1145/3290605.3300830. URL https://doi.org/10.1145/3290605.3300830.

Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pages 994–1006, New York, NY, USA, August 2021. Association for Computing Machinery. ISBN 978-1-4503-8562-6. doi: 10.1145/3468264.3468565.

Rashidul Islam, Shimei Pan, and James R. Foulds. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462614.

Shahin Jabbari, Han-Ching Ou, Himabindu Lakkaraju, and Milind Tambe. An empirical study of the trade-offs between interpretability and fairness. In *ICML Workshop on Human Interpretability in Machine Learning, International Conference on Machine Learning (ICML)*, 2020.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease data set. UCI Machine Learning Repository, 1988. URL https://archive.ics.uci.edu/ml/datasets/heart+Disease.

Sérgio Jesus, Pedro Saleiro, Beatriz M Jorge, Rita P Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani, et al. Aequitas Flow: Streamlining fair ML experimentation. *arXiv preprint arXiv:2405.05809*, 2024.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

Christoph Kern, Ruben Bach, Hannah Mautner, and Frauke Kreuter. When small decisions have big impact: Fairness implications of algorithmic profiling schemes. *ACM Journal on Responsible Computing*, 1(4), November 2024. doi: 10.1145/3689485. URL https://doi.org/10.1145/3689485.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*, 2020. URL https://www.tidymodels.org.

Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 437–444,

New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314262.

Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. arXiv preprint arXiv:1907.01439, 2019.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Nianyun Li, Naman Goel, and Elliott Ash. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2022.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 31, 2018.

Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. *Advances in neural information processing systems*, 33: 18445–18456, 2020.

Charlie Marsh. uv: An extremely fast python package and project manager, 2024. URL `https://github.com/astral-sh/uv`.

Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. Fairness and missing values. *arXiv preprint arXiv:1905.12728*, 2019.

Norman Matloff and Wenxi Zhang. A novel regularization approach to fair ml. *arXiv preprint arXiv:2208.06557*, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Wes McKinney. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, 445:51–56, 2010.

Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. Detecting risk of biased output with balance measures. *ACM J. Data Inf. Qual.*, 14(4):25:1–25:7, 2022. doi: 10.1145/3530787. URL `https://doi.org/10.1145/3530787`.

Mariachiara Mecati, Marco Torchiano, Antonio Vetrò, and Juan Carlos De Martin. Measuring imbalance on intersectional protected attributes and on target variable to forecast unfair classifications. *IEEE Access*, 11:26996–27011, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, January 2018.

Weiwen Miao. Did the results of promotion exams have a disparate impact on minorities? Using statistical evidence in Ricci v. DeStefano. *Journal of Statistics Education*, 18(3), 2010.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. Why AI is WEIRD and shouldn't be this way: Towards AI for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670, 2025.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

New York City Police Department. The nypd stop, question, and frisk database, 2012. URL `https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page`.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019. doi: 10.1126/science.aax2342. URL `https://www.science.org/doi/abs/10.1126/science.aax2342`.

Thomas Lin Pedersen. *patchwork: The Composer of Plots*, 2024. URL `https://CRAN.R-project.org/package=patchwork`. R package version 1.3.0.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL `https://doi.org/10.1145/3494672`.

Dana Pessach and Erez Shmueli. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer, 2023.

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019. URL `https://www.python.org/`. Python version 3.10.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

Vladislav Rajkovic. Nursery. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5P88W.

Karthik Ram and Hadley Wickham. *wesanderson: A Wes Anderson Palette Generator*, 2023. URL `https://CRAN.R-project.org/package=wesanderson`. R package version 0.3.7.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ML toolkits. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 236:1–236:13. ACM, 2021. doi: 10.1145/3411764.3445604. URL `https://doi.org/10.1145/3411764.3445604`.

Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International conference on machine learning*, pages 8316–8325. PMLR, 2020.

Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. WEIRD FAccTs: How western, educated, industrialized, rich, and democratic is FAccT? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 160–171, 2023.

Jan Simson. Multiversum: A helper package to conduct multiverse analyses in python, 2024. URL `https://github.com/jansim/multiversum`.

Jan Simson, Alessandro Fabris, and Christoph Kern. Lazy data practices harm fairness research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 642–659, New York, NY, USA, June 2024a. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658931.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1305–1320, New York, NY, USA, June 2024b. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658974.

Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly*, 38(4):101619, 2021.

Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR, 2019.

Yanchen Wang and Lisa Singh. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119, 2021.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

Austin Wehrwein. *awtools: misc tools and themes for austinwehrwein.com*, 2025. URL `https://github.com/awhstin/awtools`. R package version 0.2.1.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Linda F. Wightman. Lsac national longitudinal bar passage study. 1998.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Dora Zhao, Morgan Klaus Scheuerman, Pooja Chitre, Jerone Theodore Alexander Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen H. Pine, and Alice Xiang. A taxonomy of challenges to curating fair datasets. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/b142e78db191e19b17e60c1425a28b52-Abstract-Datasets_and_Benchmarks_Track.html`.

Appendix A. Supplementary Tables

This section includes supplementary tables that provide additional information supporting the results presented in the main text.

Table A1: Overview of datasets in the corpus. Row and column counts apply to the prepared data prior to further transformations.

	Name and Citation	Rows	Columns	License
1	Adult (Kohavi, 1996)	32,560	15	CC BY 4.0
2	Arrhythmia (Guvenir et al., 1998)	451	280	CC BY 4.0
3	Bank (additional + full) (Moro et al., 2014)	41,188	21	CC BY 4.0
4	Bank (additional) (Moro et al., 2014)	4,119	21	CC BY 4.0
5	Bank (full) (Moro et al., 2014)	45,211	17	CC BY 4.0
6	Bank (Moro et al., 2014)	4,521	17	CC BY 4.0
7	Communities (Redmond, 2009)	1,993	128	CC BY 4.0
8	Communities (unnormalized) (Lahoti et al., 2019)	2,214	147	CC BY 4.0
9	COMPAS (2 years) (Angwin et al., 2016)	6,172	53	?
10	COMPAS (2 years, violent) (Angwin et al., 2016)	4,743	54	?
11	COMPAS (Angwin et al., 2016)	11,757	47	?
12	CreditCard (Yeh, 2009)	30,000	25	CC BY 4.0
13	Drug (Fehrman et al., 2015)	1,885	32	CC BY 4.0
14	Dutch (Le Quy et al., 2022)	60,420	12	[a]
15	German Credit (Hofmann, 1994)	1,000	21	CC BY 4.0
16	German Credit (numeric) (Hofmann, 1994)	1,000	25	CC BY 4.0
17	South German Credit (Grömping, 2019b)	1,000	21	CC BY 4.0
18	German Credit (onehot) (Hofmann, 1994)	1,000	65	Apache License
19	Heart Disease (Janosi et al., 1988)	303	14	CC BY 4.0
20	HMDA (Consumer Financial Protection Bureau, 2022)	2,000,000	19	?
21	Law School (tensorflow) (Wightman, 1998)	22,407	39	CC BY-SA 4.0
22	Law School (LeQuy) (Wightman, 1998; Le Quy et al., 2022)	18,692	12	CC BY-SA 4.0
23	MEPS (Panel 19, FY2015) (Agency for Healthcare Research and Quality, 2018)	15,830	1,831	[b]
24	MEPS (Panel 20, FY2015) (Agency for Healthcare Research and Quality, 2018)	17,570	1,831	[b]
25	MEPS (Panel 21, FY2016) (Agency for Healthcare Research and Quality, 2018)	15,675	1,941	[b]

	Name and Citation	Rows	Columns	License
26	Nursery (Rajkovic, 1989)	12,960	9	CC BY 4.0
27	ricci (Miao, 2010)	118	5	?
28	Stop, Question and Frisk Data (New York City Police Department, 2012)	8,947	83	[c]
29	Chicago Strategic Subject List (Chicago Data Portal, 2020)	398,684	48	NA
30	Student (Cortez and Silva, 2008)	395	33	CC BY 4.0
31	Student (Language) (Cortez and Silva, 2008)	649	33	CC BY 4.0
32	generate_synthetic_data (Zafar et al., 2017)	2,000	4	GPL-3.0
33	Lipton synthetic hiring dataset (Lipton et al., 2018)	2,000	4	CC 0
34	synth (Donini et al., 2018)	6,400	4	?
35	Folktables ACSIncome (Ding et al., 2021)	1,664,500	11	CC 0
36	Folktables ACSPublicCoverage (Ding et al., 2021)	1,138,289	20	CC 0
37	Folktables ACSMobility (Ding et al., 2021)	620,937	22	CC 0
38	Folktables ACSEmployment (Ding et al., 2021)	3,236,107	17	CC 0
39	Folktables ACSTravelTime (Ding et al., 2021)	1,466,648	17	CC 0
40	Folktables ACSIncome (small) (Ding et al., 2021)	245,673	11	CC 0
41	Folktables ACSPublicCoverage (small) (Ding et al., 2021)	174,178	20	CC 0
42	Folktables ACSMobility (small) (Ding et al., 2021)	98,081	22	CC 0
43	Folktables ACSEmployment (small) (Ding et al., 2021)	478,236	17	CC 0
44	Folktables ACSTravelTime (small) (Ding et al., 2021)	216,385	17	CC 0

[a] Copyright 2001, Centraal Bureau voor de Statistiek (CBS) (Statistics Netherlands) and Minnesota Population Center.

[b] See `https://meps.ahrq.gov/data_stats/data_use.jsp`.

[c] "All rights reserved", see `https://www.nyc.gov/home/terms-of-use.page`.

Appendix B. Supplementary Figures

This section contains supplementary figures that complement the primary results and provide further context for the analyses discussed in the main manuscript.

Table A2: Countries represented in fair ML data. Each count represents a dataset that includes data from the specified country. There is one dataset representing data from across the world and one representing data from Hungary, Switzerland and the United States.

Country	Count	Percentage (%)
United States	28	59.57
Portugal	6	12.77
Germany	4	8.51
N/A	3	6.38
Hungary, Switzerland & United States	1	2.13
Netherlands	1	2.13
Slovenia	1	2.13
Taiwan	1	2.13
Turkey	1	2.13
World	1	2.13

Table A3: Quantitative comparison of datasets available in different fairness libraries. *FairGround allows for the input of any custom fairness methods by users.

		Number of		Meta-	Collec-
Library	**Main Focus**	**Datasets**	**Methods**	**Features**	**tions**
ABCFair	methods, metrics	7 (5)	10	✗	✗
Aequitas Flow	methods, metrics, guides	11 (11)	10	✗	✗
AIF360	methods, metrics	8 (8)	15	✗	✗
Fairlearn	methods, metrics, guides	6 (4)	6	✗	✗
FairGround (ours)	data	44	7*	✓	✓

Table A4: Comparison of dataset collections in FairGround and other work, showing whether a debiasing method is ever the best performing method for any of the datasets for Equalized Odds Difference (left) and Demographic Parity Difference (right). For outside collections the closest matching scenarios within FairGround are selected.

	FairGround							
	All	**Open (all)**	**Open (lg.)**	**Open (sm.)**	**ABC Fair[a]**	**AIF 360[b]**	**Fried-ler[c]**	**Typ. 3[d]**
DisparateImpactRemover (pre)	✓/ ✗	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✓	✓/ ✓	✗/ ✗
LFR (pre)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✓/ ✓	✓/ ✓	✓/ ✓
GridSearchReduction (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
AdversarialDebiasing (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
MetaFairClassifier (in)	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓
GerryFairClassifier (in)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✗/ ✗
CalibratedEqOdds (post)	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✓/ ✓	✗/ ✗	✗/ ✗	✗/ ✗
No. of Datasets	44	32	16	5	5	7	5	3

[a] Five out of seven datasets in ABCFair (Defrance et al., 2024) are used.
[b] Seven out of eight datasets in AIF360 (Bellamy et al., 2018) are used, the skipped dataset is available in FairGround, but used as a regression dataset in AIF360.
[c] Friedler et al. (2019)
[d] "Typical 3" refers to Adult, Compas and German Credit, the three most commonly used datasets in fairness research (Fabris et al., 2022).

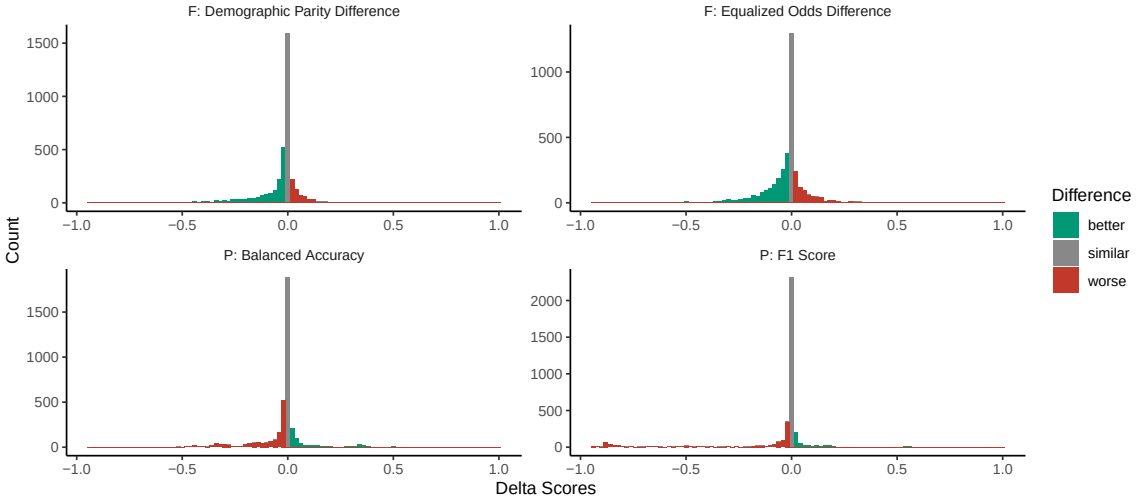

Figure B1: Delta scores across all four metrics are highly variable. Distribution of delta values for metrics of performance and fairness across different processing algorithms. Color-coding indicates whether the change is sizable (above an absolute threshold of 0.01) and corresponds to better (green) or worse (red) scores. For algorithmic fairness metrics lower scores are more desirable, whereas for metrics of performance higher scores are more desirable.

Table A5: Scenarios in the *De-Correlated Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$. The larger collection encompasses the smaller one. Scenarios are listed based on insertion order.

	C	Dataset	Sens. Attributes	Domain
1	k	folktables_acspubliccoverage	RAC1P	economics
2	k	heart_disease	sex	cardiology
3	k	hmda	applicant_sex_name; applicant_race_name_1	finance
4	k	stop_question_and_frisk_data	SUSPECT_SEX; SUSPECT_RACE_DESCRIPTION; SUSPECT_REPORTED_AGE	law
5	k	folktables_acsemployment_small	RAC1P	economics
6	τ	folktables_acstraveltime	RAC1P	economics
7	τ	compas	sex; age	law
8	τ	folktables_acsincome_small	RAC1P	economics
9	τ	compas_2_years	age	law
10	τ	communities_unnormalized	pct12-21	law
11	τ	arrhythmia	sex	cardiology
12	τ	folktables_acspubliccoverage_small	RAC1P	economics
13	τ	compas_2_years_violent	age	law
14	τ	south_german_credit	age; foreign_worker	finance
15	τ	dutch	age	demography
16	τ	folktables_acsmobility_small	RAC1P	economics
17	τ	law_school_tensorflow	gender	education
18	τ	german_credit_onehot	<= 25 years	finance
19	τ	communities	racePctAsian	law
20	τ	nursery	finance	education
21	τ	german_credit_numeric	age	finance
22	τ	chicago_strategic_subject_list	RACE CODE CD	law

Table A6: Scenarios in the *Permissively Licensed Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$; an empty value corresponds to the full collection. The larger collections encompass the smaller ones. Scenarios are ordered based on when they were added to the collection.

	C	Dataset	Sens. Attributes	license
1	k	folktables_acspubliccoverage	RAC1P	CC 0
2	k	heart_disease	sex	CC BY 4.0
3	k	communities_unnormalized	pct12-21	CC BY 4.0
4	k	lipton_synthetic_hiring_dataset	sex	CC 0
5	k	bank	age; marital	CC BY 4.0
6	τ	german_credit_onehot	$>$ 25 years	Apache License
7	τ	folktables_acsincome	RAC1P	CC 0
8	τ	south_german_credit	age	CC BY 4.0
9	τ	folktables_acsemployment_small	RAC1P	CC 0
10	τ	german_credit_numeric	age	CC BY 4.0
11	τ	student	sex; age	CC BY 4.0
12	τ	folktables_acstraveltime_small	RAC1P	CC 0
13	τ	folktables_acspubliccoverage_small	RAC1P	CC 0
14	τ	communities	agePct16t24	CC BY 4.0
15	τ	folktables_acsmobility	RAC1P	CC 0
16	τ	law_school_tensorflow	gender	CC BY-SA 4.0
17		arrhythmia	sex	CC BY 4.0
18		adult	race	CC BY 4.0
19		nursery	finance; parents	CC BY 4.0
20		folktables_acsincome_small	RAC1P	CC 0
21		creditcard	SEX	CC BY 4.0
22		folktables_acsmobility_small	RAC1P	CC 0
23		student_language	age	CC BY 4.0
24		drug	ethnicity	CC BY 4.0
25		law_school_lequy	racetxt; male	CC BY-SA 4.0
26		folktables_acstraveltime	RAC1P	CC 0
27		bank_additional_full	age; marital	CC BY 4.0
28		german_credit	foreign_worker	CC BY 4.0
29		generate_synthetic_data	s1	GPL-3.0
30		bank_additional	age	CC BY 4.0
31		folktables_acsemployment	RAC1P	CC 0
32		bank_full	age	CC BY 4.0

Table A7: Scenarios in the *Geographically Diverse Datasets* collection. Column C denotes collection membership: k corresponds to the small collection with a cutoff value of $k = 5$; τ corresponds to the bigger collection with a cutoff value of $\tau = 0$; an empty value corresponds to the full collection. The larger collections encompass the smaller ones. Scenarios are ordered based on when they were added to the collection.

	C	Dataset	Sens. Attributes	country
1	k	folktables_acspubliccoverage	RAC1P	USA
2	k	heart_disease	sex	HUN;CHE;USA
3	k	dutch	age; citizenship	NLD
4	k	creditcard	SEX	TWN
5	k	german_credit_onehot	$>$ 25 years	DEU
6	τ	student	sex	PRT
7		arrhythmia	sex	TUR
8		nursery	finance; parents	SVN
9		synth	sensible_feature	NA
10		drug	ethnicity	WORLD

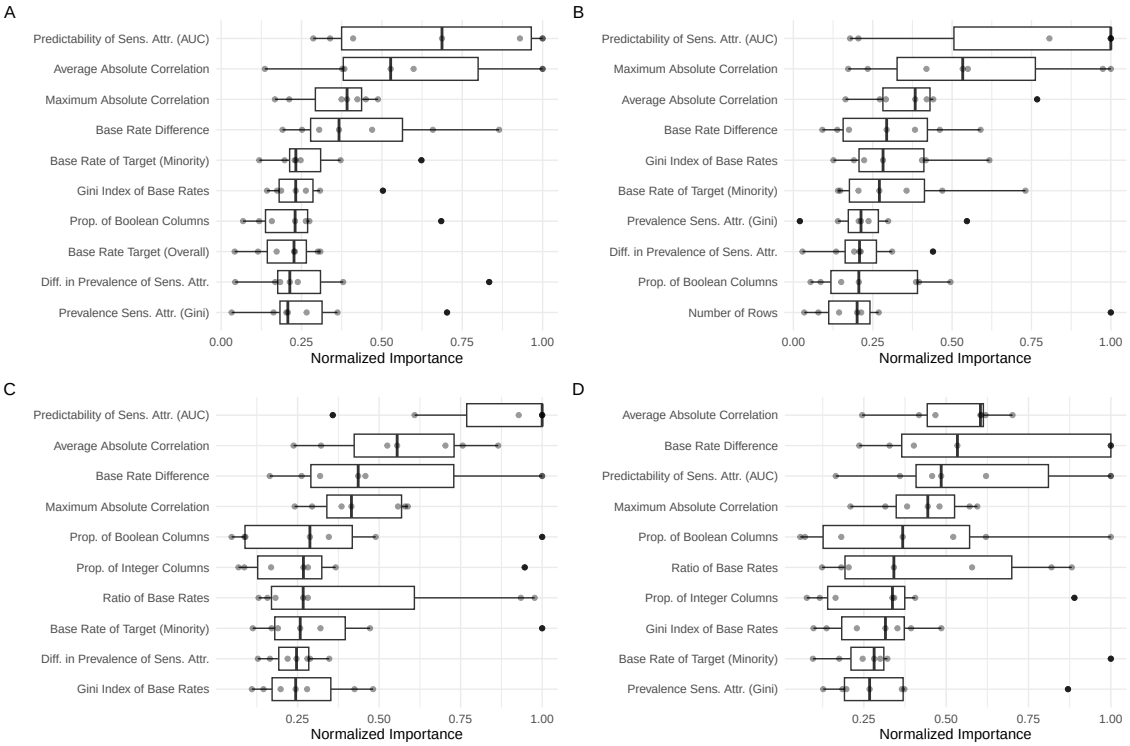

Figure B2: Normalized feature importance of the 10 most important computed metadata features to predict the difference in *Balanced Accuracy* (A), *F1 Score* (B), *Equalized Odds Difference* (C) and *Demographic Parity Difference* (D) across different processing methods.

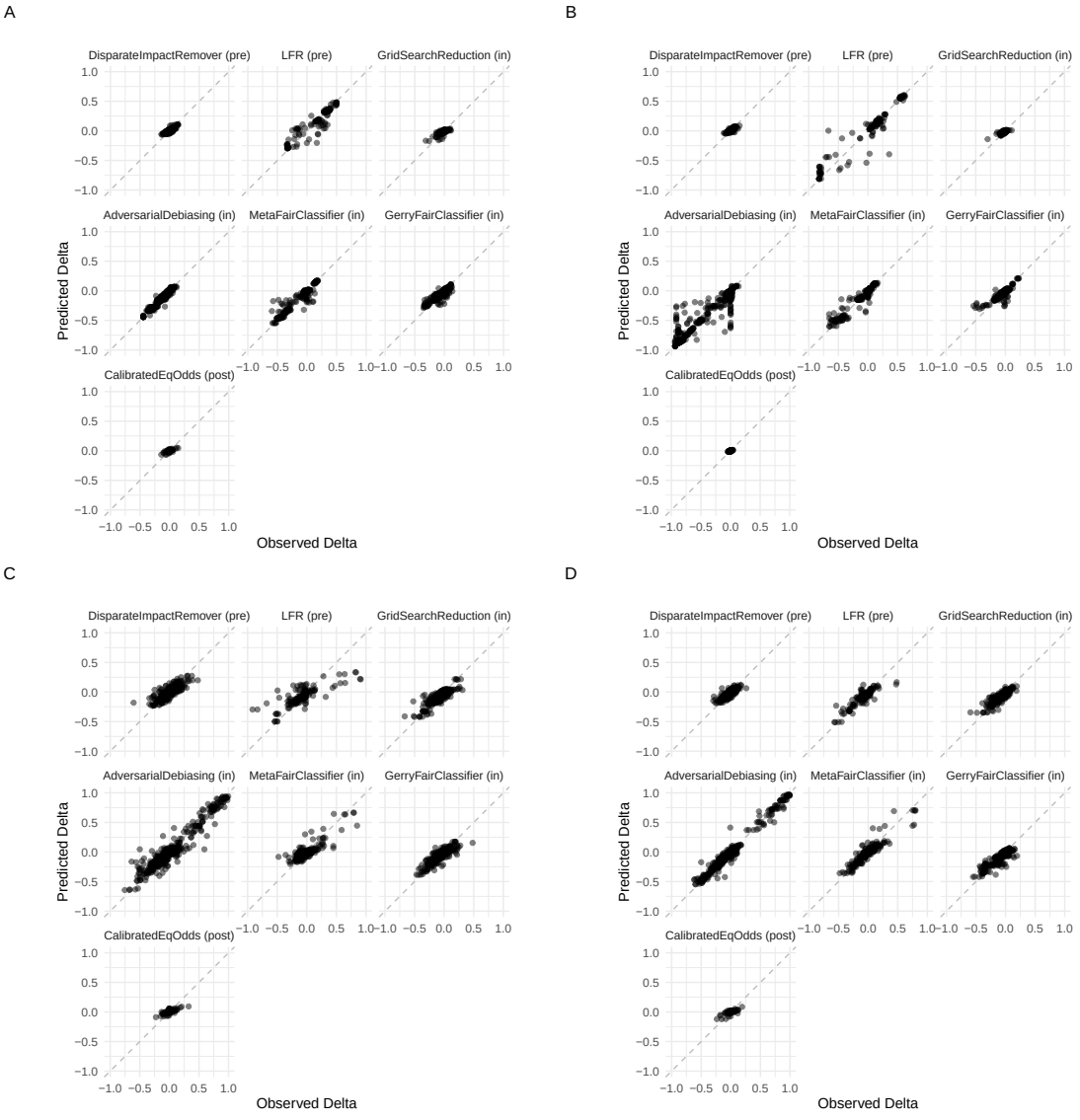

Figure B3: Comparison between observed and model-predicted values for *Balanced Accuracy* (A), *F1 Score* (B), *Equalized Odds Difference* (C) and *Demographic Parity Difference* (D) across different processing methods.

Appendix C. Technical Appendix

C.1 Metrics

$$\begin{aligned} \text{Precision} &= \Pr(y = 1 | \hat{y} = 1) \\ \text{Recall} &= \Pr(\hat{y} = 1 | y = 1) \\ \text{Specificity} &= \Pr(\hat{y} = 0 | y = 0) \end{aligned}$$

We use Balanced Accuracy (bAcc; Eq. 1) and F1 Score (Eq. 2) as measures of performance. The two performance metrics are defined as follows:

$$\text{bACC} = \frac{\text{Specificity} + \text{Recall}}{2} \tag{1}$$

$$\text{F1 Score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} \tag{2}$$

We use Equalized Odds Difference (EOD; Eq. 3) and Demographic Parity Difference (DPD; Eq. 4) as measures of algorithmic fairness. The two fairness metrics are defined as follows:

$$\text{EOD} = \max_g \Pr(\hat{y} = 1 | y = 1, S = g) - \min_g \Pr(\hat{y} = 1 | y = 1, S = g) \tag{3}$$

$$\text{DPD} = \max_g \Pr(\hat{y} = 1 | S = g) - \min_g \Pr(\hat{y} = 1 | S = g) \tag{4}$$

When comparing different fairness aware methods, we use delta scores ($\Delta_{a,b}$) for their comparison. These scores are computed for each performance and fairness metric and are defined as follows:

$$\Delta_{a,b} = score_{a,b} - score_{a,\text{baseline}} \tag{5}$$

C.2 Selection Algorithm

Given the corpus of datasets and their associated scenarios $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$, where each dataset D_i consists of a set of scenarios $D_i = \{s_{i1}, s_{i2}, \ldots\}$, the goal is to construct a collection of scenarios $\mathcal{C}$ such that the pairwise spearman correlations of delta scores (Eq. 5), across *Balanced Accuracy* (Eq. 1), *F1 Score* (Eq. 2), *Equalized Odds Difference* (Eq. 3) and *Demographic Parity Difference* (Eq. 4) between members of $\mathcal{C}$ are as low as possible across different families of fair ML algorithms (*Learning Fair Representations* (Zemel et al., 2013), *Disparate Impact Remover* (Feldman et al., 2015), *Adversarial Debiasing* (Zhang et al., 2018), *Meta-Algorithm* (Celis et al., 2019), *Rich Subgroup Fairness / GerryFair* (Kearns et al., 2018), *Grid Search Reduction* (Agarwal et al., 2018), *Group-Specific Thresholds* (Hardt et al., 2016)). To control the number of scenarios in $\mathcal{C}$, we use either a fixed number k or a correlation threshold τ. While this work uses only one of these constraints at a time, they can be combined if desired. The algorithm proceeds as follows:

1. Let r_{ab} denote the *Spearman rank correlation* between scenarios s_a and s_b, where $s_a, s_b \in \bigcup_{i=1}^{N} D_i$.

2. For each scenario s_a, compute the average Spearman correlation to all other scenarios:

$$\bar{r}_a = \frac{1}{M-1} \sum_{b \neq a} r_{ab}$$

   where M is the total number of scenarios in the corpus. Select the scenario s_m with the lowest average correlation:

$$m = \arg\min_a \bar{r}_a$$

   Initialize the selected set $\mathcal{C} = \{s_m\}$, and the remaining pool $\mathcal{R} = \left(\bigcup_{i=1}^{N} D_i\right) \setminus D_{i(m)}$, where $D_{i(m)}$ is the dataset containing scenario s_m.

3. Repeat the following until $|\mathcal{C}| = k$ or no candidate in $\mathcal{R}$ has an average Spearman correlation strictly less than τ with all members of $\mathcal{C}$:

   (a) For each scenario $s_j \in \mathcal{R}$, compute the average correlation with the current set $\mathcal{C}$:

$$\bar{r}_{j\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{s_i \in \mathcal{C}} r_{ij}$$

   (b) Identify the scenario s_{j^*} with the lowest such average:

$$j^* = \arg\min_{j \in \mathcal{R}} \bar{r}_{j\mathcal{C}}$$

   (c) If $\bar{r}_{j^*\mathcal{C}} < \tau$, add s_{j^*} to $\mathcal{C}$, and remove all scenarios in the same dataset $D_{i(j^*)}$ from $\mathcal{R}$.

4. The algorithm terminates when $|\mathcal{C}| = k$ or no remaining scenario has an average Spearman correlation below τ with the current set $\mathcal{C}$. The resulting subset $\mathcal{C}$ is returned as the final collection of minimally correlated scenarios.

C.3 Example Code using the Package

The following subsection contains exemplary code illustrating the usage of the Python package. We recommend readers to review the online package documentation at `https://brave-ocean-078c2100f.6.azurestaticapps.net/` for a more in-depth description of the package's functions.

C.3.1 Using a Dataset

```
from fairml_datasets import Dataset

# Get the dataset
dataset = Dataset.from_id("folktables_acsemployment")
```

```
# Load as pandas DataFrame
df = dataset.load()  # or df = dataset.to_pandas()
print(f"Dataset shape: {df.shape}")

# Get the target column
target_column = dataset.get_target_column()
print(f"Target column: {target_column}")

# Get sensitive attributes (before transformation)
sensitive_columns_org = dataset.sensitive_columns

# Transform to e.g. impute missing data
df_transformed, transformation_info = dataset.transform(df)
# Sensitive columns may change due to transformation
sensitive_columns = transformation_info.sensitive_columns

# Split into train and test sets
train_df, test_df = dataset.train_test_split(df, test_size=0.3)

# Run analyses on the data
```

C.3.2 Using a Collection of Datasets / Scenarios

```
from fairml_datasets.collections import DeCorrelatedSmall

collection = DeCorrelatedSmall()

# The collection consists of scenarios
for scenario in collection:
    # Each scenario behaves just like a dataset

    # Load as pandas DataFrame
    df = scenario.load()  # or df = scenario.to_pandas()
    print(f"Dataset shape: {df.shape}")

    # Get the target column
    target_column = scenario.get_target_column()
    print(f"Target column: {target_column}")

    # Get sensitive attributes (before transformation)
    sensitive_columns_org = scenario.sensitive_columns

    # Transform to  e.g. impute missing data
```

```
df_transformed, transformation_info = scenario.transform(df)
# Sensitive columns may change due to transformation
sensitive_columns = transformation_info['sensitive_columns']

# Split into train and test sets
train_df, test_df = scenario.train_test_split(df, test_size=0.3)

# Run analyses on the data
```

C.4 Annotation Procedure

We started the annotation process by collecting all tabular datasets used for fair classification tasks in a large survey of fair ML datasets (Fabris et al., 2022). This provided a list of $n = 37$ unique datasets. Additionally, we added the folktables (Ding et al., 2021) collection of datasets, due to its recent popularity and as the datasets specifically try to address issues in the most popular dataset in the survey: *Adult* (Kohavi, 1996).

For each dataset, we annotated the information required to practically use the dataset in a fair classification task, as well as key qualitative and quantitative data regarding the information represented in each dataset. During this process, a critical issue quickly became apparent: While datasets are commonly referenced by name as if they were uniquely identified, this is often not the case in practice. A striking example is the widely used Bank dataset, one of the most frequently cited datasets in Fair ML (Fabris et al., 2022). Although typically referred to as Bank or Bank Marketing, the primary source[1] actually comprises four distinct datasets, each differing in their respective number of instances and attributes. Recognizing this ambiguity, we adapted our annotation methodology to explicitly capture dataset variants, significantly increasing the number of distinct datasets in the corpus. In our framework, we treat these variants as separate datasets while preserving their connection to maintain clarity and traceability.

When collecting the information required to download and load datasets, we were forced to exclude $n = 11$ datasets due to data not being publicly available or with restricted access. We excluded a further $n = 18$ datasets, if there were issues with recreating how a dataset was generated or the dataset's usage did not fit into schema of a "classic" fairML classification task including features, a target column and sensitive attribute(s). A detailed breakdown of excluded datasets and the reasons for their exclusion is available in Section C.7.

After exclusion of non-eligible datasets and inclusion of different variants, we arrive at a list of $N = 44$ datasets.

Datasets were annotated by two of the authors with help from research assistants. A random subset of annotations was reviewed by a third author.

C.5 Annotated Columns

The following section provides descriptions of columns which were manually annotated for each dataset in the corpus.

1. https://archive.ics.uci.edu/dataset/222/bank+marketing

new_dataset_id A unique identifier for each dataset. Usually derived from the dataset name.

dataset_name An official, common, or known name of the dataset that is unique across datasets.

base_dataset_name In case there are different variants of the same dataset, this field holds a common name to group all these variants together.

description_public This is a free-text field reporting (1) the aim/purpose of a data artifact (i.e., why it was developed/collected), as stated by curators or inferred from context; (2) a high-level description of the available features; (3) the labeling procedure for annotated attributes, with special attention to sensitive ones, if any; (4) the envisioned ML task, if any.

notes_public Any notes or comments regarding this dataset / task combination.

dataset_aliases Any names that this dataset is called by. While 'dataset_name' only contains the single most common name, this field holds possible aliases used to reference this dataset.

affiliation Affiliation of the creators of the dataset. Based on reports, articles, or official web pages presenting the dataset.

domain_class_main The main field where the data is used (e.g., computer vision for ImageNet) or the field studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert).

domain_class_multi The primary fields where the data is used (e.g., computer vision for ImageNet) or the fields studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert). Multiple domains are possible in this feature.

domain_freetext Fine-grained domain of the prediction task. Summarized with 1 - 2 words.

sample_size Dataset cardinality. Rough estimate of the size of the dataset.

year_last_updated The last known update to the dataset. For resources whose collection and curation are ongoing (e.g., HMDA), we write “present”.

years_data The timespan covered in the data. This refers to the "social realities" captured in the data i.e., data from which year(s) is present in the data.

citation The main / official source to cite this dataset in BibTeX format. For synthetic datasets, this refers to the original paper where the dataset was first introduced.

main_url The main landing page or website related to the dataset. This is a website with information on the dataset and not the dataset itself, which is referenced via 'download_url'.

related_urls List of related links and resources to the dataset.

license Under which license is the dataset made available? A "?" indicates that no license was found.

continent Continent(s) where the dataset is sourced. In two-letter format. If "n/a", this concept is not applicable for a dataset (e.g., a synthetic one).

country Countrie(s) where the dataset is sourced. In ISO3 format. If "n/a", this concept is not applicable for a dataset (e.g., a synthetic one).

dataset_variant_id This ID is used to identify different datasets belonging to the same original dataset e.g., COMPAS has 3 unique smaller datasets belonging to this one bigger one. In cases like this, each smaller dataset gets its own dataset_variant_id.

dataset_variant_description Description outlining how this "sub-dataset" is different from the others. Only filled out if there are multiple "dataset_variant_ids".

is_accessible Is the dataset publicly accessible? "Manual download" indicates that an automated download is not possible.

download_url URL to the dataset file itself, if it is publicly accessible.

custom_download Are there some extra steps needed to download the dataset itself, e.g., unpacking a ZIP archive?

filename_raw Filename of the dataset for downloading it or finding it in a ZIP archive.

format Format of the dataset. Corresponds to the format the data is in, not the extension of the dataset e.g., CSV for comma-separated-values, TSV for tab-separated-values, FIXED-WIDTH for fixed-width formats etc.

colnames Column names to use if the dataset file does not include them.

processing Does the dataset need some special pre-processing to be in the correct format?

sensitive_attributes Sensitive attributes that are *available* in the dataset. Supports multiple entries, separated with a semicolon and a space: '; '.

typical_col_sensitive All columns containing available sensitive attributes and the information they contain in a categorical fashion. Covering the attributes listed in 'sensitive_attributes'. Formatted as a JSON dictionary.

typical_col_features All columns typically used as features / predictors. Either a list of column names indicating a positive selection or a list of column names prefixed with a - indicating a negative selection i.e. all columns except the listed ones. A - indicates using all available columns (except the target).

typical_col_target Column(s) which are being predicted. If more than one, separated by semicolons.

target_lvl_good Which value of the target variable is considered desirable? Desirable here means good for any person impacted by a system built using this data.

target_lvl_bad Which value of the target variable is considered undesirable? Undesirable here means bad for any person impacted by a system built using this data.

dataset_size Whether a dataset is exceptionally large.

C.6 Computed Metadata

The following section provides descriptions of the computed metadata features which are implemented in the Python package and computed for each of the datasets in the corpus. The technical implementation can be reviewed in the publicly available source code of the package.

Size As Ding et al. (2021) note, increasing dataset size does not necessarily reduce observed disparities due to persistent structural inequalities. We try to cover a broad range of dataset sizes in our corpus and compute dataset sizes by rows (samples) and columns (attributes) of both prepared (`meta_pretrans_n_rows, meta_pretrans_n_cols`) and transformed datasets (`meta_n_rows, meta_n_cols`).

Missing values To address potential bias from missing data (e.g. see Pessach and Shmueli, 2022; Wang and Singh, 2021; Martínez-Plumed et al., 2019), we

calculate the fraction of missing data per dataset. Metadata was computed prior to processing to assess the proportion of rows (`meta_pretrans_prop_NA_rows`), columns (`meta_pretrans_prop_NA_cols`) and cells (`meta_pretrans_prop_NA_cells`) that contain missing values. We further calculate missingness within each group of the protected attribute (only when binarizing; `meta_prop_NA_sens_minority, meta_prop_NA_sens_majority`).

Attribute types We calculate the proportions of different numeric (`meta_prop_cols_float, meta_prop_cols_int`) and logical (`meta_prop_cols_bool`) data types in the data to assess their potential influence.

Sensitive AUC Non-sensitive attributes can act as proxies for sensitive ones (e.g. see Pessach and Shmueli, 2022; Mehrabi et al., 2021; Fawkes et al., 2024). Identifying and addressing such proxies can help mitigate unfairness (Pessach and Shmueli, 2022; Matloff and Zhang, 2022). To assess this, we define *Sensitive AUC* as the ROC-AUC of a random forest model (Ho, 1995) trained to predict the sensitive attribute using only non-sensitive features (`meta_sens_predictability_roc_auc`). A higher Sensitive AUC suggests that non-sensitive attributes may encode sensitive information.

Bivariate correlations Serving as an additional indicator of potential proxy variables, we computed the correlation between each non-sensitive feature and the sensitive attribute, using the average and maximum correlation values (`meta_average_absolute_correlation, meta_maximum_absolute_correlation`).

Number of protected groups Some fairness methods require binary representations of protected attributes, leading to the binarization of categorical or numerical sensitive attributes during preprocessing. Documenting the original number of protected groups before processing (`meta_pretrans_unique_group_counts_pre_agg`) helps track this process and may provide insight into how such simplifications affect the performance and suitability of fairness methods.

Prevalence We computed the proportions of minority and majority groups within the dataset (only when binarizing; `meta_prev_sens_minority, meta_prev_sens_majority`), along with the absolute difference between them (`meta_prev_sens_difference`) and the imbalance ratio (`meta_prev_sens_ratio`). A smaller absolute difference and an imbalance ratio closer to 1 indicate a more balanced distribution of the sensitive attribute.

Base Rate Similar to prevalence, we computed the probability of the favorable outcome overall (`meta_base_rate_target`) and for each group (only when binarizing; `meta_base_rate_target_sens_minority`, `meta_base_rate_target_sens_majority`) along with the absolute difference (`meta_base_rate_difference`) and ratio (`meta_base_rate_ratio`) between them.

Gini-Simpson Index The Gini-Simpson Index measures the probability that two randomly selected individuals belong to different groups. Similar indices have been previously used by Mecati et al. (2023) and Vetrò et al. (2021) to assess balance and detect potential unfairness in datasets. We compute the Gini-Simpson Index for both group prevalence and base rates

$$GS = 1 - \sum_i p_i^2,$$

where p_i is the proportion of instances in group $i \in \{1, 2\}$ (protected or non-protected). For prevalence, this is is the proportion of individuals per group relative to the entire dataset (`meta_prev_sens_gini`). For base rates, p_i denotes the proportion of favorable outcomes within each group (`meta_base_rate_sens_gini`).

C.7 Excluded Datasets

This subsection contains explanations for additional datasets that were excluded from the corpus. The annotation procedure is described in detail in Section C.4.

2016 Presidential Elections (2 datasets) This dataset from the FiveThirtyEight 2016 Election Forecast was developed with the goal of providing an aggregated estimate of the probability that Trump/Clinton wins the 2016 election based on multiple polls, weighting each input according to sample size, recency, and historical accuracy of the polling organization. For each poll, the dataset provides the period of data collection, its sample size, the pollster conducting it, their rating, and a url linking to the source data. The dataset does not contain any sensitive attributes and was therefore excluded. One annotated but excluded dataset came from ABC News, and another, potentially deviating, from (Sabato and Yom-Tov, 2020).

Cancer Cases and Deaths (3 datasets) The main dataset reports state-level cancer prevalence for 18 cancer types, based on data from the CDC's NPCR and the NCI's SEER program. Mortality data come from the CDC's National Vital Statistics System. As it contains only aggregated data on state-level, it was excluded from our analysis. Two additional datasets provided the source data on new cases and deaths. As neither was used in isolation in our annotations, both were excluded with the main dataset.

Clinical Annotations / Warfarin Dosage / PharmaGKB (4 datasets) The data, collected by the International Warfarin Pharmacogenetics Consortium and co-curated by PharmGKB, was used to study algorithmic estimation of optimal warfarin dosage. The original data includes thousands of patient demographics, comorbidities, medications, genetics, and effective warfarin doses. However, the available datasets do not contain demographic details and only a specialty group column indicates few pediatric cases. Due to the absence of sensitive attributes, these datasets were excluded. The excluded datasets comprised: 1) meta-data for each clinical annotation; 2) genotype/allele-based annotation text with CPIC-assigned function, if available; 3) supporting annotation details (variant, guideline, label); and 4) clinical annotation history with creation and update dates.

COMPAS (4 datasets) We retain the original COMPAS data published by ProPublica Angwin et al. (2016). Specific versions of the COMPAS dataset were excluded, including an unofficial version published on Kaggle, used in one reviewed study (Jabbari et al., 2020), and two others, each appearing in a single paper (Wang et al., 2019; Mandal et al., 2020), due to a lack of clarity in the differences and processing from the original ProPublica release. The COMPAS repository[2] also includes a file with "raw" scores, named `compas-scores-raw.csv`, which we decided not to include, as it is not further utilized in the analysis.

FICO Credit Score, Credit Score Performance (2 datasets) The dataset originates from a 2007 Federal Reserve report to the US Congress on credit scoring and its

2. `https://github.com/propublica/compas-analysis`

effects on the availability and affordability of credit. The collection, creation, processing, and aggregation was carried out by the working group and is based on a sample of 301,536 TransUnion TransRisk scores from 2003. The dataset contains only aggregated statistics per FICO score and race/ethnicity group and was therefore excluded. A second version with unclear differences was also excluded.

Fifa 20 Complete Player This dataset was scraped by Stefano Leone and shared on Kaggle. It contains player data from FIFA Career Mode (FIFA 15-20). We excluded this dataset, because relevant sensitive attributes and target variables were unclear. A paper by Awasthi et al. (2021) created a sensitive attribute by predicting nationality from player names using LSTM, an approach that could introduce unnecessary uncertainty and therefore may have reduced comparability.

Pima Diabetes This dataset was derived from a medical study of Native Americans from the Gila River Community, often called Pima. Conducted by the National Institute of Diabetes and Digestive and Kidney Diseases since the 1960s, the study found a large prevalence of *diabetes mellitus* in this population. The dataset includes a subset of the original study, focusing on women of age 21 or older. It reports diabetes test results and eight key risk factors, such as number of pregnancies, skin thickness, and BMI. Relevant sensitive attributes were not clear based on the papers we reviewed, so we decided to exclude the dataset.

US Census (1990) This dataset is derived from the 1990 US Census. In the reviewed literature, the classification task was often unclear or unsuitable for our analysis goals (e.g., Sabato and Yom-Tov, 2020). Another meta-analysis referenced 25 selected numeric attributes without specifying them.

C.8 Computational Infrastructure

Experiments were run on a shared Linux compute cluster with partitions and compute infrastructure chosen based on availability of resources. Experiments were run as four consecutive jobs, the first running experiments at high concurrency and the later re-running errored out experiments at lower concurrency.

The first job was run on a node with access to 76 CPU cores and 512 GB of memory over a duration of 11 hours. Later jobs were run on a node with 96 CPUs and 1 TB of memory, using 5-fold parallelism and a maximum execution time of 2.5 hours for the second and third run and 5 hours for the last run. Experiments were conducted using only CPU compute.

C.9 Software

Simulation experiments were conducted using Python (Python Core Team, 2019) version 3.10 and the Python package `multiversum` (Simson, 2024) version 0.7.0. We used the implementations of fairness-aware processing methods from the package `AIF360` (Bellamy et al., 2018) and used `scikit-learn` (Pedregosa et al., 2011) to fit logistic regressions. Data were processed using the newly developed `fairml_datasets` package, utilizing `pandas` (McKinney, 2010), `fastparquet` (Durant) and `scikit-learn`. Multiple other packages were utilized as (peer) dependencies of the named packages. We use `uv` (Marsh, 2024) for virtual environment management.

Results from the experiments were analysed using R version 4.4.1 (R Core Team, 2024) with packages from the `tidyverse` (Wickham et al., 2019), `patchwork` (Pedersen, 2024) and `tidymodels` (Kuhn and Wickham, 2020). Color schemes are used from the R packages `awtools` (Wehrwein, 2025) and `wesanderson` (Ram and Wickham, 2023). We use `renv` for virtual environment management.

Lockfiles for both Python and R packages are provided with the codebase.

Experiments were executed using a docker container converted to the `enroot` format[3].

3. https://github.com/NVIDIA/enroot