Is it Still Fair? A Comparative Evaluation of Fairness Algorithms through the Lens of Covariate Drift-Supplementary

Anonymous Author(s)*

CCS CONCEPTS

• **Computing methodologies** → **Regularization**; Multi-agent systems; • **Applied computing** → Law; *Interactive learning environments*; • **Information systems** → **Clustering and classification**.

KEYWORDS

Algorithmic Fairness, Covariate Drift, Fairness Algorithms, Robustness

ACM Reference Format:
Anonymous Author(s). 2024. Is it Still Fair? A Comparative Evaluation of Fairness Algorithms through the Lens of Covariate Drift-Supplementary. In *Proceedings of The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'24).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 COVARIATE RANKING

1.1 NWF Dataset

Table 1: Ranking of Covariates for NWF dataset.

Covariate	Co-eff/Cov-imp avg	shap avg	Overall Average
Quiz	0.273	0.257	0.265
Last login	0.267	0.247	0.257
Assignment	0.136	0.189	0.162
Folder	0.086	0.087	0.086
Forum	0.078	0.085	0.082
Resource	0.058	0.056	0.057
Url	0.039	0.027	0.033
Course	0.029	0.035	0.032
Study period	0.024	0.010	0.017
SMS	0.010	0.006	0.008

Co-eff/Cov-imp avg is the average of the co-efficient weights and covariate importance scores across all 4 models whereas shap avg is the average of SHAP values across all 4 models.

1.2 ITF Dataset

Table 2: Ranking of Covariates for ITF dataset.

Covariate	Co-eff/Cov-imp avg	shap avg	Overall Average
Last Login	0.479	0.402	0.440
Page	0.097	0.140	0.119
Course	0.088	0.104	0.096
Quiz	0.085	0.097	0.091
Resource	0.066	0.092	0.079
Forum	0.062	0.082	0.072
Study Period	0.053	0.039	0.046
Url	0.030	0.034	0.032
SMS	0.042	0.011	0.027

Co-eff/Cov-imp avg is the average of the co-efficient weights and covariate importance scores across all 4 models whereas shap avg is the average of SHAP values across all 4 models.

1.3 BAF Dataset

Table 3: Ranking of Covariates for BAF dataset.

Covariate	Co-eff/Cov-imp	shap avg	OVR AVG
keep_alive_session	0.098	0.120	0.109
phone_home_valid	0.081	0.104	0.092
has_other_cards	0.060	0.076	0.068
prev_address_months_count	0.057	0.073	0.065
credit_risk_score	0.061	0.066	0.064
income	0.066	0.054	0.060
name_email_similarity	0.069	0.047	0.058
date_of_birth_distinct_emails_4w	0.050	0.049	0.050
current_address_months_count	0.042	0.057	0.050
phone_mobile_valid	0.041	0.037	0.039
email_is_free	0.036	0.042	0.039
device_distinct_emails_8w	0.040	0.030	0.035
bank_branch_count_8w	0.030	0.035	0.032
velocity_4w	0.029	0.027	0.028
proposed_credit_limit	0.026	0.027	0.026
customer_age	0.024	0.027	0.026
days_since_request	0.029	0.021	0.025
intended_balcon_amount	0.028	0.021	0.024
zip_count_4w	0.025	0.018	0.022
velocity_24h	0.024	0.017	0.020
bank_months_count	0.021	0.018	0.020
session_length_in_minutes	0.022	0.015	0.018
velocity_6h	0.022	0.015	0.018
foreign_request	0.016	0.006	0.011
device_fraud_count	0.000	0.000	0.000

Co-eff/Cov-imp avg is the average of the co-efficient weights and covariate importance scores across all 4 models whereas shap avg is the average of SHAP values across all 4 models.

SIGIR'24, July 14–18, 2024, Washington D.C., USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 COVARIATE DRIFTS

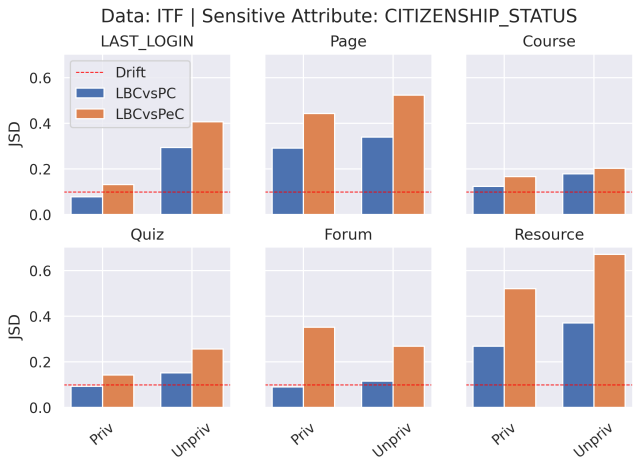

Figure 1: Drifts of top-6 important covariates for ITF dataset. P = privileged group, UP = unprivileged group . LBC= long before covid (t_0). PC = pre-covid (t_1), and PeC = peri-covid (t_2).

3 DCD VS. FAIRNESS

3.1 NWF Dataset

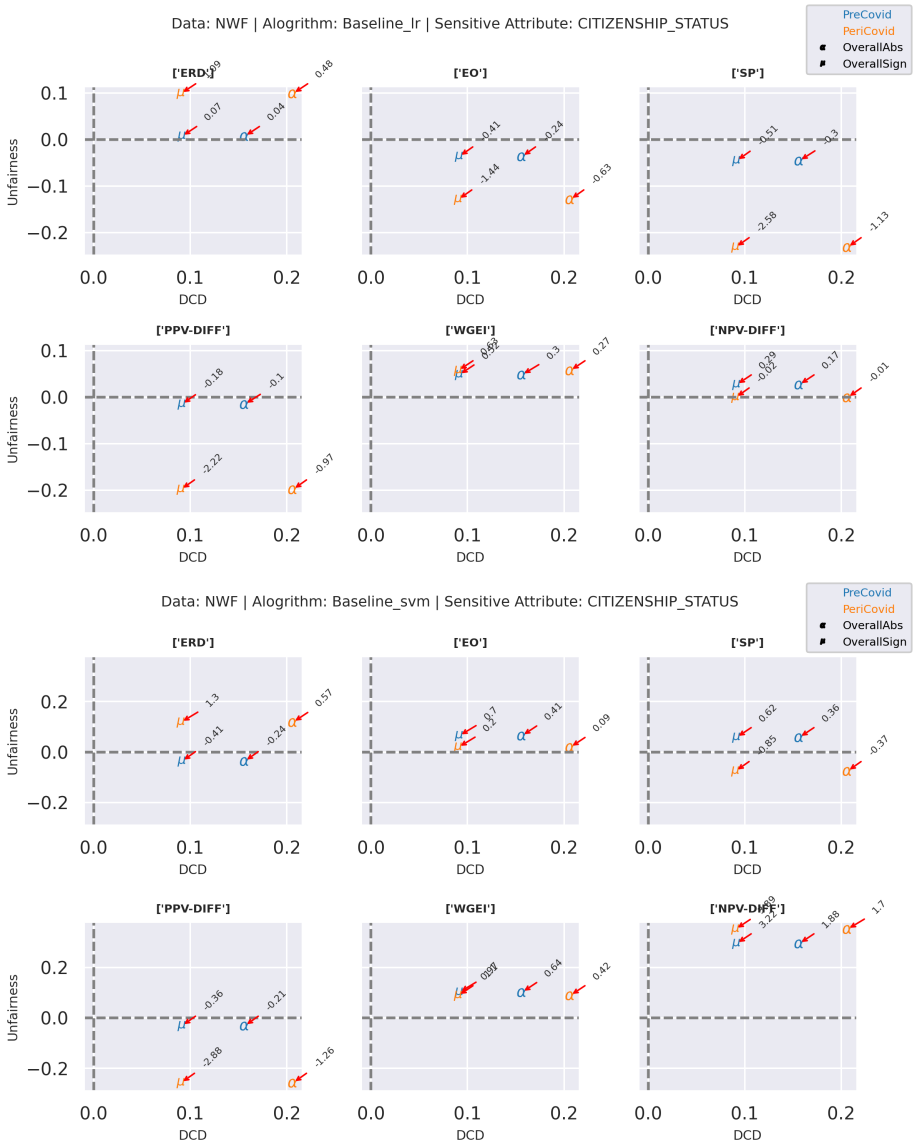

Figure 2: DCD vs unfairness for baseline models for NWF dataset.

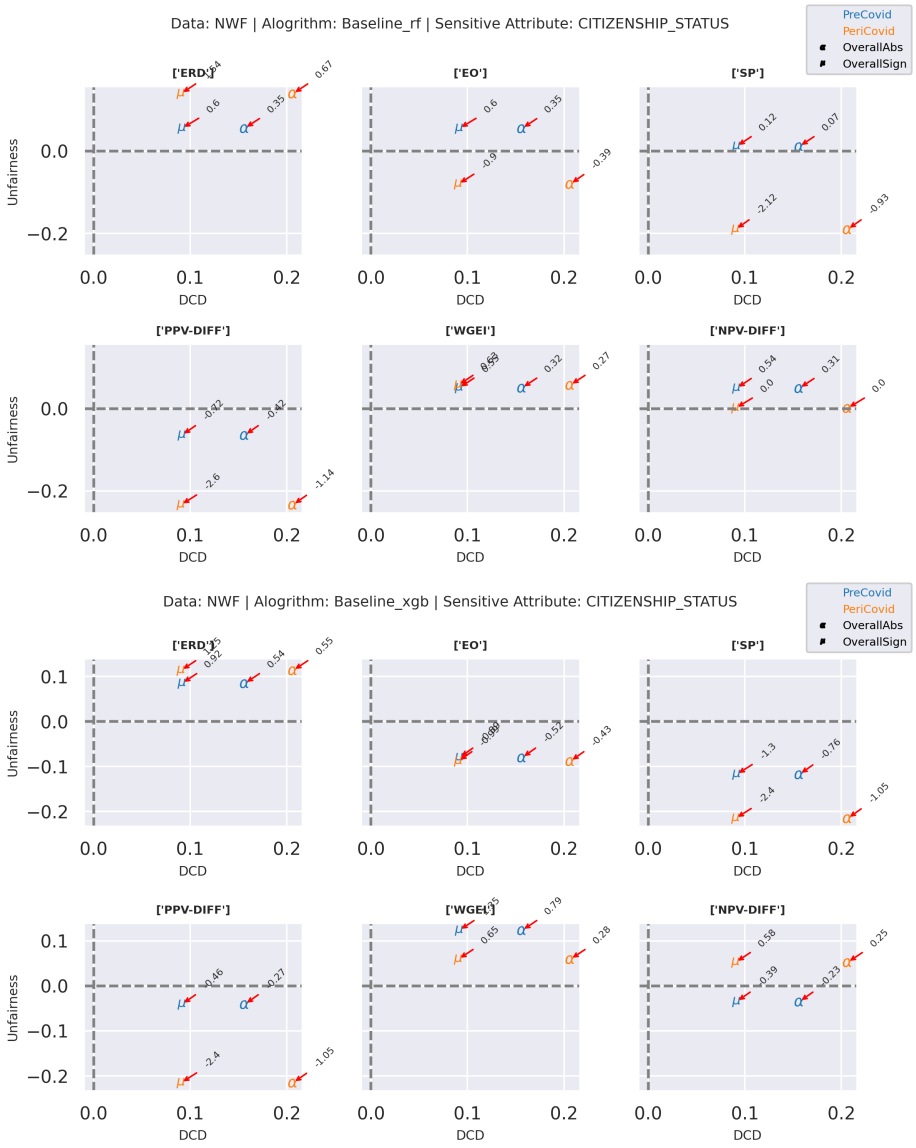

Figure 3: DCD vs unfairness for baseline models for NWF dataset.

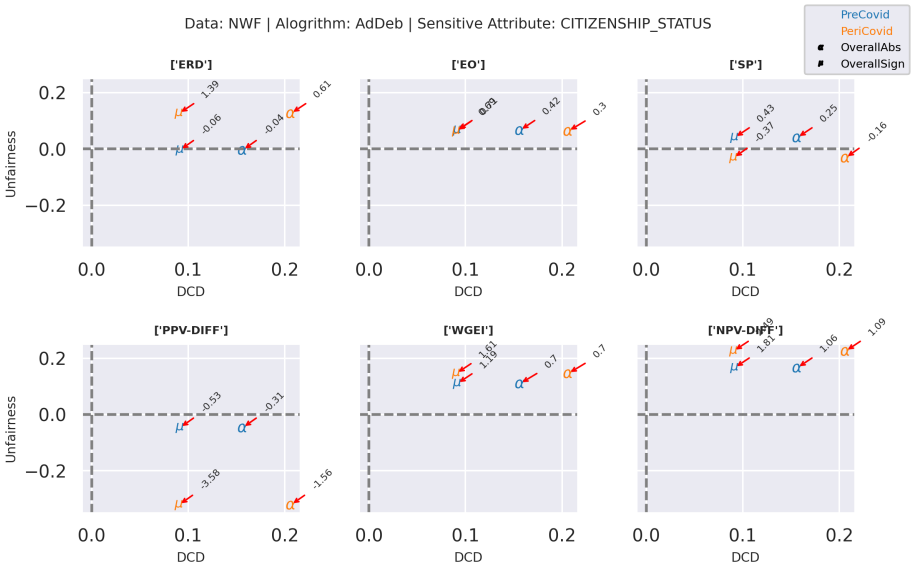

Figure 4: DCD vs unfairness for AdDeb for NWF dataset.

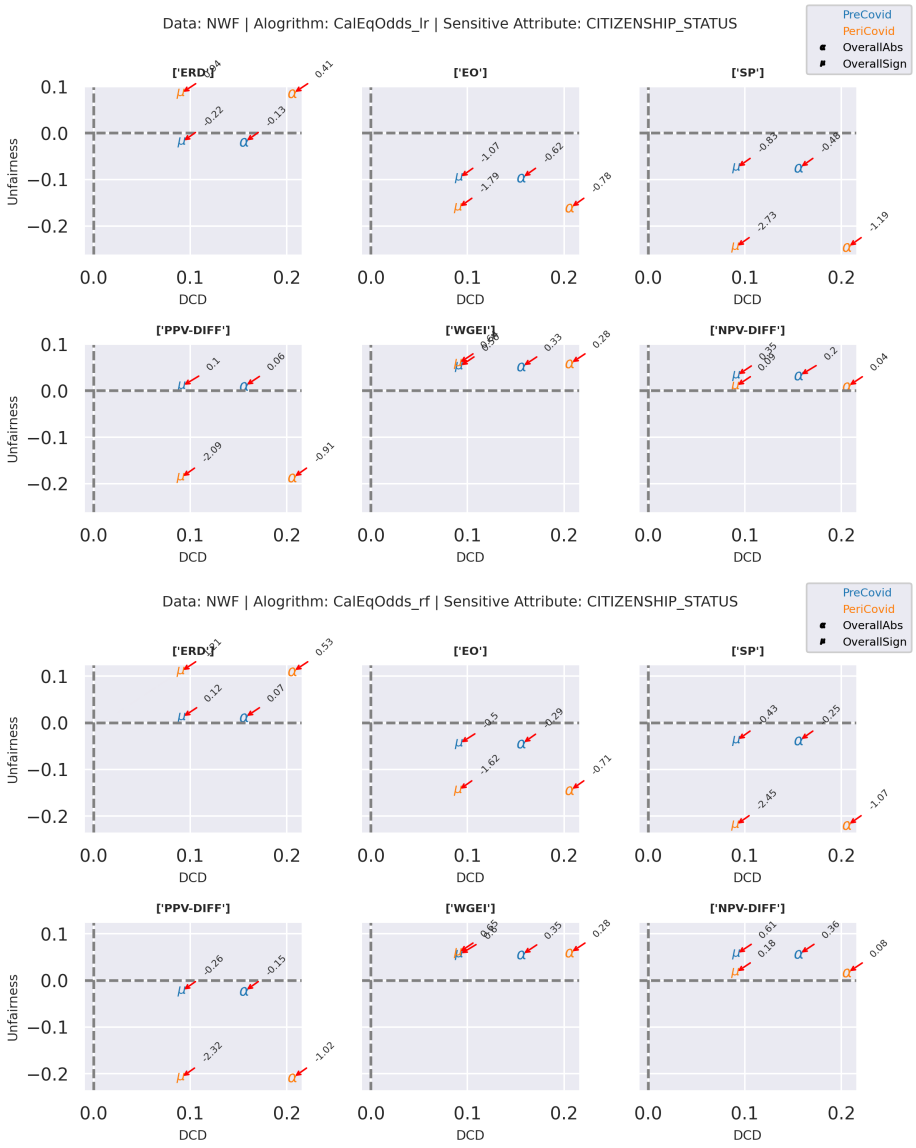

Figure 5: DCD vs unfairness for CalEqOdds for NWF dataset.

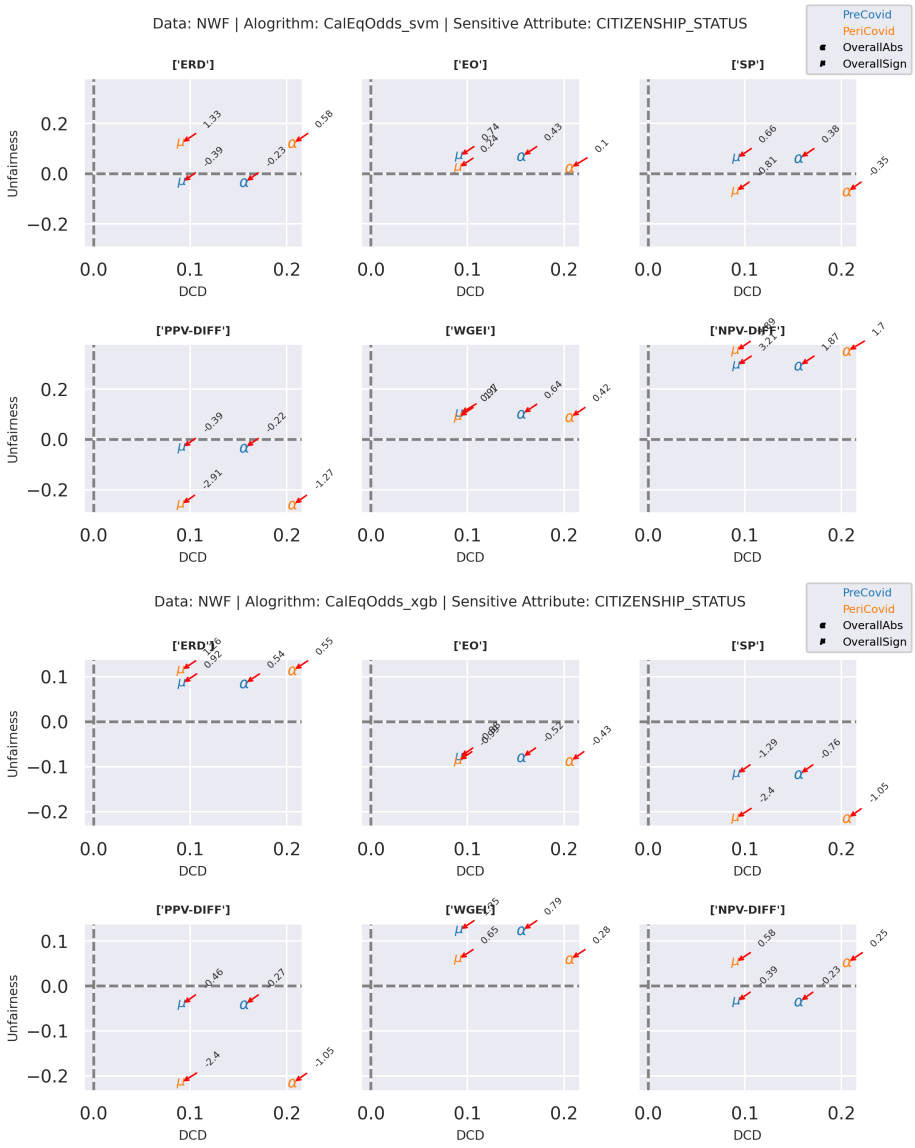

Figure 6: DCD vs unfairness for CalEqOdds for NWF dataset.

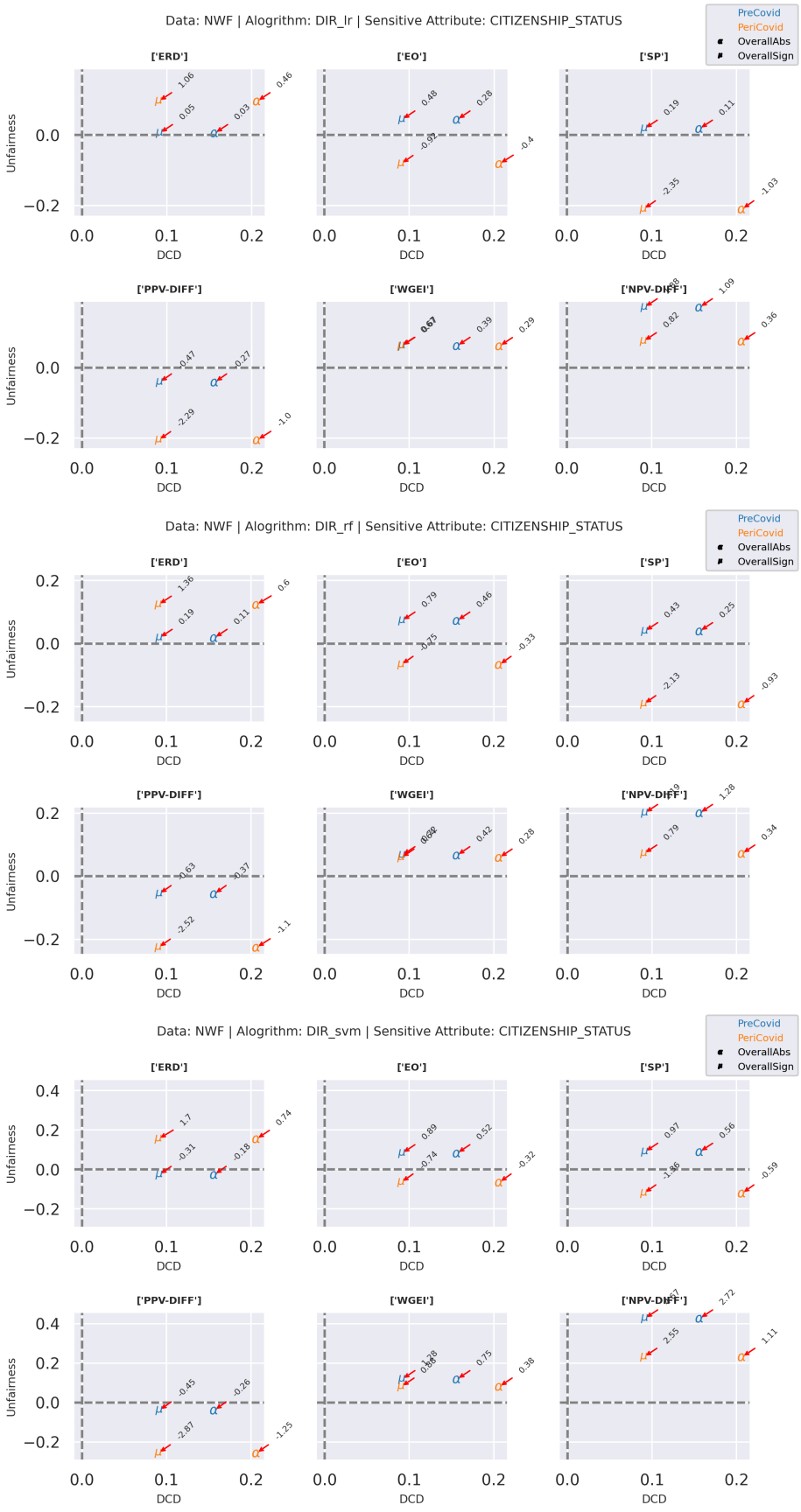

Figure 7: DCD vs unfairness for DIR for NWF dataset.

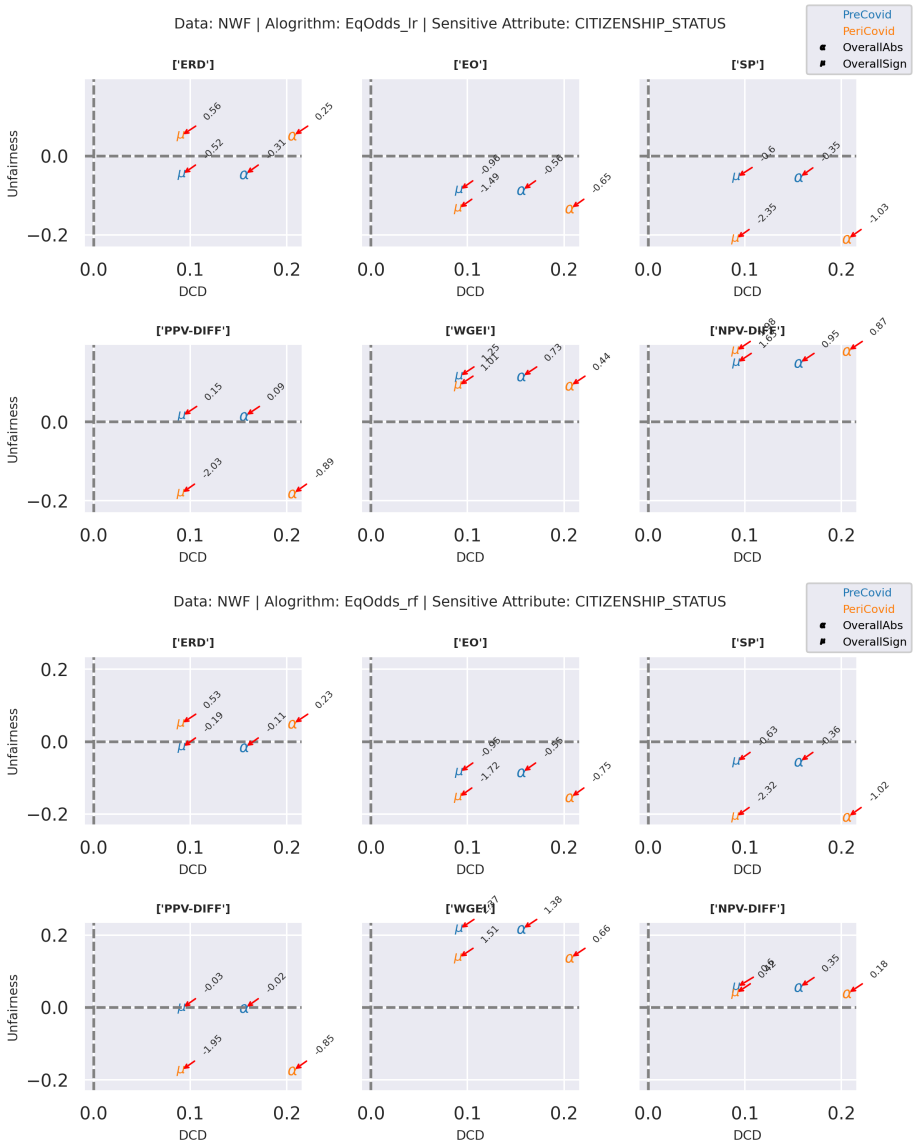

Figure 8: DCD vs unfairness for EqOdds for NWF dataset.

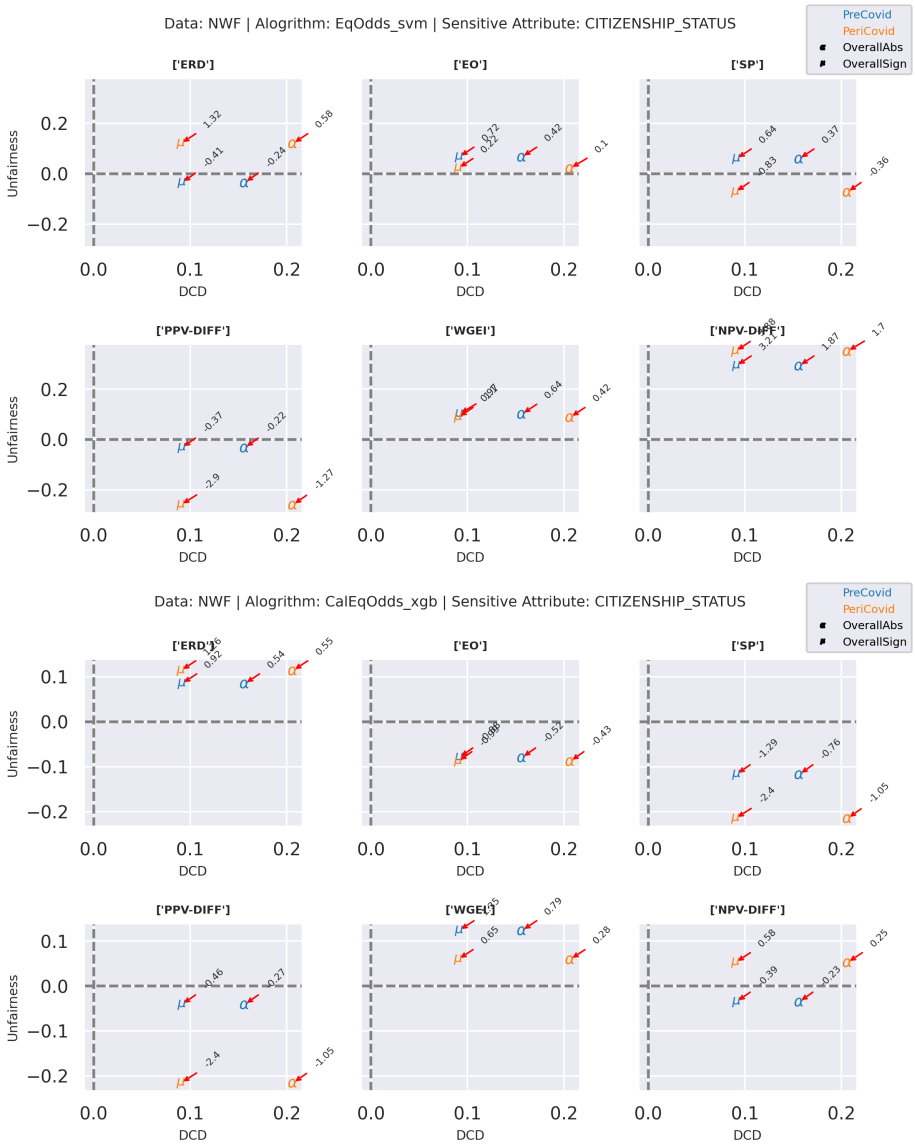

Figure 9: DCD vs unfairness for EqOdds for NWF dataset.