Is Underrepresentation Overemphasized? A Study of Data Bias and Algorithmic Fairness

Marina Ceccon[1,*], Giandomenico Cornacchia[2], Davide Dalle Pezze[1], Alessandro Fabris[3,4,*] and Gian Antonio Susto[1]

[1]*University of Padova, Italy*
[2]*IBM Research, Dublin, Ireland*
[3]*University of Trieste, Italy*
[4]*Max Planck Institute for Security and Privacy, Bochum, Germany*

Abstract

Undesirable biases embedded in data are major contributors to algorithmic discrimination. Their significance is well acknowledged in both algorithmic fairness research and in emerging legal and regulatory frameworks for anti-discrimination in AI. However, despite this awareness, data biases are still insufficiently explored, limiting progress toward effective computational methods for their identification and mitigation. This article focuses on three prevalent forms of data bias–underrepresentation, proxy variables, and label bias–and their influence on the fairness of models trained on such data. Through extensive experimentation across a variety of datasets, models, and fairness measures, we show that combinations of proxies and label biases are particularly harmful, while underrepresentation in training sets is less impactful than commonly affirmed. We link algorithmic fairness research with anti-discrimination policy through a focus on data. This work is an extended abstract of Ceccon et al. [1].

Keywords

algorithmic fairness, anti-discrimination, bias detection, data bias, AI Act

1. Introduction

Monitoring data bias is a crucial computational approach in combating algorithmic discrimination. Since biases in datasets are core contributors to discriminatory outcomes [2, 3], their management features prominently in recent standards and regulations addressing algorithmic fairness [4, 5, 6].

However, establishing clear criteria for effective bias management raises several key questions that remain largely unexplored in the fairness literature. First, multiple types of data bias contribute to discrimination—but which combinations have the most significant impact? Second, while data biases are often described in qualitative terms, there is a lack of robust quantitative frameworks. Can we reliably and distinctly monitor different forms of bias? Third, although many fairness interventions have been proposed, there is little guidance on how to prioritize them. Ultimately, how can researchers and practitioners approach bias mitigation in a systematic and informed way?

2. Related work

For brevity, we summarize key related work on linking fairness with data properties. Recent research has focused on quantifying data biases and understanding their impact on model fairness. Studies have introduced causal frameworks, examined different bias types, and evaluated fairness metrics under varying data conditions [7, 3]. Notably, Vetrò et al. [2] pioneered predicting discrimination risk from underrepresentation, influencing follow-up work. Building on this, the current study expands the

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
marina.ceccon@phd.unipd.it (M. Ceccon); giandomenico.cornacchia1@ibm.com (G. Cornacchia); davide.dallepezze@unipd.it (D. Dalle Pezze); alessandro.fabris@units.it (A. Fabris); gianantonio.susto@unipd.it (G. A. Susto)
0009-0006-5892-4080 (M. Ceccon); 0000-0001-5448-9970 (G. Cornacchia); 0000-0002-4741-1021 (D. Dalle Pezze); 0000-0001-6108-9940 (A. Fabris); 0000-0001-5739-9639 (G. A. Susto)

scope by analyzing three types of data bias–underrepresentation, label bias, and proxies–and their combined effect on fairness. Unlike earlier work, we evaluate models on unbiased test sets and find that underrepresentation is often overstated, while other biases may pose greater risks to fairness.

3. Data Bias & Injection Protocols

We present three key types of data bias and briefly introduce the bias injection mechanisms we use to study their impact.

3.1. Underrepresentation

Underrepresentation of disadvantaged groups is widely cited in media and fairness research as a major cause of algorithmic discrimination, leading to poor model performance for those groups [8]. This issue is studied as a key predictor of unfair outcomes and is acknowledged in major regulations and standards, which require efforts to document and mitigate it [5, 4].

We vary the underrepresentation of disadvantaged groups in training sets as $u = 1 - r$ where r is the proportion of that group's instances retained. u ranges from 0 (no underrepresentation) to 1 (complete exclusion), allowing systematic exploration of its effects.

3.2. Label bias

Labels are essential in AI as they represent the "ground truth" for model learning, but they often reflect societal biases [9]. Measurement methods can be biased across protected groups, leading to unfair outcomes, for example, crime and medical data may misrepresent or underdiagnose disadvantaged populations [10, 11].

To inject label bias in training sets we flip a proportion f of positive labels ($y = \oplus$) to negative ($y = \ominus$) for instances from the disadvantaged group. The flip factor f ranges from 0 (no label bias) to 1 (all positives flipped).

3.3. Proxies

Proxies are features that reveal or correlate with protected attributes, such as gender or race, even when those attributes are not explicitly included. This undermines the "fairness through unawareness" approach, as protected information often remains encoded in other features [12]. Removing proxy features is a common strategy to improve fairness, and strong proxies are widely recognized as contributors to algorithmic bias [13]. As a result, policymakers increasingly expect practitioners to detect and remove such features from decision-making models.

The strength of proxy features is measured by the proxy factor (sAUC), the AUC of a classifier trained to predict protected attributes from input features. Two methods are used to vary proxy strength. (1) Additive protocol: introduces a new feature correlated with the sensitive attribute, with noise controlling the strength of the correlation (lower noise = stronger proxy). (2) Subtractive protocol: iteratively removes features most similar to the protected attribute (e.g., by correlation), reducing the model's ability to infer sensitive information.

4. Effect of Data Bias

4.1. Setup

The study uses five tabular and two medical imaging datasets commonly used in fairness research, covering key protected attributes like gender, age, ethnicity, and marital status. Tabular datasets include employment, income, recidivism, crime, and credit risk tasks; medical datasets focus on chest X-rays and skin conditions. Models range from traditional machine learning (random forest, SVC, logistic

Table 1
Model fairness remains largely stable despite underrepresentation in the training set. Equal Opportunity (EO) is shown as the underrepresentation factor u increases from 0 (no bias) to 1 (full underrepresentation of the minority group). Results reflect the mean and standard deviation across 10 runs. Asterisks (*) and (**) indicate statistically significant differences from $u = 0$ at the $p = 0.05$ and $p = 0.01$ levels, respectively, based on an unpaired t-test.

Dataset	**sensitive**	**model**	EO $u = 0$ (no bias)	EO $u = 0.2$	EO $u = 0.8$	EO $u = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	0.09 ± 0.02	0.09 ± 0.02	$0.21 \pm 0.07^{**}$
	marital-status	LR	0.36 ± 0.03	0.36 ± 0.03	0.37 ± 0.03	$0.29 \pm 0.04^{**}$
Compas	race	LR	0.17 ± 0.03	0.17 ± 0.03	0.16 ± 0.02	0.16 ± 0.02
Crime	race	LR	0.33 ± 0.11	0.36 ± 0.12	0.30 ± 0.14	0.28 ± 0.13
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04	0.05 ± 0.04
German	age	LR	0.07 ± 0.13	0.06 ± 0.11	0.07 ± 0.09	0.06 ± 0.07
NIH	gender	DenseNet	0.01 ± 0.02	$0.04 \pm 0.01^{**}$	0.03 ± 0.02	$0.06 \pm 0.02^{**}$
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.02	0.11 ± 0.02	0.11 ± 0.02	0.11 ± 0.02

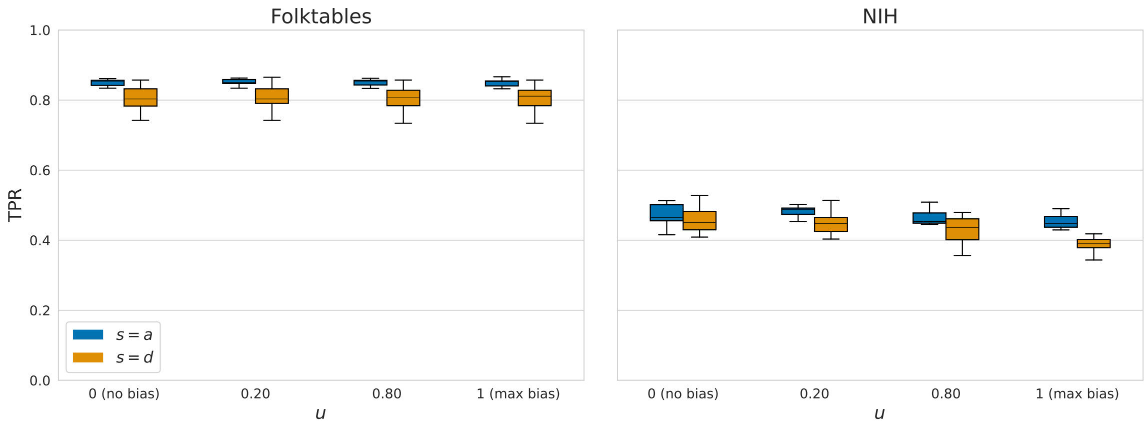

Figure 1: Severe underrepresentation causes only slight changes in True Positive Rates (TPR) across groups. Boxplots show the TPR for the advantaged ($s = a$) and disadvantaged ($s = d$) groups as the underrepresentation factor u varies.

regression) for tabular data to deep learning (DenseNet121, VGG16) for medical imaging. The detailed experimental setup is presented in Ceccon et al. [1].

Performance is measured via balanced accuracy (average of true positive and true negative rates). **Fairness** is evaluated with three metrics. (1) Demographic Parity (DP): difference in positive prediction rates between advantaged and disadvantaged groups, ignoring true labels. (2) Equal Opportunity (EO): difference in true positive rates between groups, emphasizing fairness when ground truth is reliable. (3) Prediction Quality Parity (PQP): difference in balanced accuracy between groups, reflecting overall prediction fairness. **Larger values correspond to more unfavorable outcomes for disadvantaged groups**. Below, we showcase a selection of results that exemplify the overall findings from our experiments.

4.2. Results

Underrepresentation. Underrepresentation of the minority group has minimal impact on algorithmic fairness across datasets, with Equal Opportunity (EO) remaining largely stable as underrepresentation increases (Table 1). Only NIH and Adult (gender) show a statistically significant rise in disparity, and only under extreme underrepresentation ($u = 1$), while some datasets (e.g., Crime, Adult marital status)

Table 2
Label bias has a significant impact on model fairness. Equal Opportunity (EO) as a function of the flip factor f, representing the proportion of positive labels flipped to negative in the disadvantaged group, from $f = 0$ (no bias) to $f = 1$ (maximum bias).

Dataset	sensitive	model	EO $f = 0$ (no bias)	EO $f = 0.2$	EO $f = 0.8$	EO $f = 1$ (max bias)
Adult	gender	LR	0.08 ± 0.02	$0.21 \pm 0.02^{**}$	$0.52 \pm 0.03^{**}$	$0.55 \pm 0.02^{**}$
	marital-status	LR	0.36 ± 0.03	$0.43 \pm 0.04^{**}$	$0.62 \pm 0.02^{**}$	$0.63 \pm 0.02^{**}$
Compas	race	LR	0.17 ± 0.03	0.21 ± 0.05	0.21 ± 0.05	0.15 ± 0.05
Crime	race	LR	0.33 ± 0.11	0.39 ± 0.11	$0.62 \pm 0.17^{**}$	$0.67 \pm 0.15^{**}$
Folktables	race	LR	0.05 ± 0.04	0.05 ± 0.04	0.08 ± 0.04	0.09 ± 0.04
German	age	LR	0.07 ± 0.13	0.10 ± 0.14	0.22 ± 0.16	$0.25 \pm 0.10^{**}$
NIH	gender	DenseNet	0.01 ± 0.02	$0.09 \pm 0.02^{**}$	$0.40 \pm 0.03^{**}$	$0.48 \pm 0.01^{**}$
Fitzpatrick17k	skin type	vgg16	0.09 ± 0.05	$0.16 \pm 0.06^{*}$	$0.21 \pm 0.08^{**}$	$0.28 \pm 0.02^{**}$

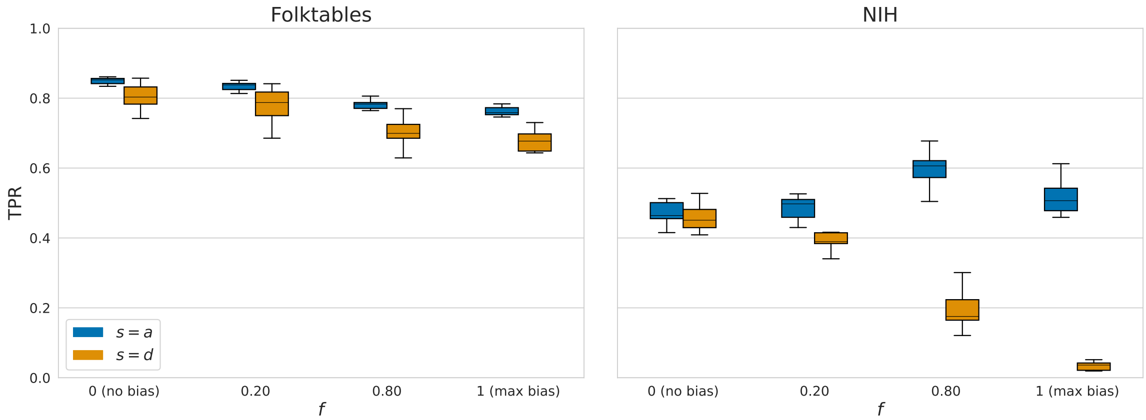

Figure 2: Label bias leads to substantial shifts in group-wise true positive rates (TPR), with stronger impact on the disadvantaged group. Boxplots show the TPR for the advantaged and disadvantaged groups (y-axis) as the proportion of flipped positive labels in the disadvantaged group increases in the training data (x-axis).

even see a slight decrease. This challenges common assumptions about underrepresentation harming fairness. Further analysis reveals that true positive rates (TPR) for both groups remain stable in most cases, with NIH showing a modest decline, especially for the disadvantaged group, due to their larger initial representation (Figure 1).

Label bias. Label bias significantly worsens algorithmic fairness, with Equal Opportunity (EO) increasing across all datasets as label flips in the disadvantaged group grow (Table 2). This effect is much stronger than that of underrepresentation. Datasets like NIH and Adult show large fairness gaps (e.g., EO +0.47 for NIH), while Folktables shows a more modest rise (+0.04). The main driver is the drop in true positive rates (TPR) for the disadvantaged group (Figure 2). In some datasets (e.g., NIH, Adult), the advantaged group's TPR remains stable, widening the gap; in others (e.g., Folktables, Compas), both groups' TPRs decline, resulting in a smaller gap. These differences relate to proxy feature strength, discussed further below.

Proxies. We analyze the interaction between label bias and proxy strength. Figure 3 shows EO for Adult (gender) and Folktables as the flip rate f increases. Each curve corresponds to a different input space, progressively reduced by removing features one at a time (the one most strongly correlated with the sensitive attribute s). In both datasets, the influence of label flips on EO, reflected in the average slope of the curves, diminishes as sAUC decreases. This effect is more pronounced in Adult, which

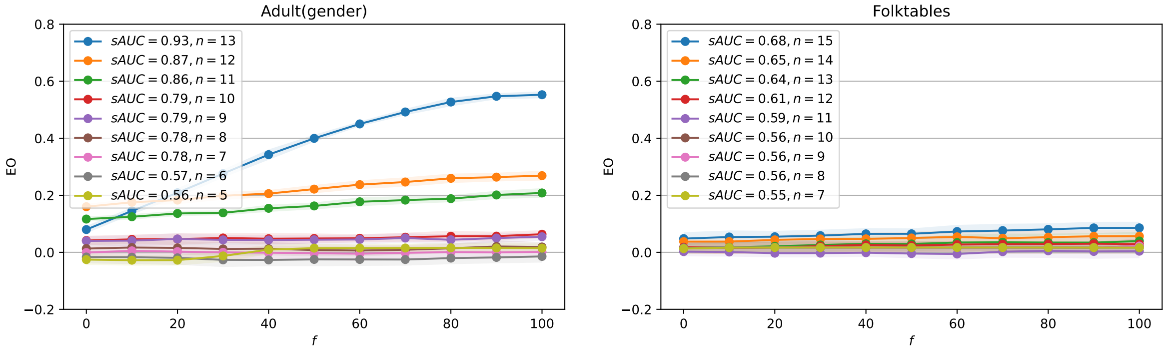

Figure 3: Proxies amplify the discriminatory impact of label bias. Equal Opportunity (EO, y-axis) rises with increasing label bias f (x-axis). The severity of this effect depends on proxy strength: weaker proxies (lower *s*AUC) result in a gentler slope, indicating a reduced influence of label bias on fairness.

begins with stronger proxies (i.e., higher *s*AUC), as indicated in the legend.

5. Discussion

Our findings challenge a widely held belief in algorithmic fairness: that increasing the representation of disadvantaged groups in training data is conducive to fair outcomes. Contrary to this view, our experiments reveal that underrepresentation has a surprisingly limited impact on model fairness across datasets, models, and metrics. While inclusion remains a critical goal for ethical AI development, these results suggest that simply increasing group prevalence in training data, without addressing deeper issues, may not lead to meaningful fairness gains. Instead, we advocate for prioritizing high-quality, well-annotated data from disadvantaged groups for evaluation purposes, enabling reliable fairness assessments beyond training performance.

In contrast, label bias (systematic errors in ground truth labels for disadvantaged groups) emerges as a dominant factor in driving unfairness. Even small amounts of biased labeling can significantly widen performance gaps, especially in True Positive Rates (TPR), and disproportionately harm disadvantaged groups. This effect is further exacerbated when models rely on strong proxy features that correlate with sensitive attributes, enabling discrimination through indirect means. We find that stronger proxies intensify the impact of label bias, while feature reduction can help mitigate this risk.

6. Conclusions & Future Work

Data biases are fundamental drivers of algorithmic discrimination. Although this is widely acknowledged, the relative impact of different biases and their interplay remain insufficiently explored. Our work addresses this gap through a systematic investigation of key bias factors, their contribution to unfair outcomes, and extends to methods for their detection [1]. These are essential steps toward building a common vocabulary for describing data biases, documenting them consistently, and connecting them to fairness interventions in a principled way. Looking ahead, we call for a community-wide effort to standardize the quantitative documentation of dataset biases into a shared standardized framework that can inform and align with international anti-discrimination policy [5].

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly for spell-checking and ChatGPT to paraphrase specific sentences. We reviewed and edited the content as needed and we take full responsibility for the publication's content.

References

[1] M. Ceccon, G. Cornacchia, D. D. Pezze, A. Fabris, G. A. Susto, Underrepresentation, label bias, and proxies: Towards data bias profiles for the eu ai act and beyond, Expert Systems with Applications (2025) 128266. URL: https://www.sciencedirect.com/science/article/pii/S0957417425018858. doi:https://doi.org/10.1016/j.eswa.2025.128266.

[2] A. Vetrò, M. Torchiano, M. Mecati, A data quality approach to the identification of discrimination risk in automated decision making systems, Gov. Inf. Q. 38 (2021) 101619. URL: https://doi.org/10.1016/j.giq.2021.101619. doi:10.1016/J.GIQ.2021.101619.

[3] D. Brzezinski, J. Stachowiak, J. Stefanowski, I. Szczech, R. Susmaga, S. Aksenyuk, U. Ivashka, O. Yasinskyi, Properties of fairness measures in the context of varying class imbalance and protected group ratios, ACM Transactions on Knowledge Discovery from Data (2024).

[4] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, P. Hall, Towards a standard for identifying and managing bias in artificial intelligence, US Department of Commerce, National Institute of Standards and Technology, 2022.

[5] E. Parliament, Artificial intelligence act, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024.

[6] ISO, Information technology — artificial intelligence (ai) — bias in ai systems and ai aided decision making, 2021. https://www.iso.org/standard/77607.html.

[7] L. Guerdan, A. Coston, Z. S. Wu, K. Holstein, Ground(less) truth: A causal framework for proxy labels in human-algorithm decision-making, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023, ACM, 2023, pp. 688–704. URL: https://doi.org/10.1145/3593013.3594036. doi:10.1145/3593013.3594036.

[8] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

[9] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.

[10] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453.

[11] M. Bao, A. Zhou, S. Zottola, B. Brubach, S. Desmarais, A. Horowitz, K. Lum, S. Venkatasubramanian, It's compaslicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html.

[12] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 3315–3323. URL: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

[13] D. Madras, E. Creager, T. Pitassi, R. S. Zemel, Learning adversarially fair and transferable representations, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 3381–3390. URL: http://proceedings.mlr.press/v80/madras18a.html.