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, 41 pages. https://doi.org/XXXXXXX.XXXXXXX

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

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

1 COVARIATE RANKING

1.1 BAF Dataset

Table 1: 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.

1.2 NWF Dataset

Table 2: 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.

1.3 ITF Dataset

Table 3: 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
Forum	0.066	0.092	0.079
Resource	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.

1.4 COMPAS Dataset

Table 4: Ranking of Covariates for COMPAS dataset.

Covariate	Co-eff/Cov-imp avg	shap avg	Overall Average
V_decile_score	0.228	0.299	0.264
Priors_count	0.183	0.234	0.209
Decile_score	0.114	0.131	0.122
Days_b_screening_arrest	0.125	0.086	0.106
Age	0.094	0.094	0.094
C_days_from_compas	0.043	0.041	0.042
Juv_misd_count	0.067	0.016	0.042
C_charge_degree	0.033	0.037	0.035
Sex	0.034	0.023	0.028
Juv_fel_count	0.026	0.011	0.018
Juv_other_count	0.018	0.004	0.011

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.

1.5 Adult Dataset

Table 5: Ranking of Covariates for Adult dataset.

Covariate	Co-eff/Cov-imp avg	shap avg	Overall Average
Marital-status	0.194	0.239	0.216
Capital-gain	0.195	0.120	0.158
Education-num	0.188	0.117	0.152
Age	0.102	0.128	0.115
Education	0.062	0.085	0.074
Hours-per-week	0.065	0.080	0.073
Occupation	0.047	0.065	0.056
Capital-loss	0.045	0.037	0.041
Relationship	0.032	0.050	0.041
Gender	0.021	0.034	0.028
Workclass	0.020	0.020	0.020
Race	0.018	0.019	0.018
Native-country	0.010	0.007	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.

2 COVARIATE DRIFTS

2.1 ITF Dataset

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).

2.2 Adult Dataset

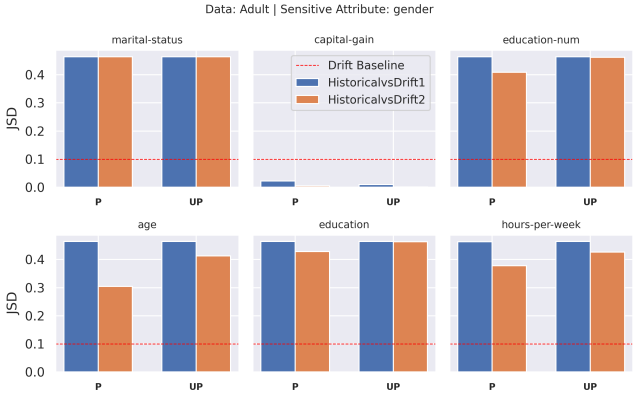

Figure 2: Drifts of top-6 important covariates for Adult dataset. P = privileged group, UP = unprivileged group.

3 DCD VS. FAIRNESS

3.1 NWF Dataset

3.2 ITF Dataset

3.3 BAF Dataset

3.4 COMPAS Dataset

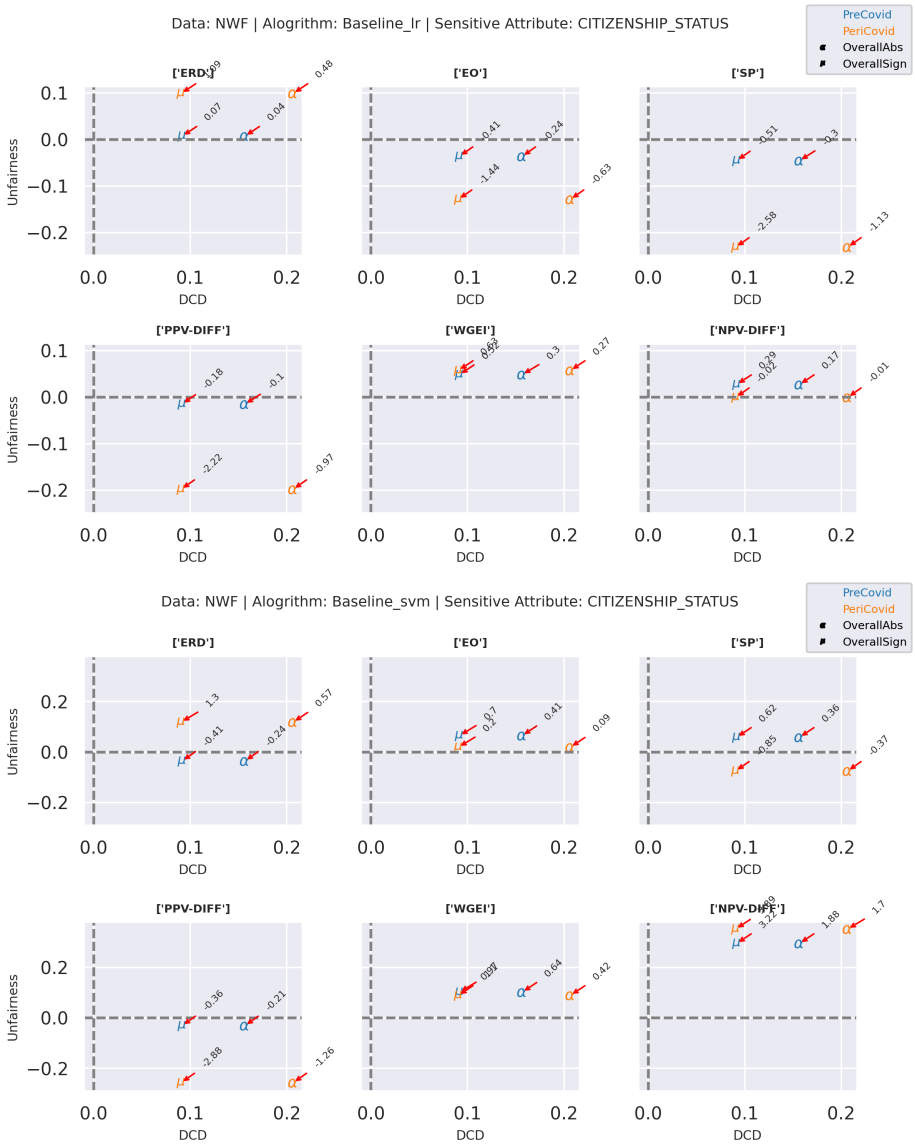

Figure 3: DCD vs unfairness for baseline models for NWF dataset.

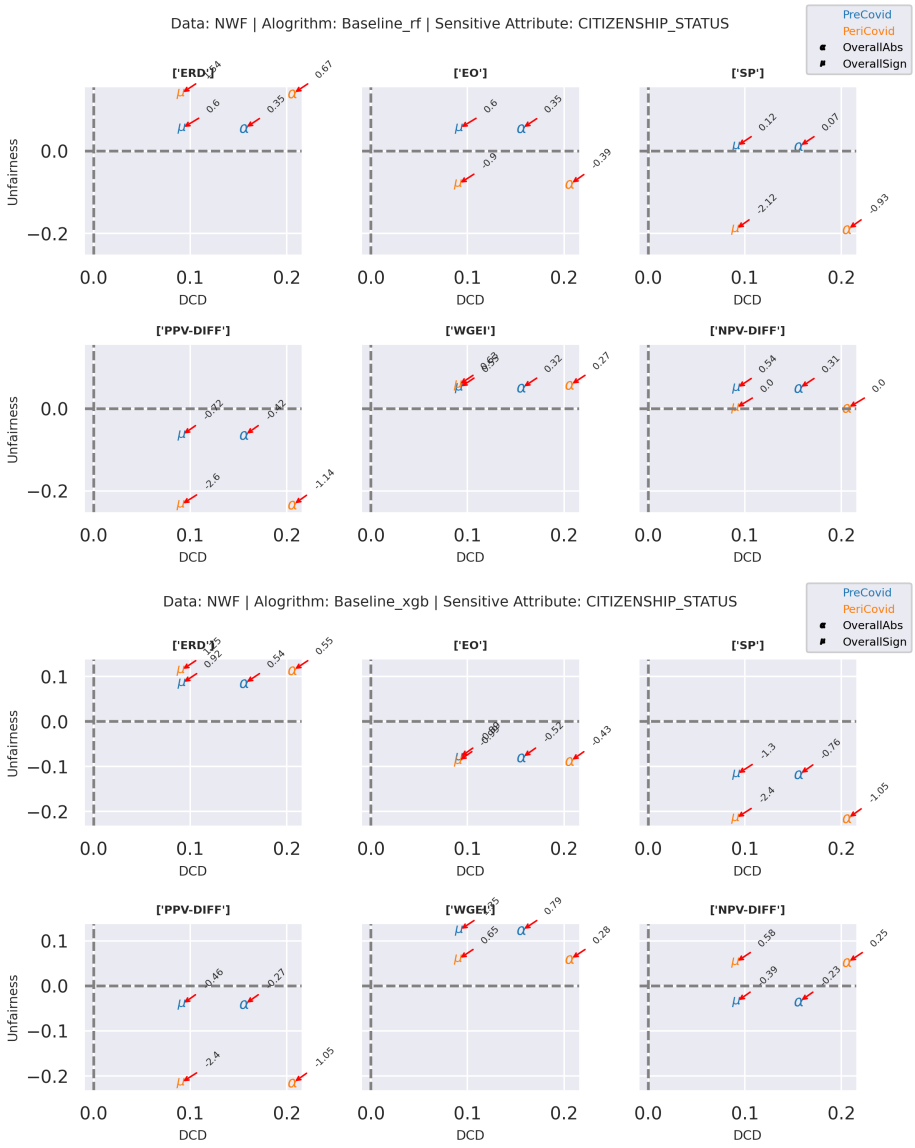

Figure 4: DCD vs unfairness for baseline models for NWF dataset.

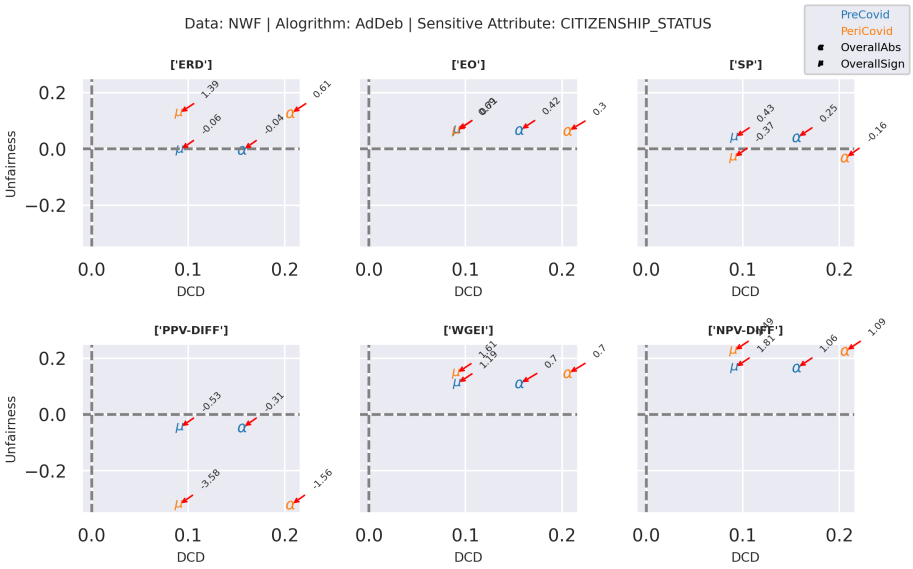

Figure 5: DCD vs unfairness for AdDeb for NWF dataset.

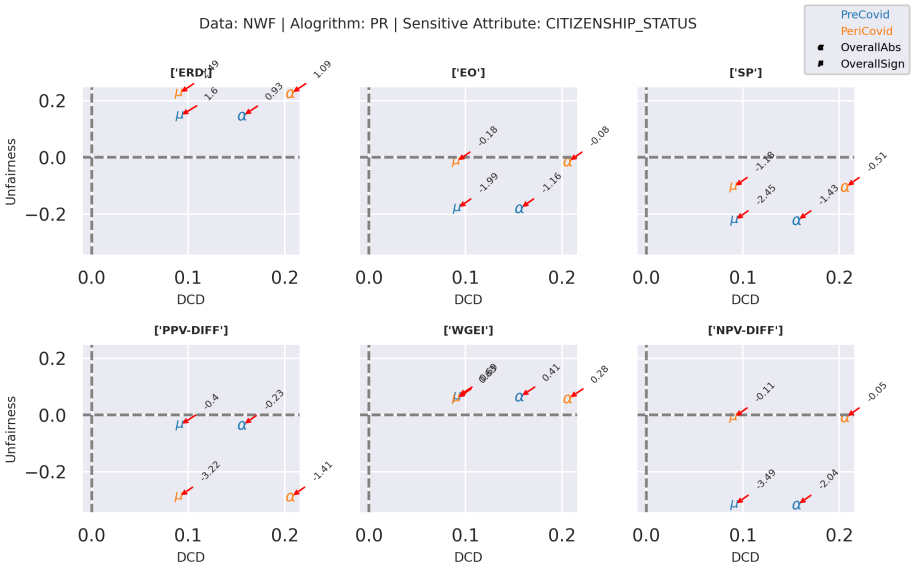

Figure 6: DCD vs unfairness for PR for NWF dataset.

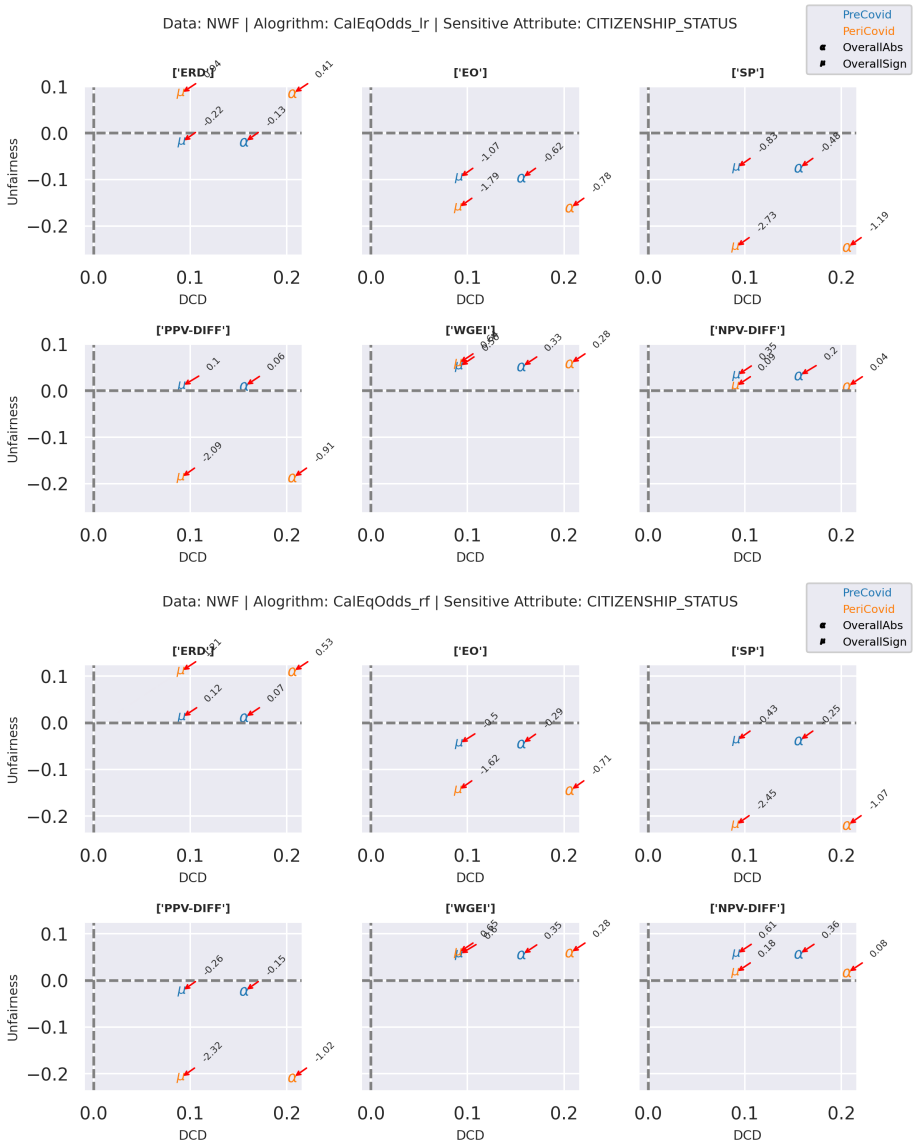

Figure 7: DCD vs unfairness for CalEqOdds for NWF dataset.

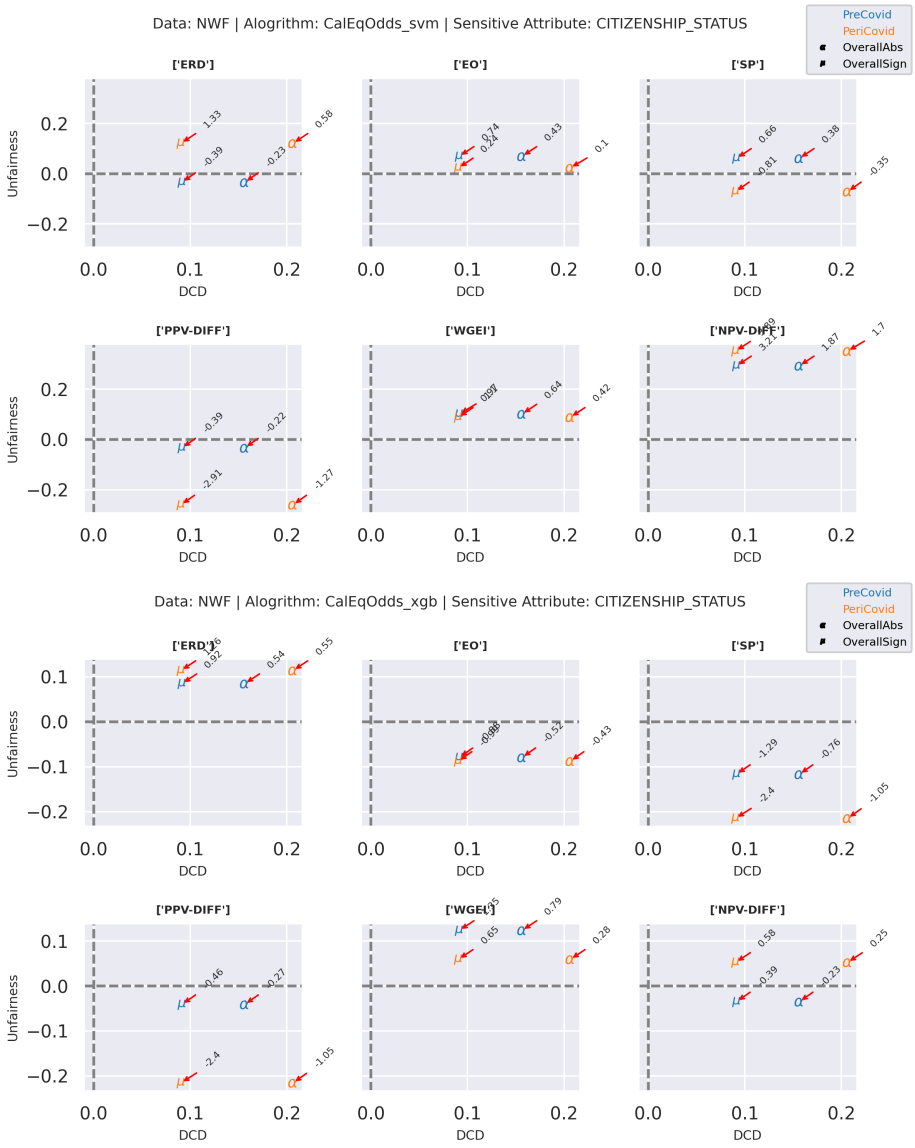

Figure 8: DCD vs unfairness for CalEqOdds for NWF dataset.

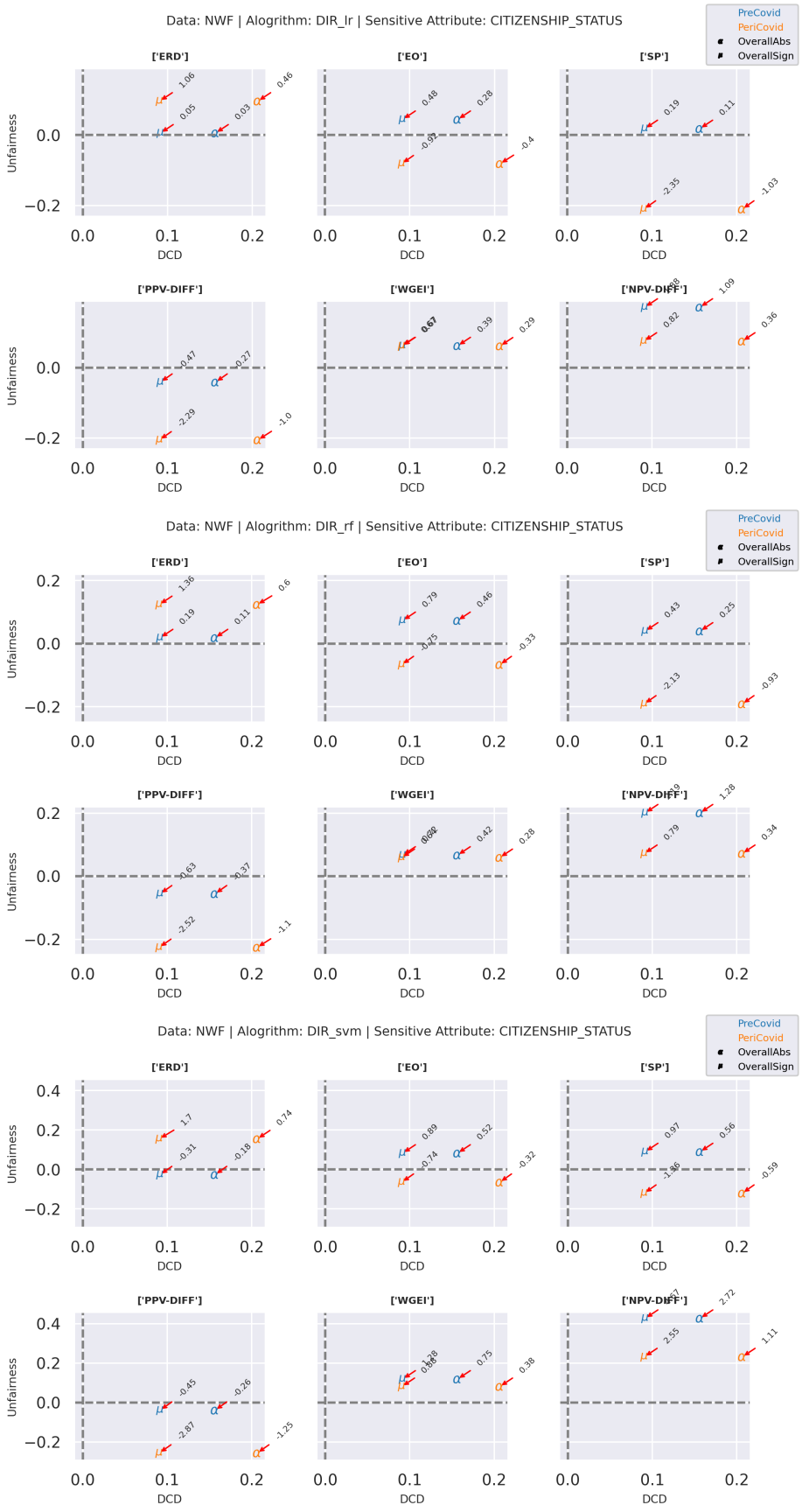

Figure 9: DCD vs unfairness for DIR for NWF dataset.

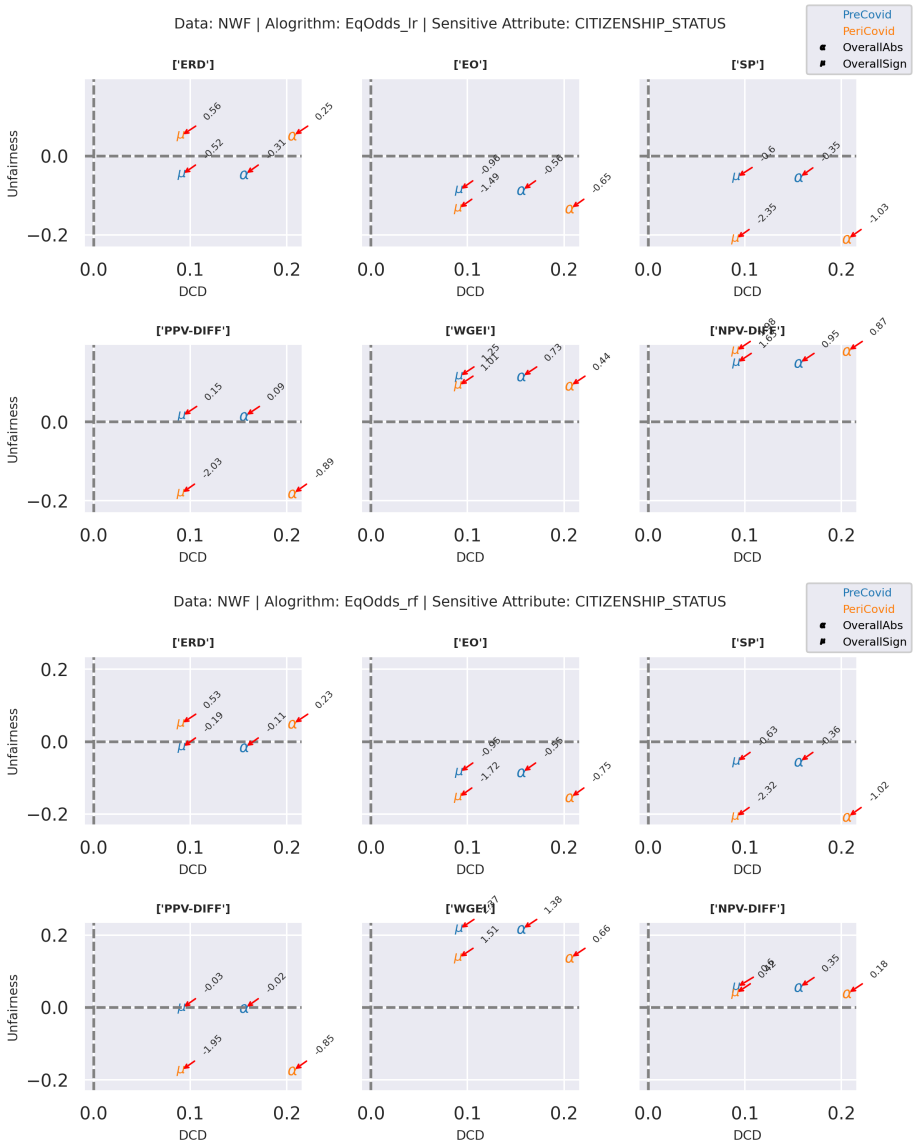

Figure 10: DCD vs unfairness for EqOdds for NWF dataset.

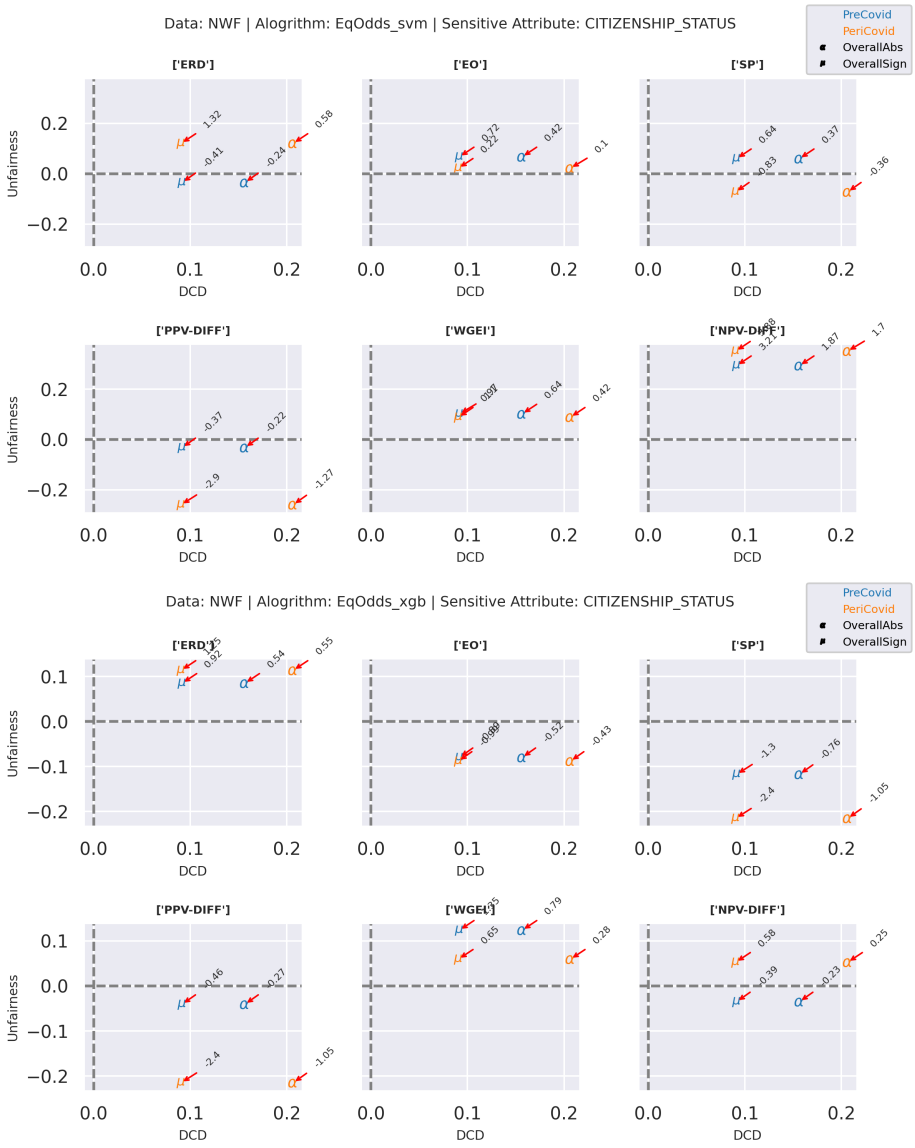

Figure 11: DCD vs unfairness for EqOdds for NWF dataset.

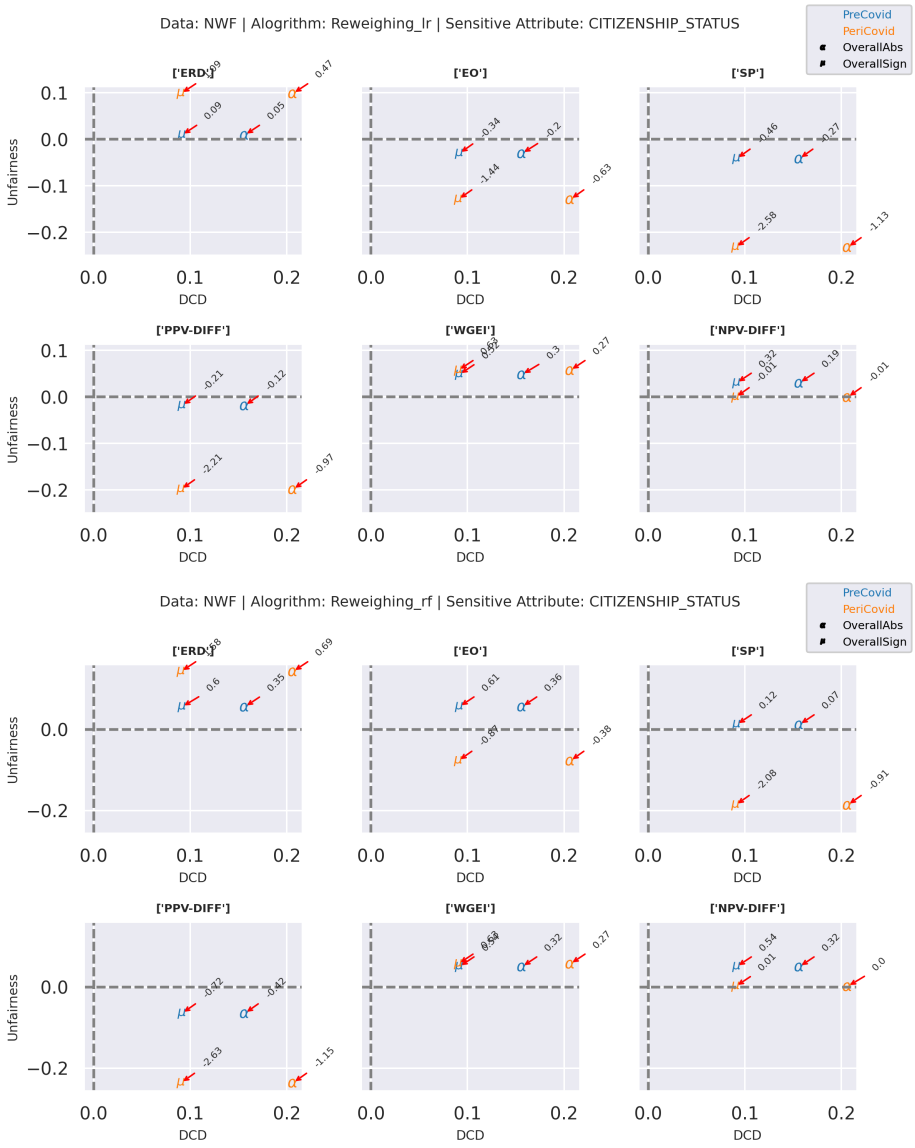

Figure 12: DCD vs unfairness for Reweighing for NWF dataset.

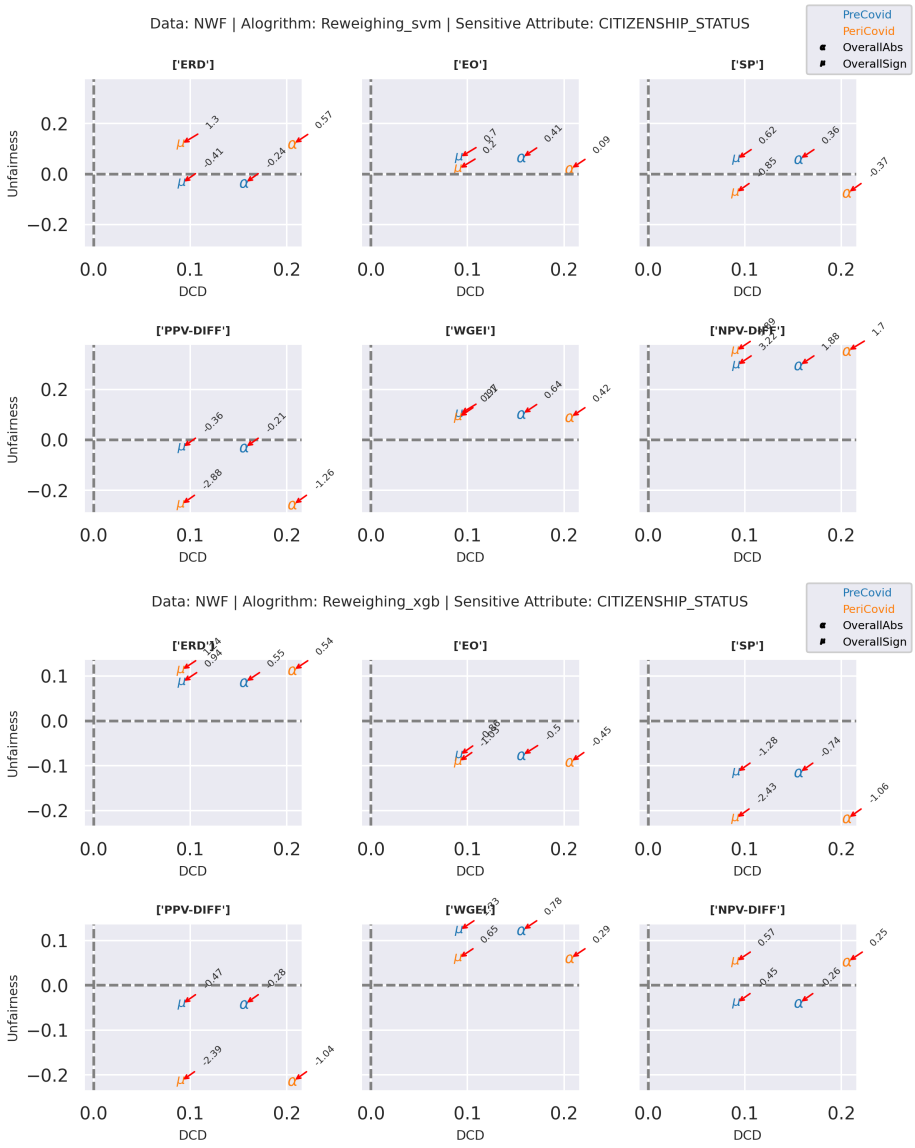

Figure 13: DCD vs unfairness for Reweighing for NWF dataset.

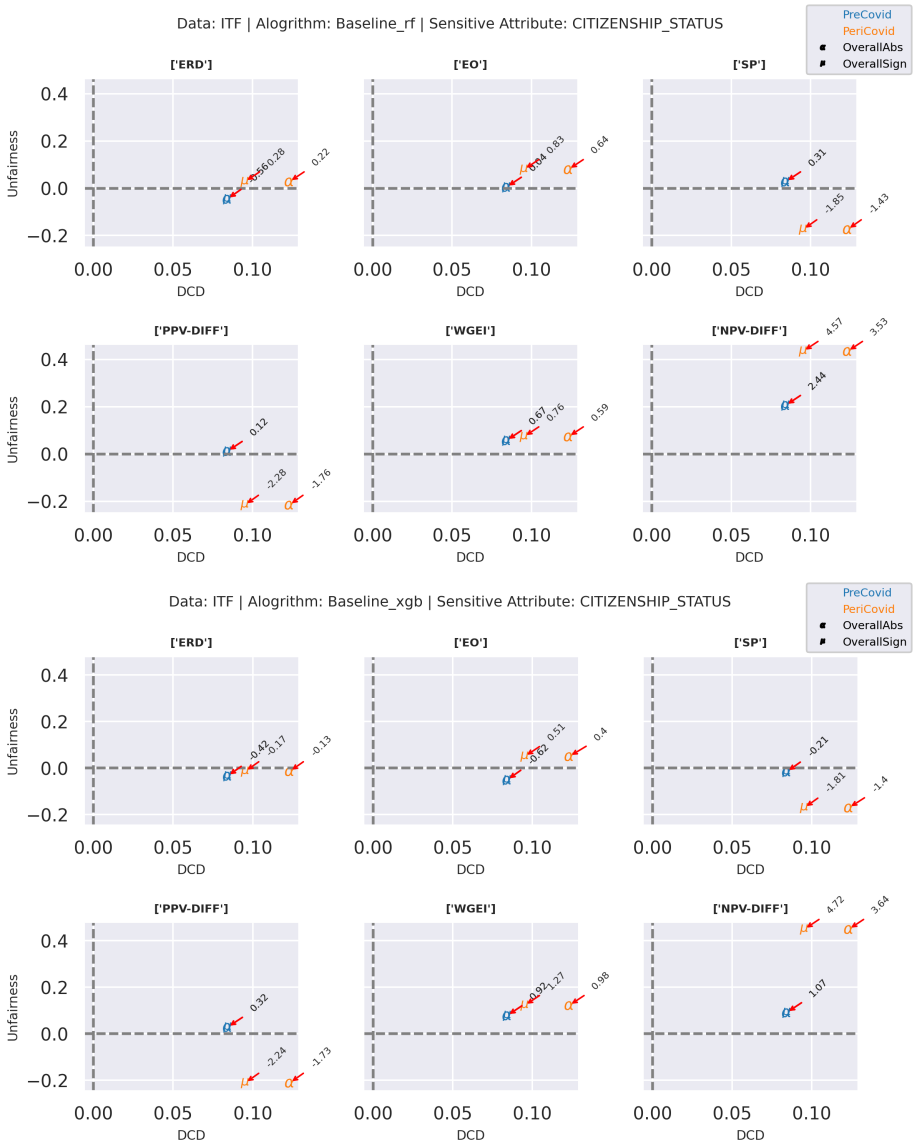

Figure 15: DCD vs unfairness for baseline models for ITF dataset.

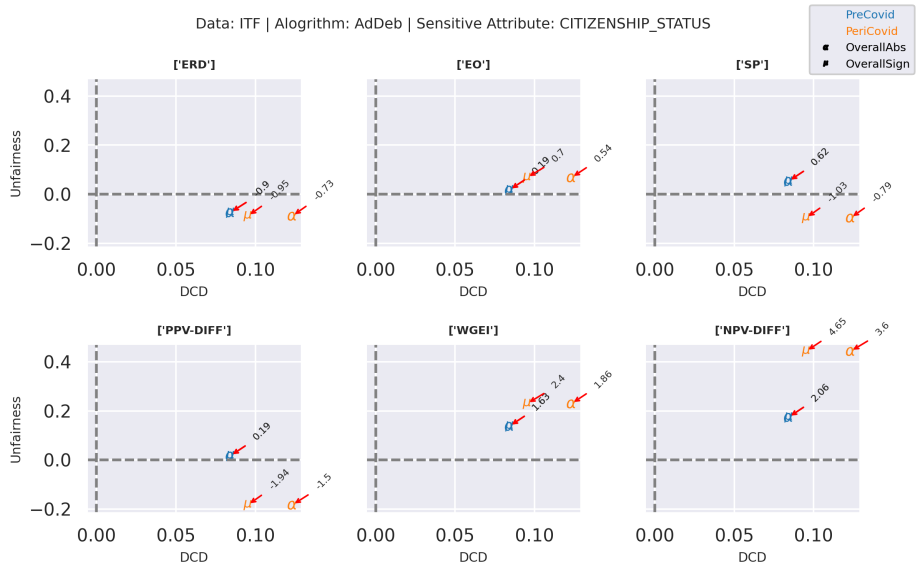

Figure 16: DCD vs unfairness for AdDeb for ITF dataset.

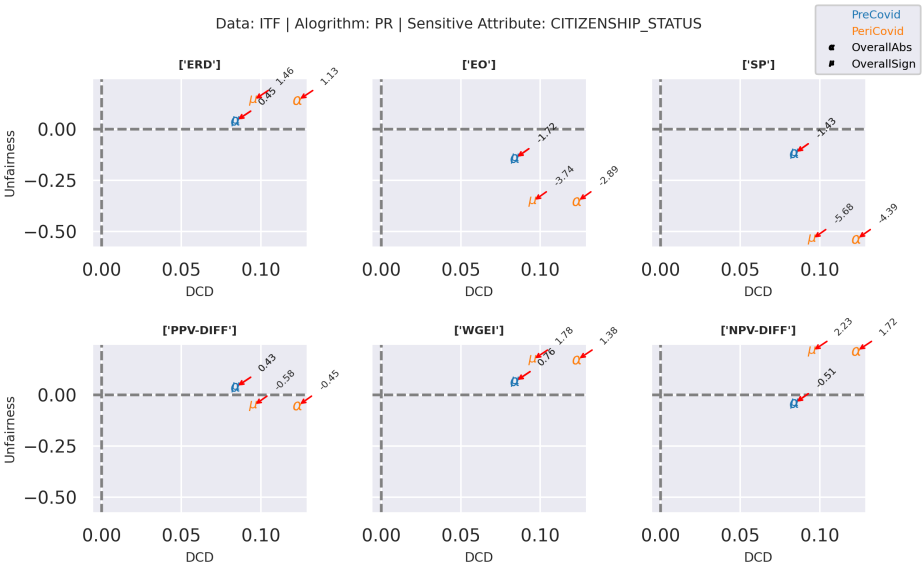

Figure 17: DCD vs unfairness for PR for ITF dataset.

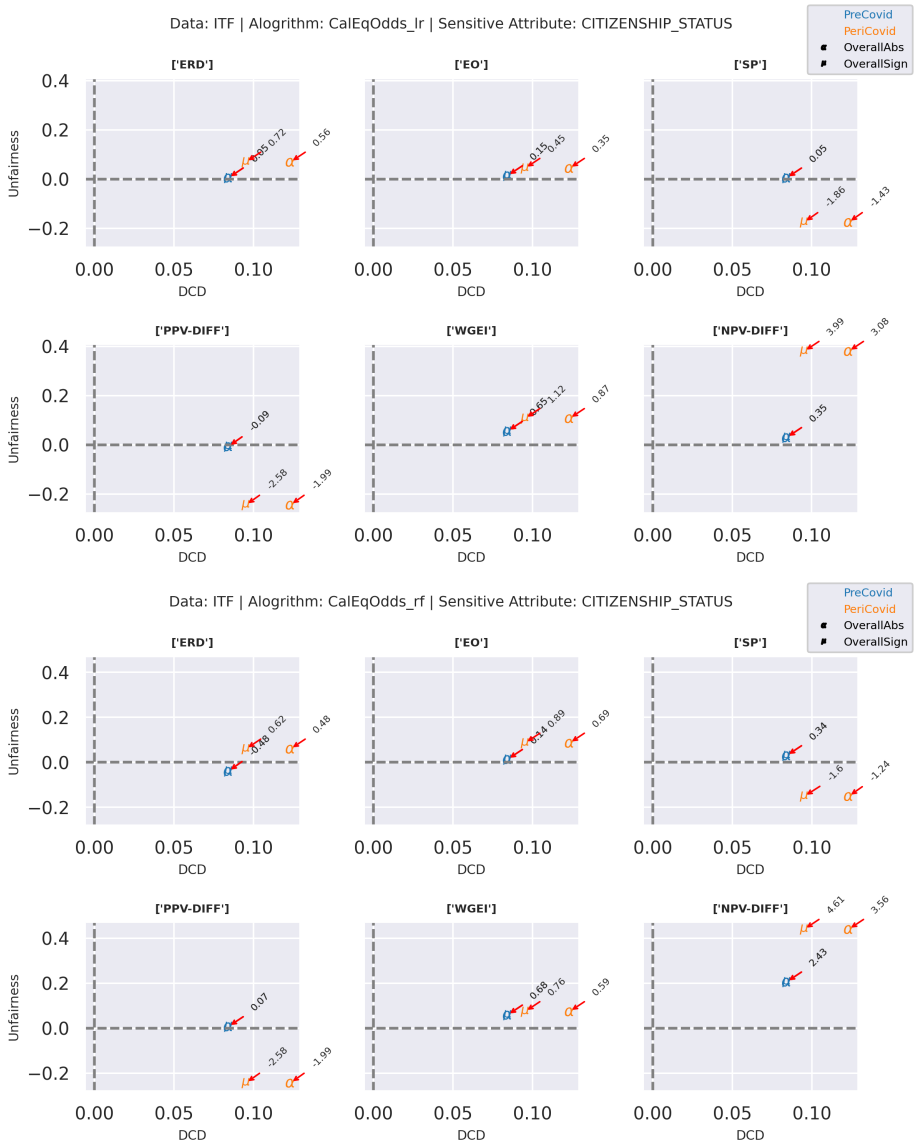

Figure 18: DCD vs unfairness for CalEqOdds for ITF dataset.

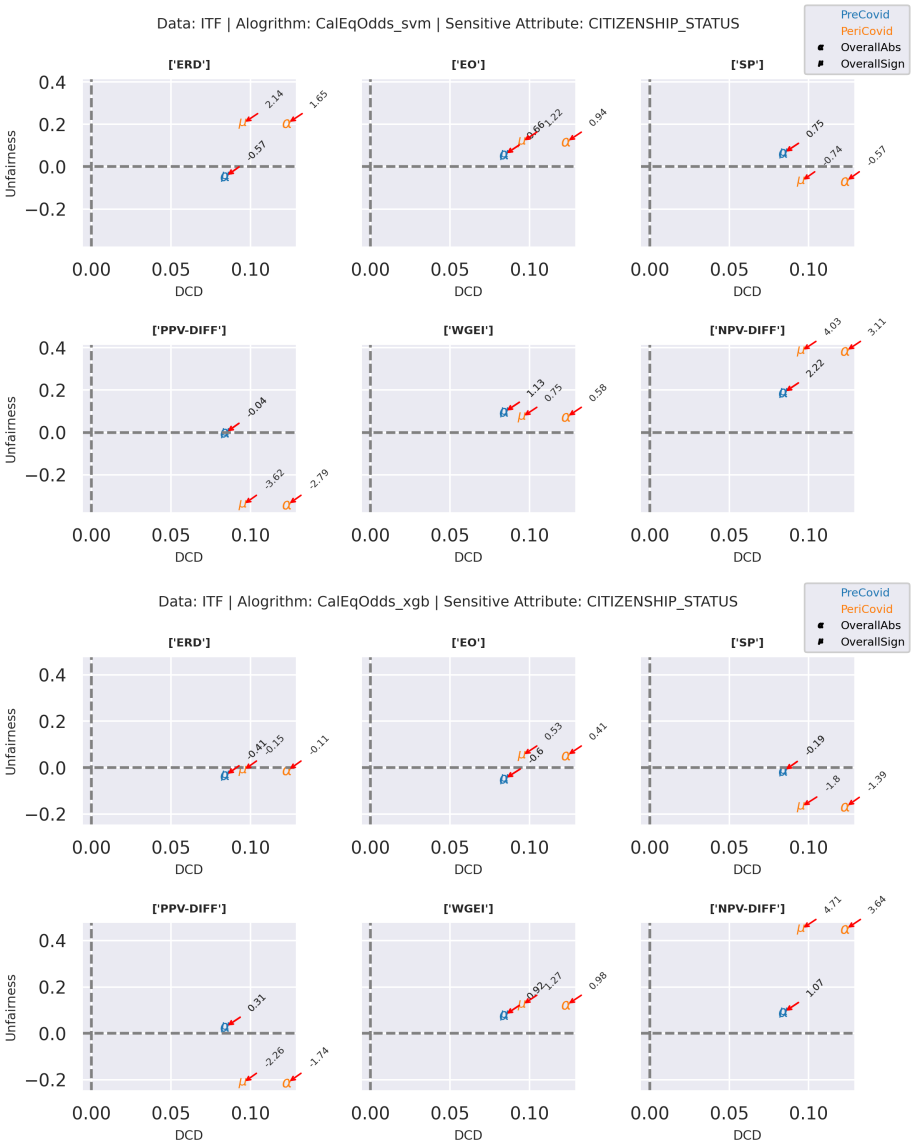

Figure 19: DCD vs unfairness for CalEqOdds for ITF dataset.

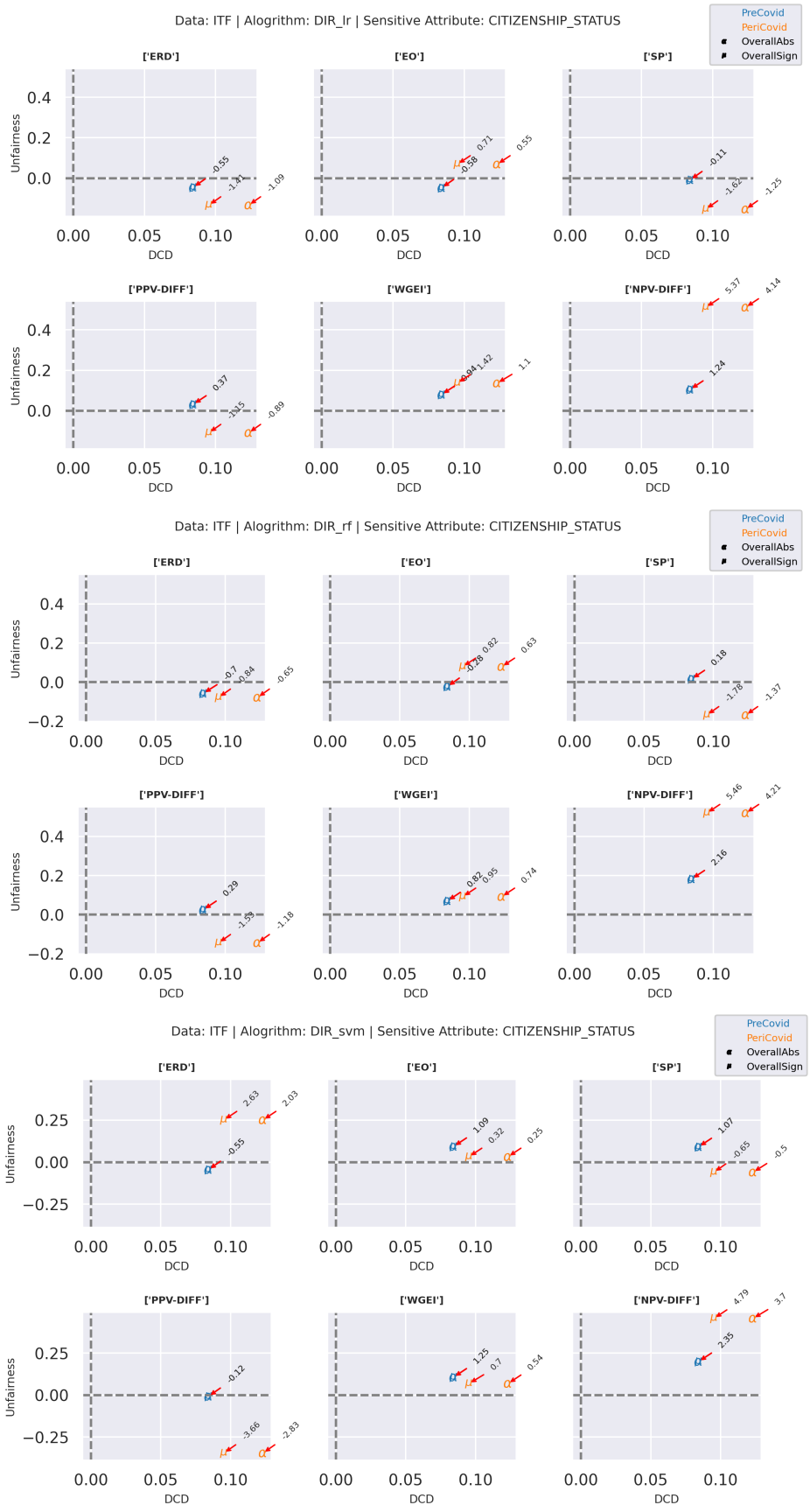

Figure 20: DCD vs unfairness for DIR for ITF dataset.

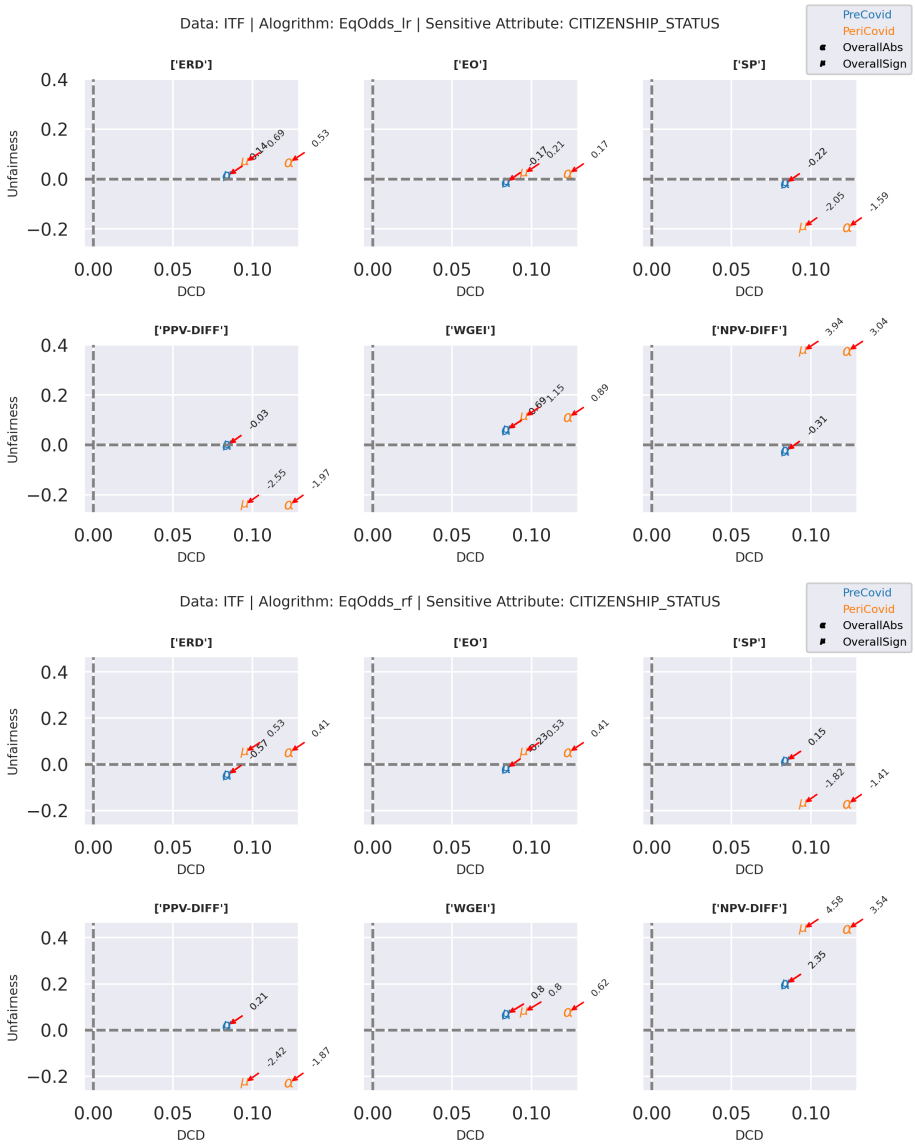

Figure 21: DCD vs unfairness for EqOdds for ITF dataset.

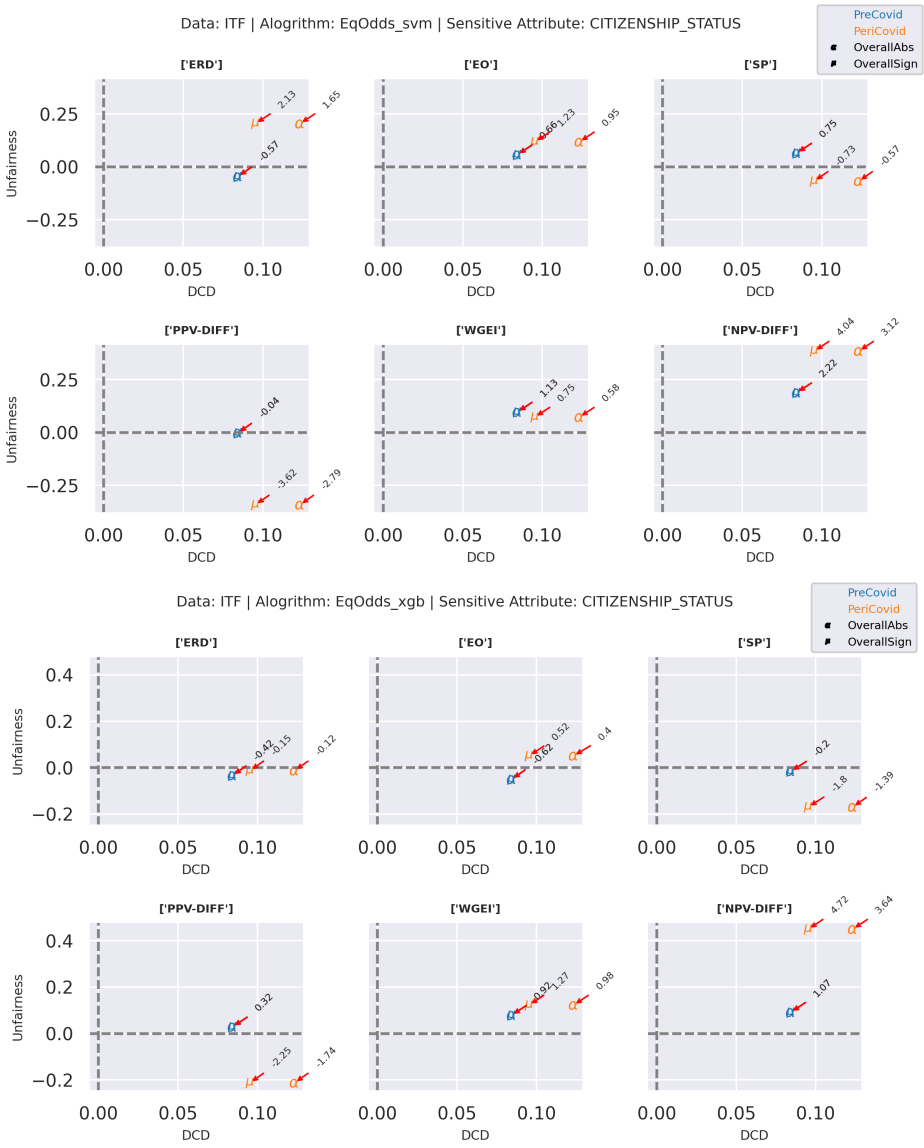

Figure 22: DCD vs unfairness for EqOdds for ITF dataset.

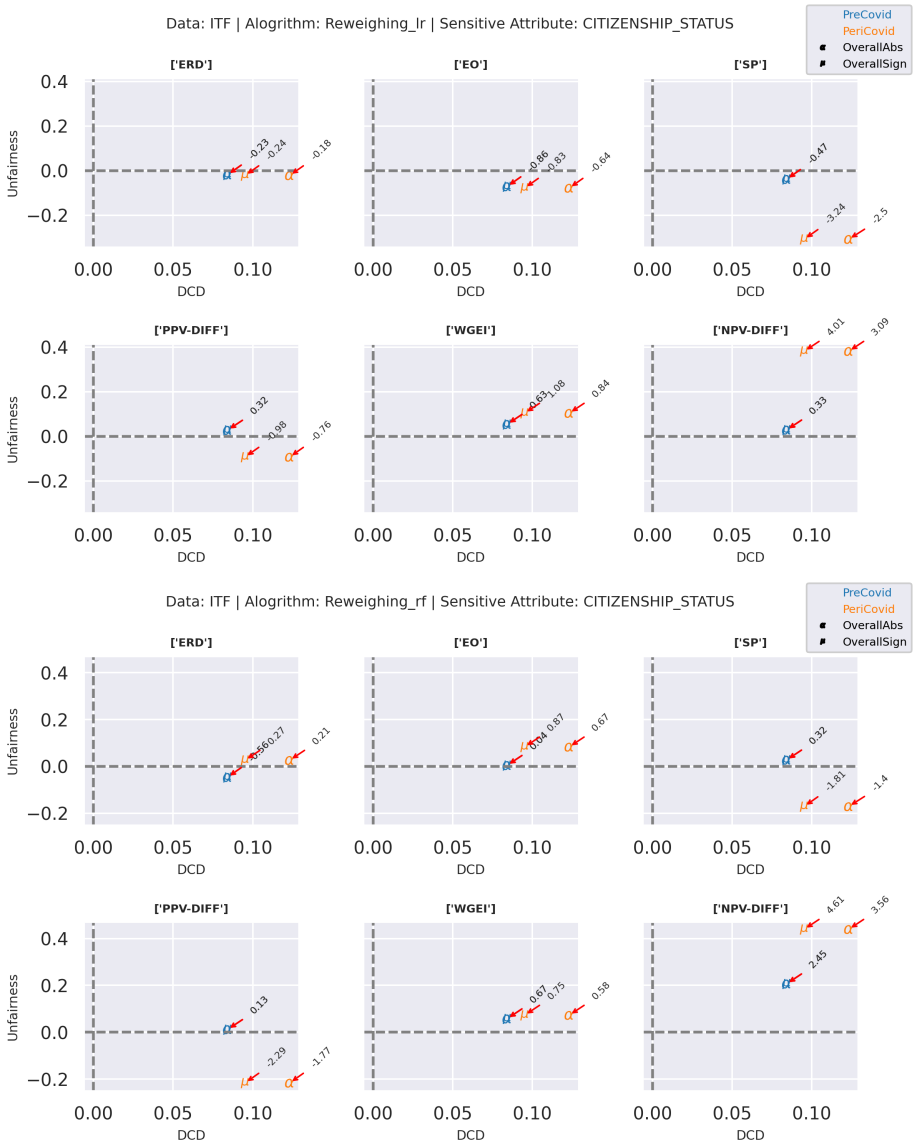

Figure 23: DCD vs unfairness for Reweighing for ITF dataset.

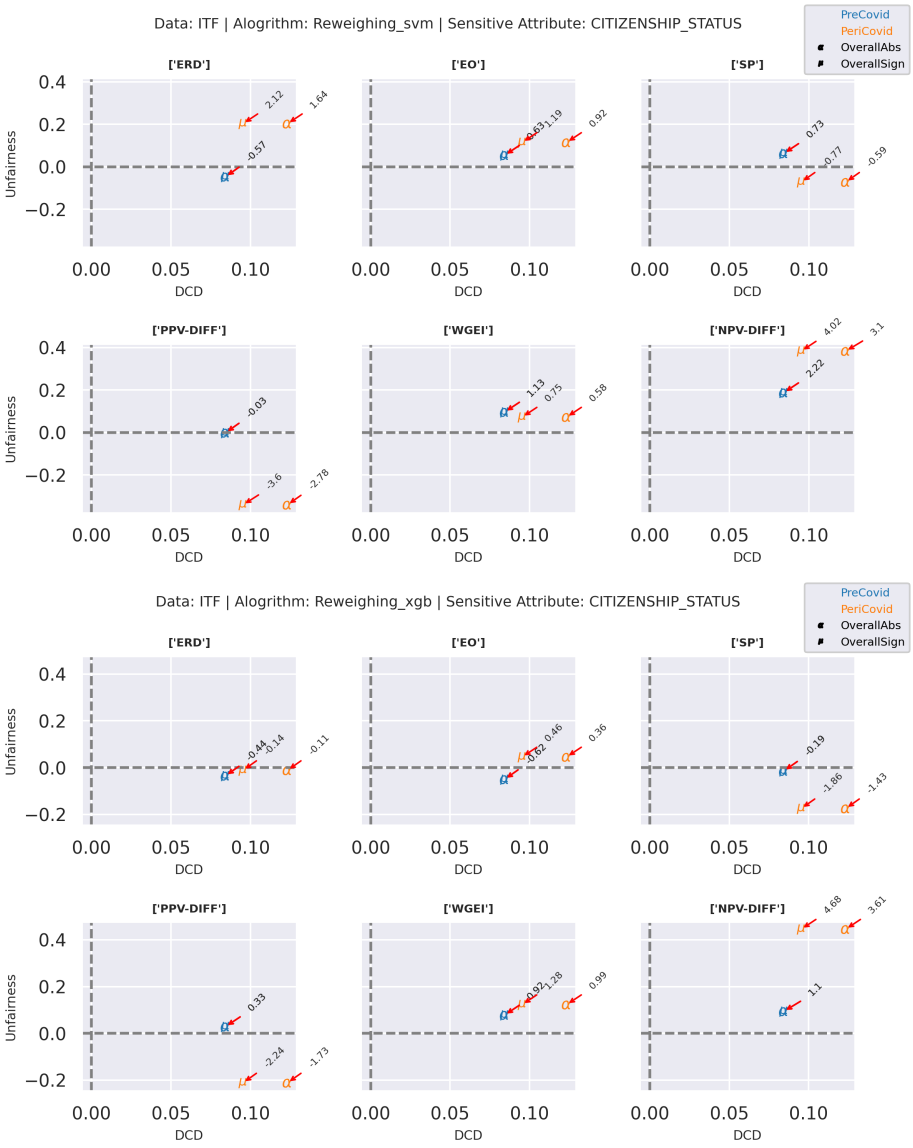

Figure 24: DCD vs unfairness for Reweighing for ITF dataset.

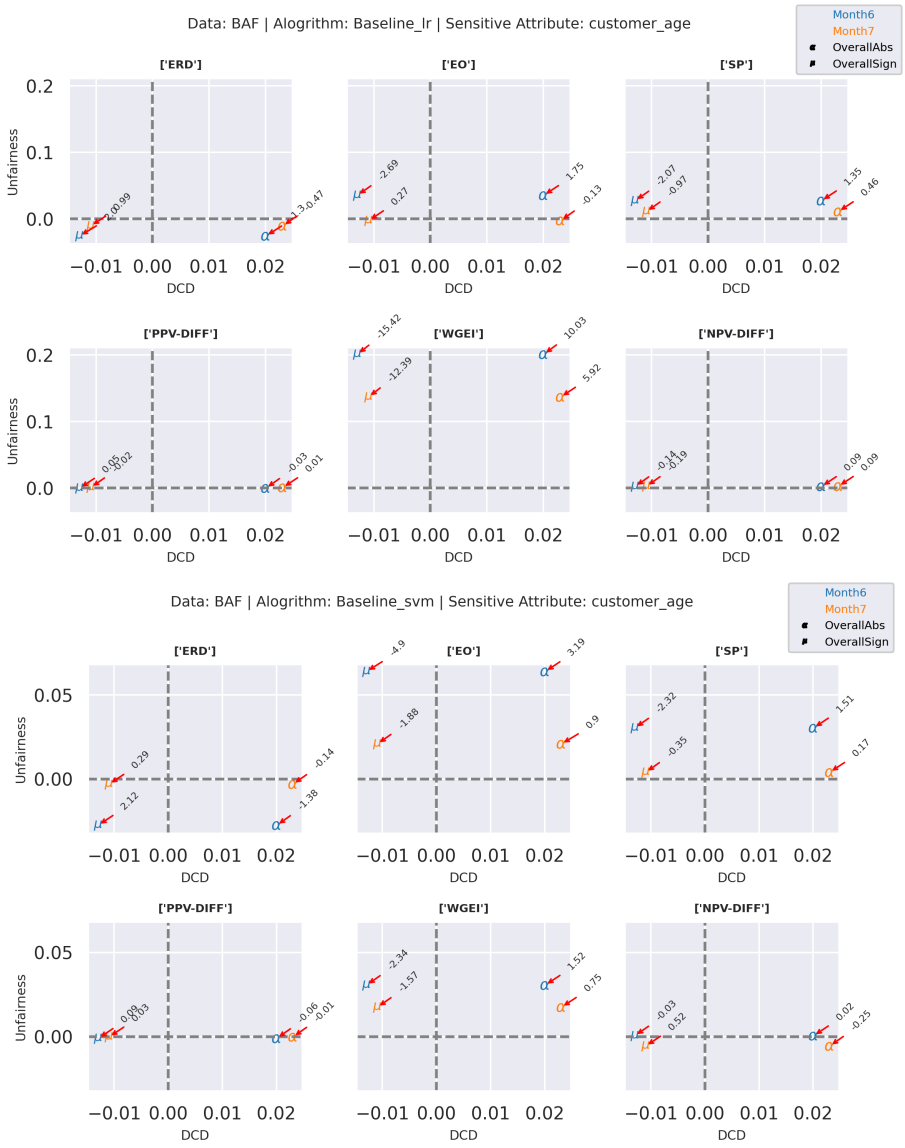

Figure 25: DCD vs unfairness for baseline models for BAF dataset.

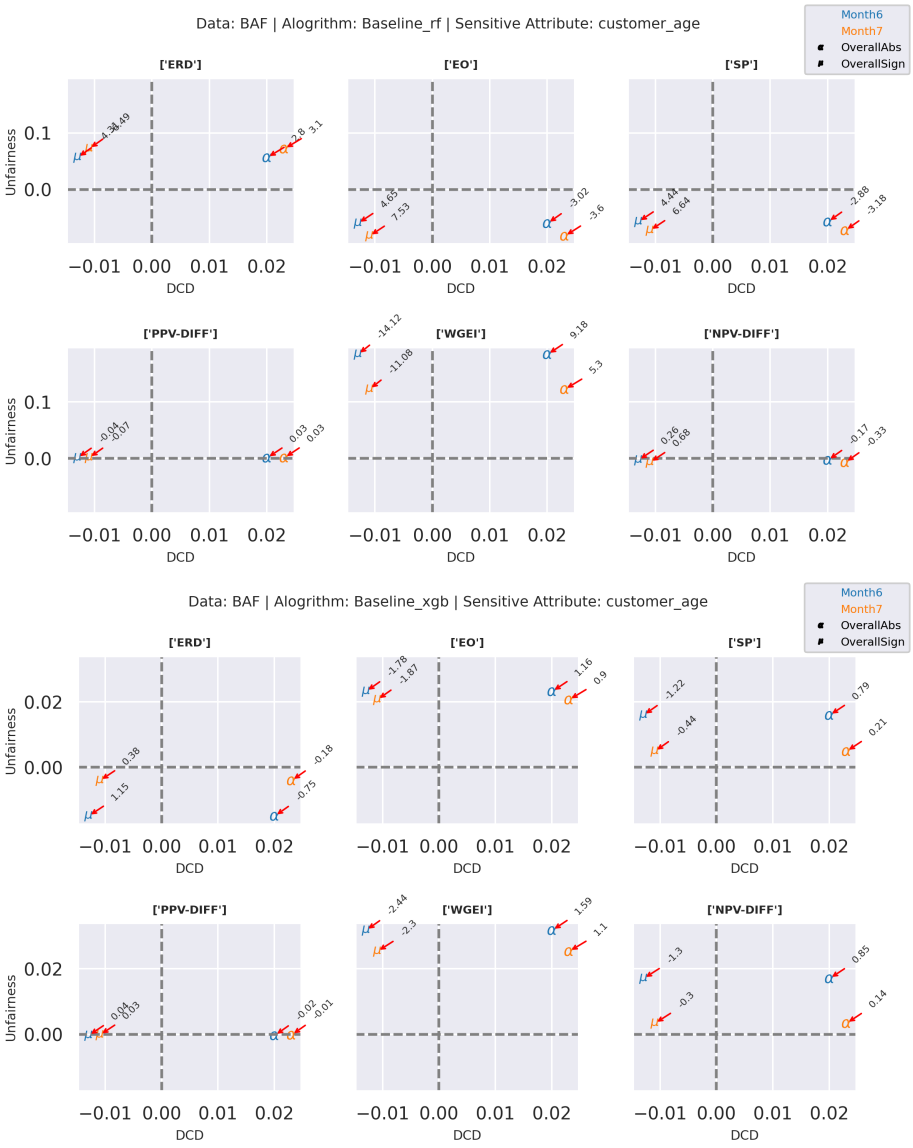

Figure 26: DCD vs unfairness for baseline models for BAF dataset.

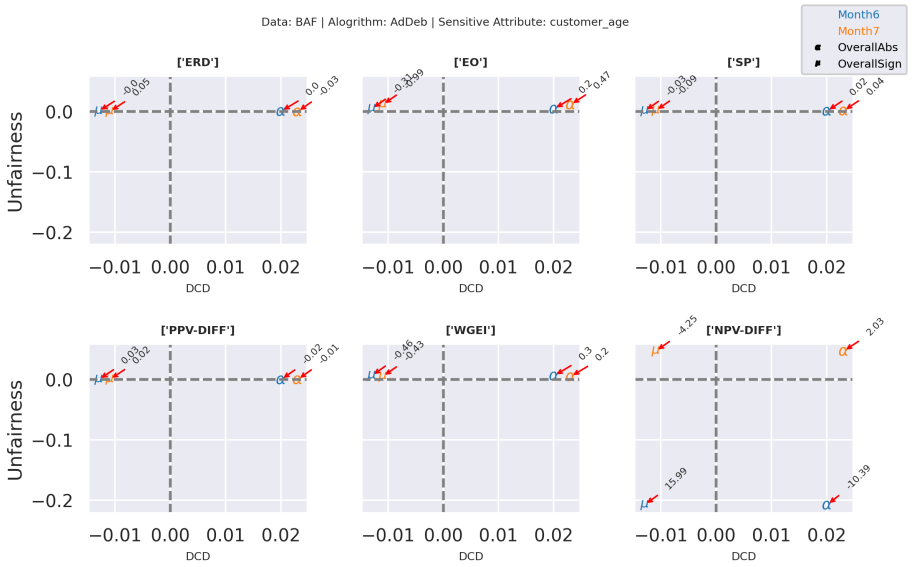

Figure 27: DCD vs unfairness for AdDeb for BAF dataset.

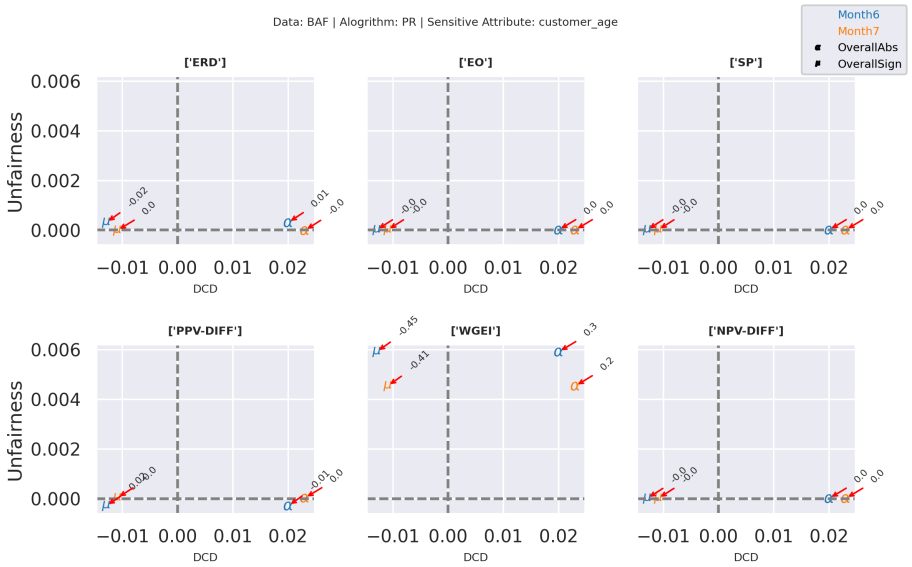

Figure 28: DCD vs unfairness for PR for BAF dataset.

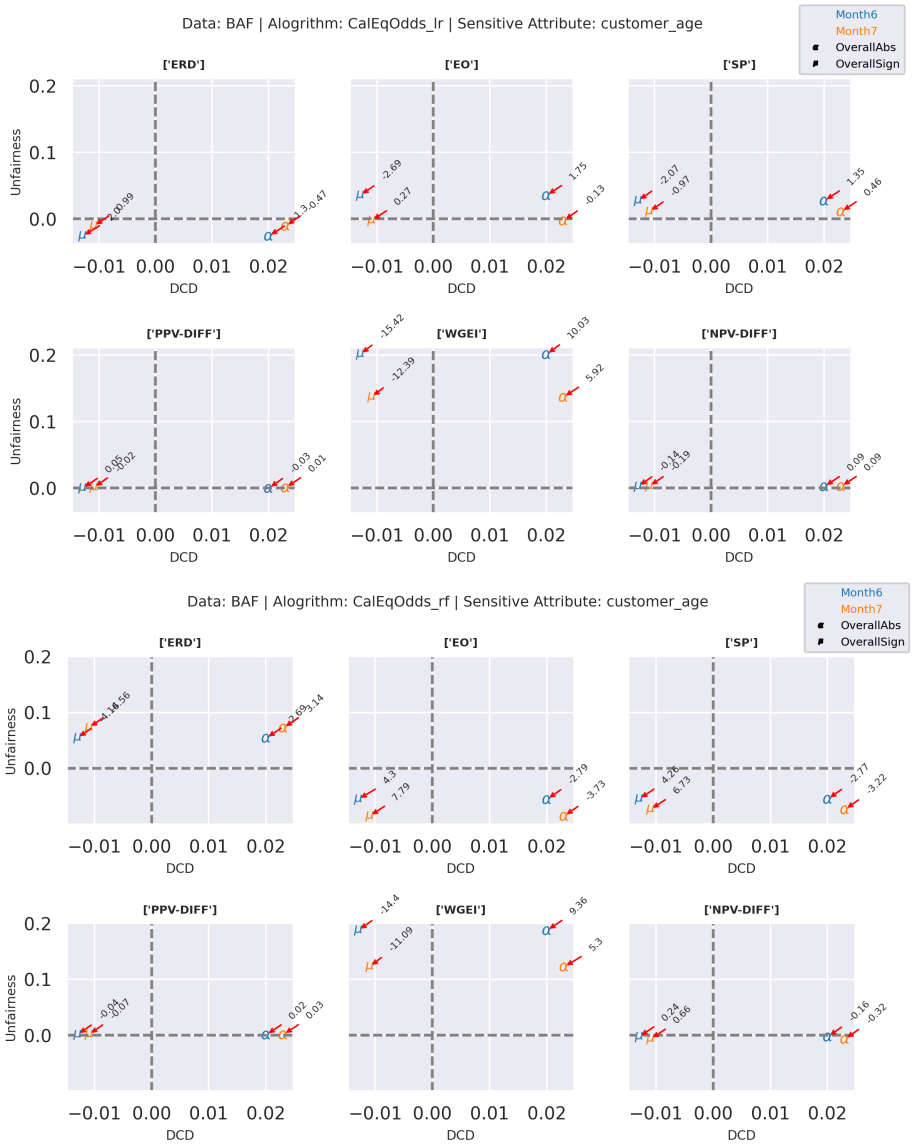

Figure 29: DCD vs unfairness for CalEqOdds for BAF dataset.

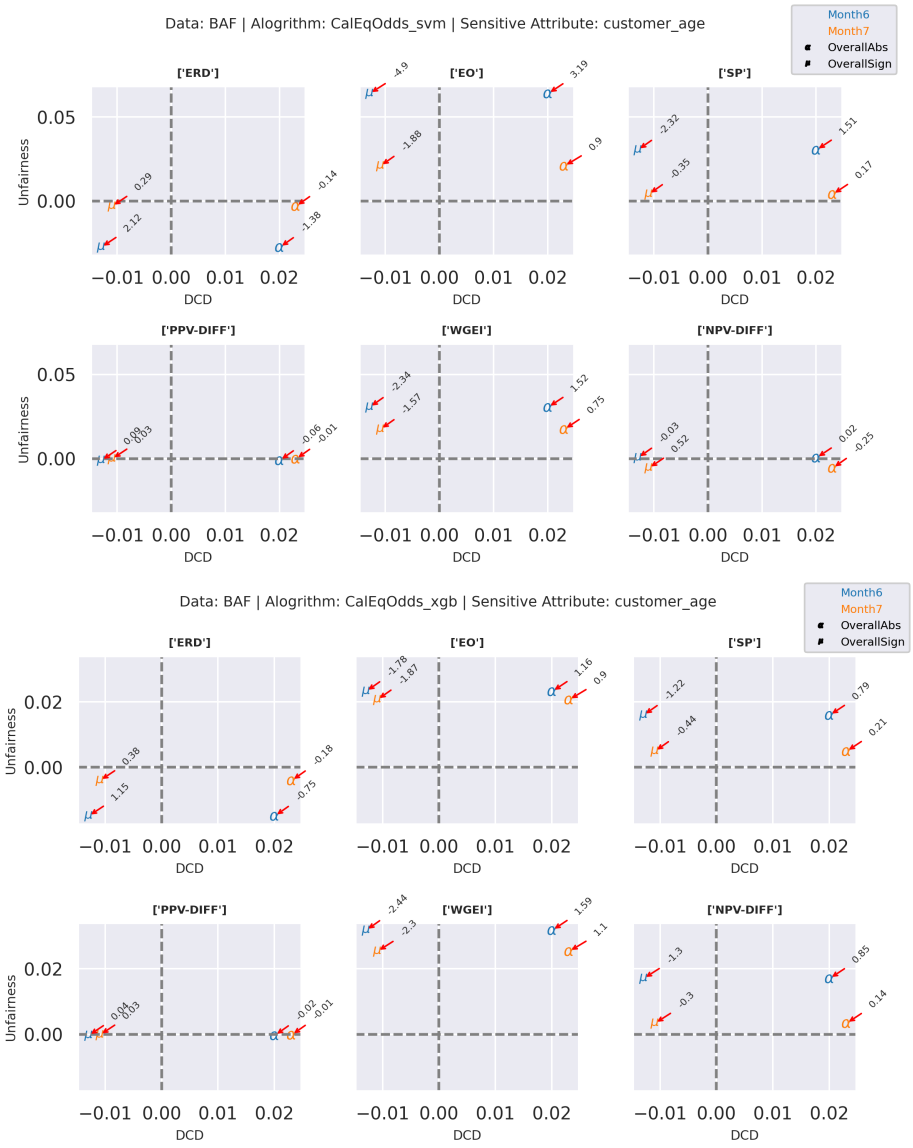

Figure 30: DCD vs unfairness for CalEqOdds for BAF dataset.

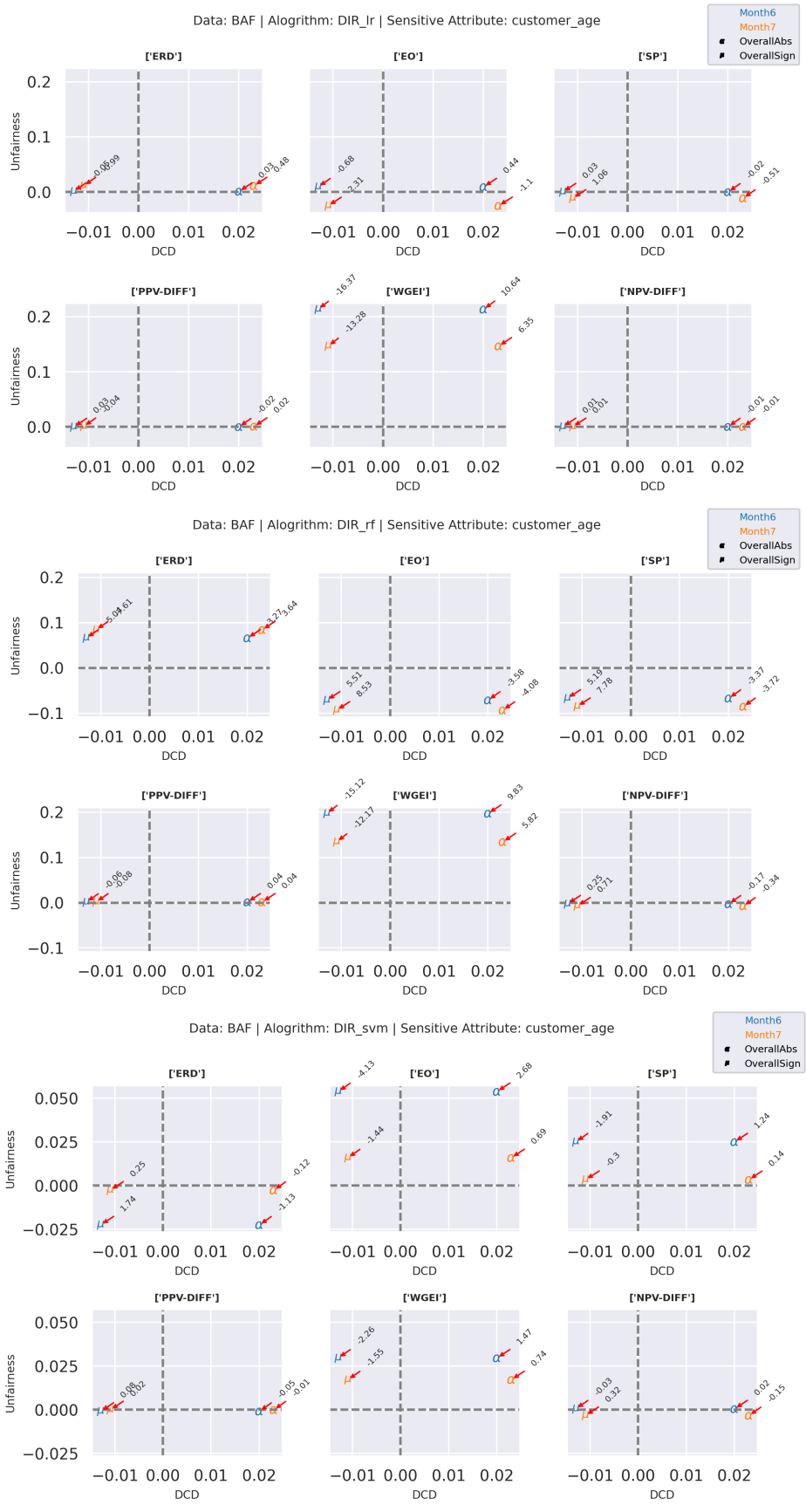

Figure 31: DCD vs unfairness for DIR for BAF dataset.

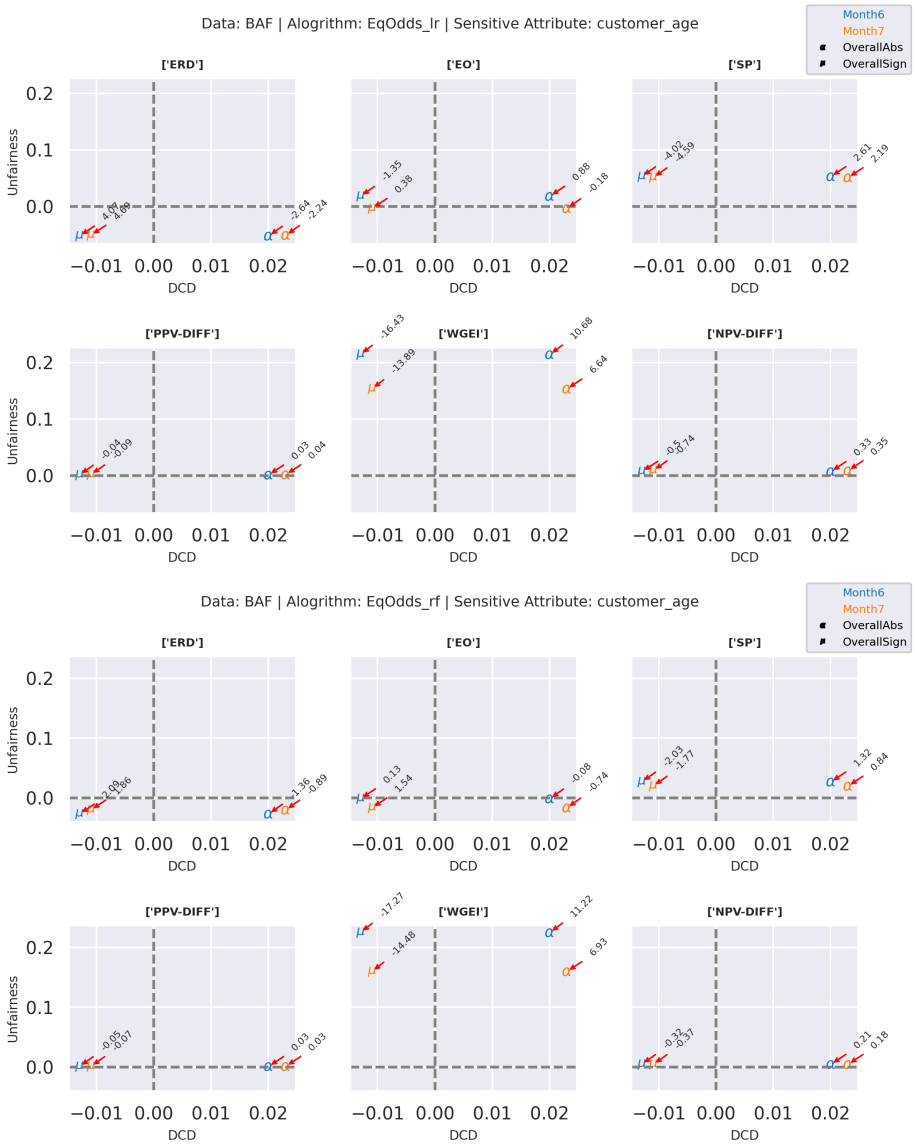

Figure 32: DCD vs unfairness for EqOdds for BAF dataset.

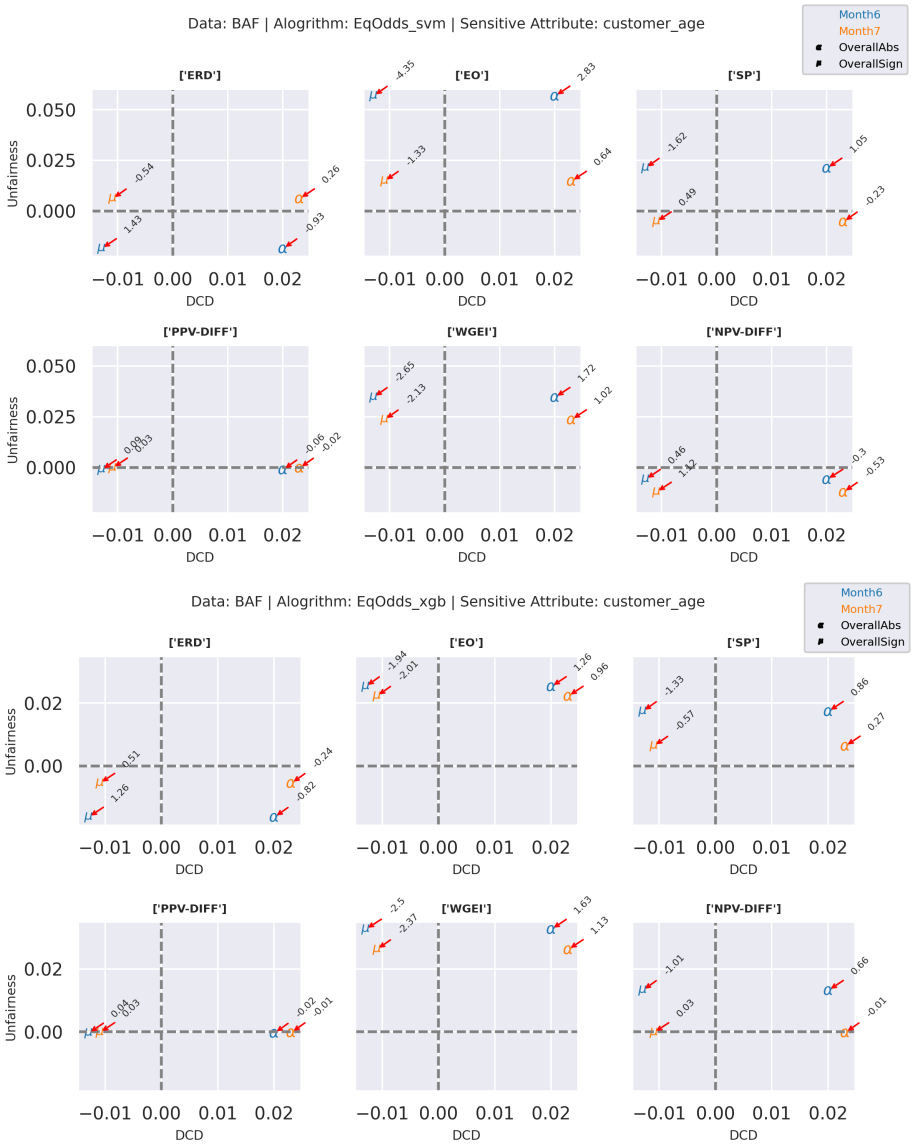

Figure 33: DCD vs unfairness for EqOdds for BAF dataset.

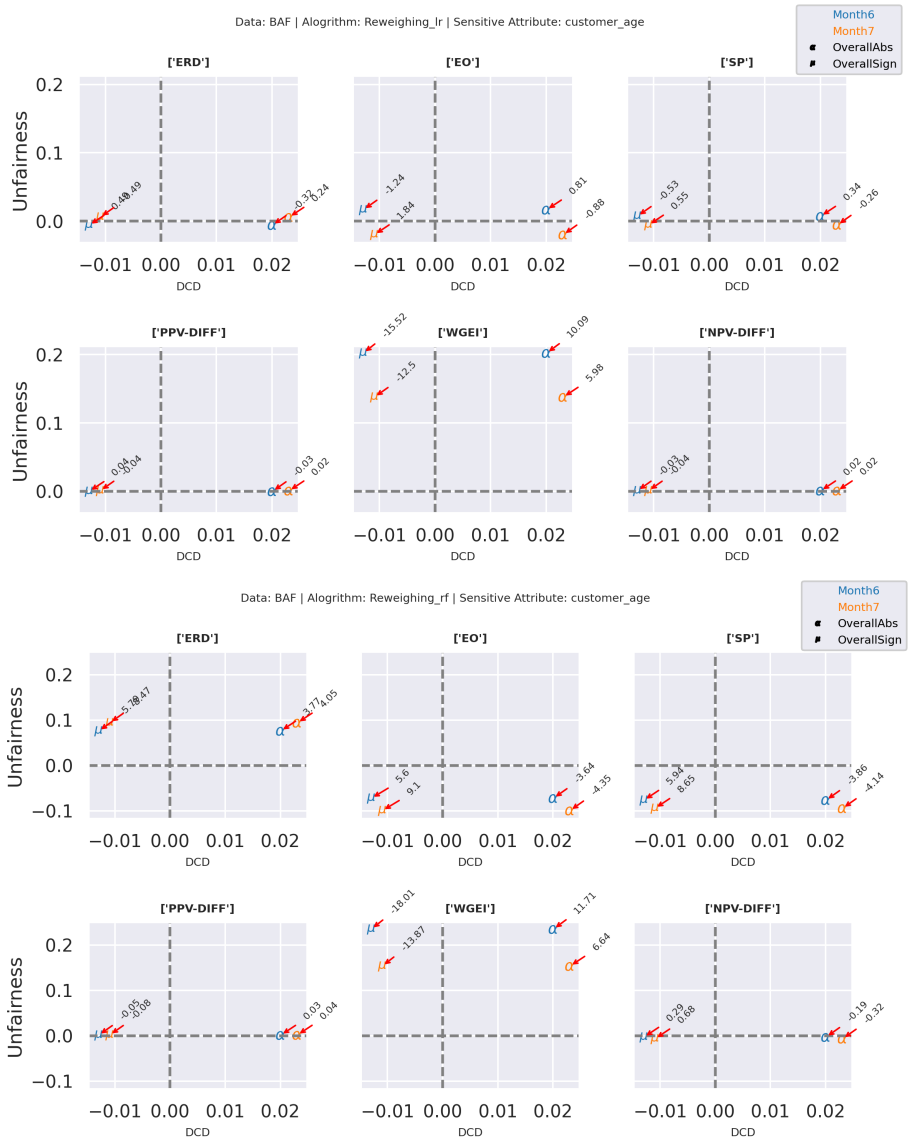

Figure 34: DCD vs unfairness for Reweighing for BAF dataset.

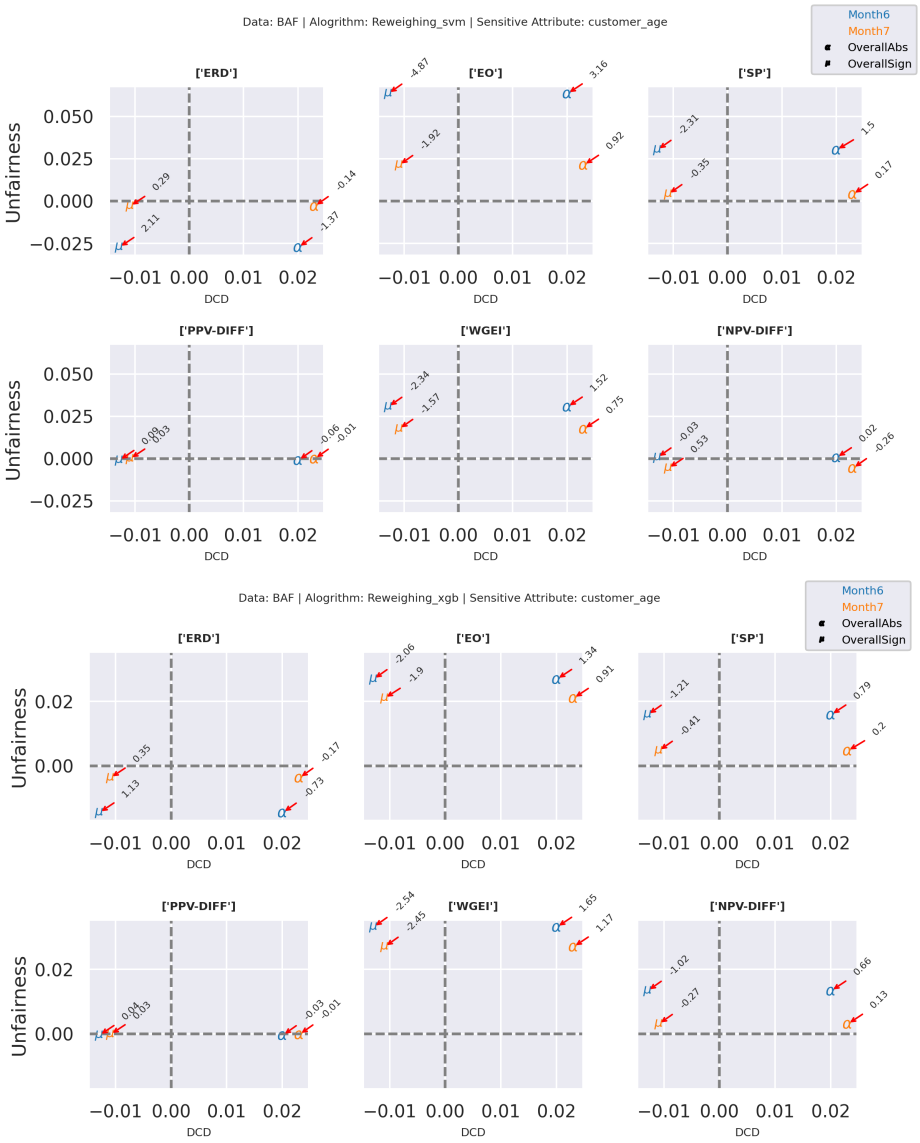

Figure 35: DCD vs unfairness for Reweighing for BAF dataset.

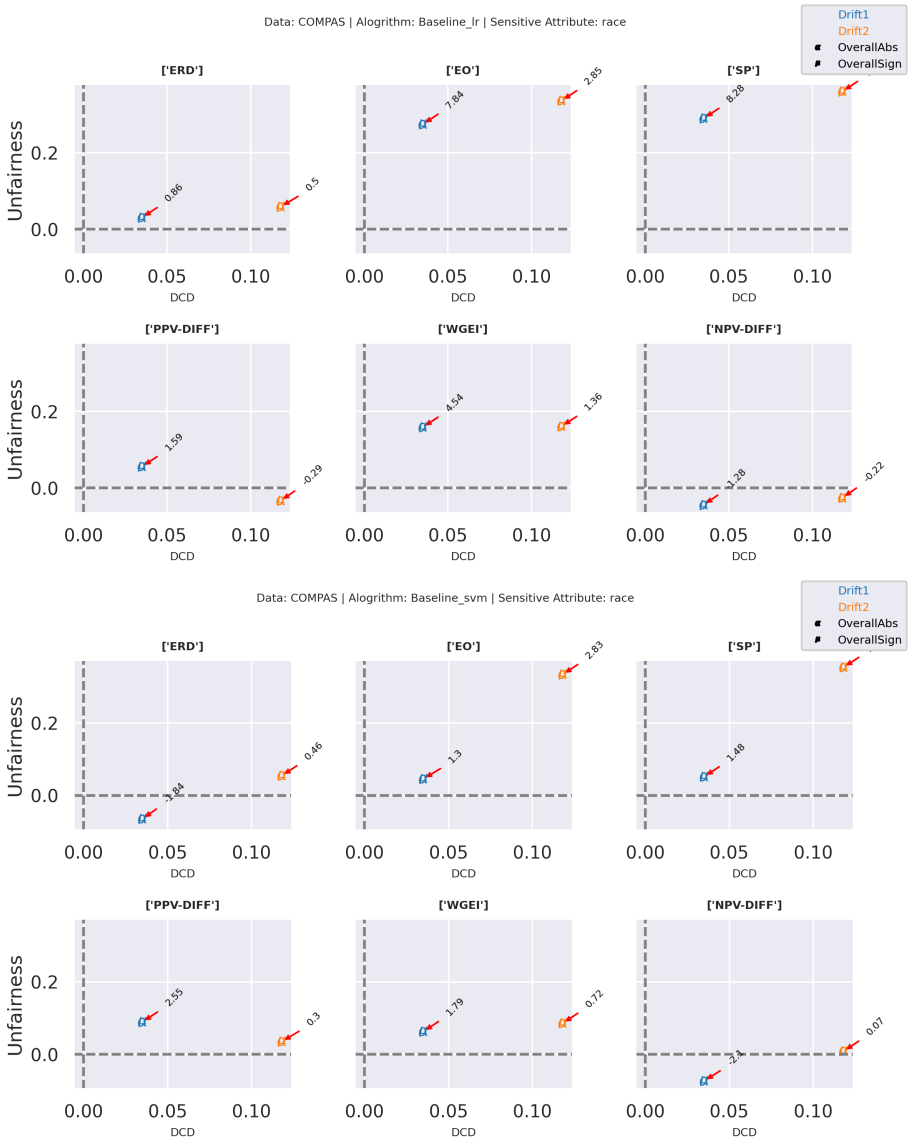

Figure 36: DCD vs unfairness for baseline models for COMPAS dataset.

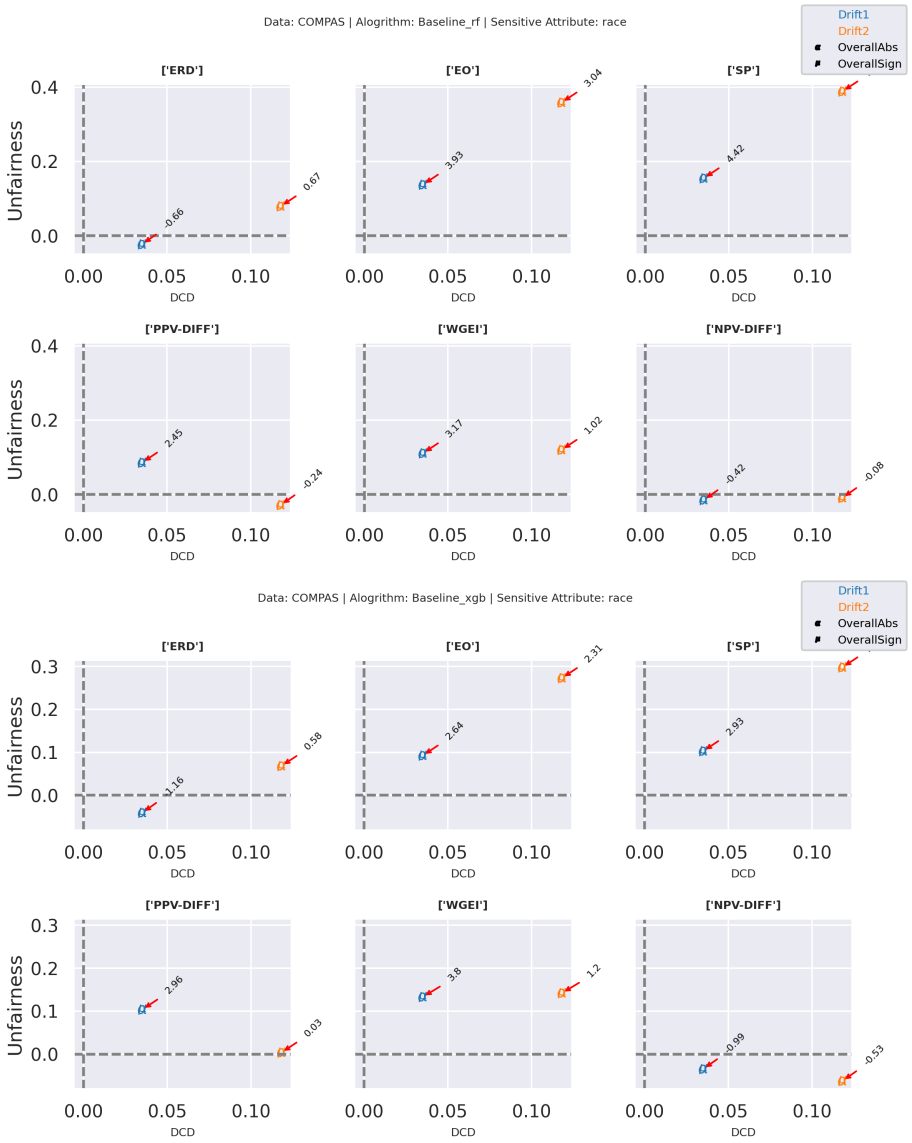

Figure 37: DCD vs unfairness for baseline models for COMPAS dataset.

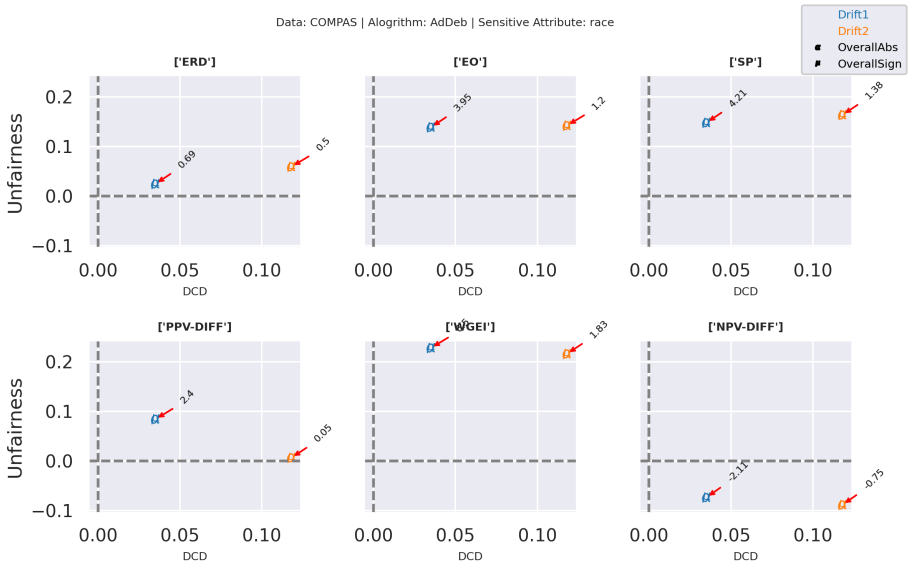

Figure 38: DCD vs unfairness for AdDeb for COMPAS dataset.

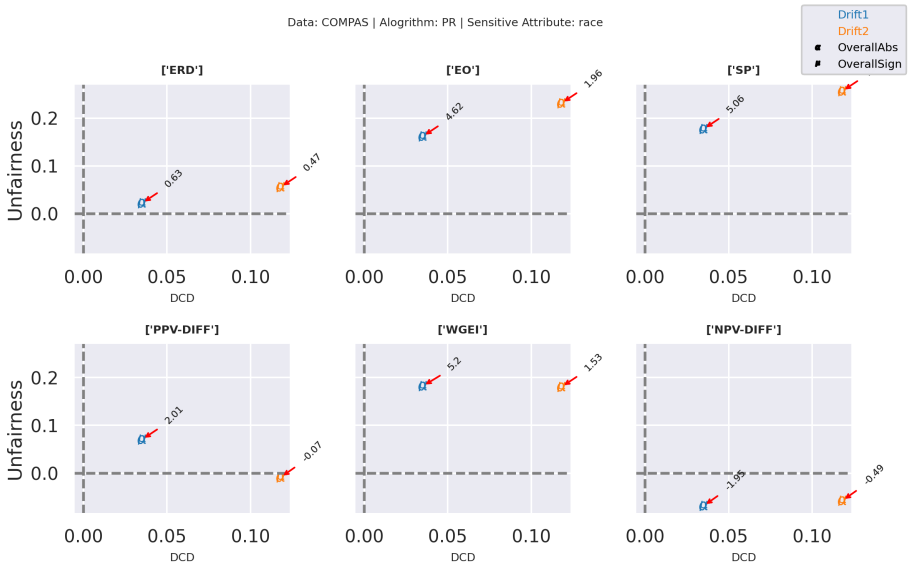

Figure 39: DCD vs unfairness for PR for COMPAS dataset.

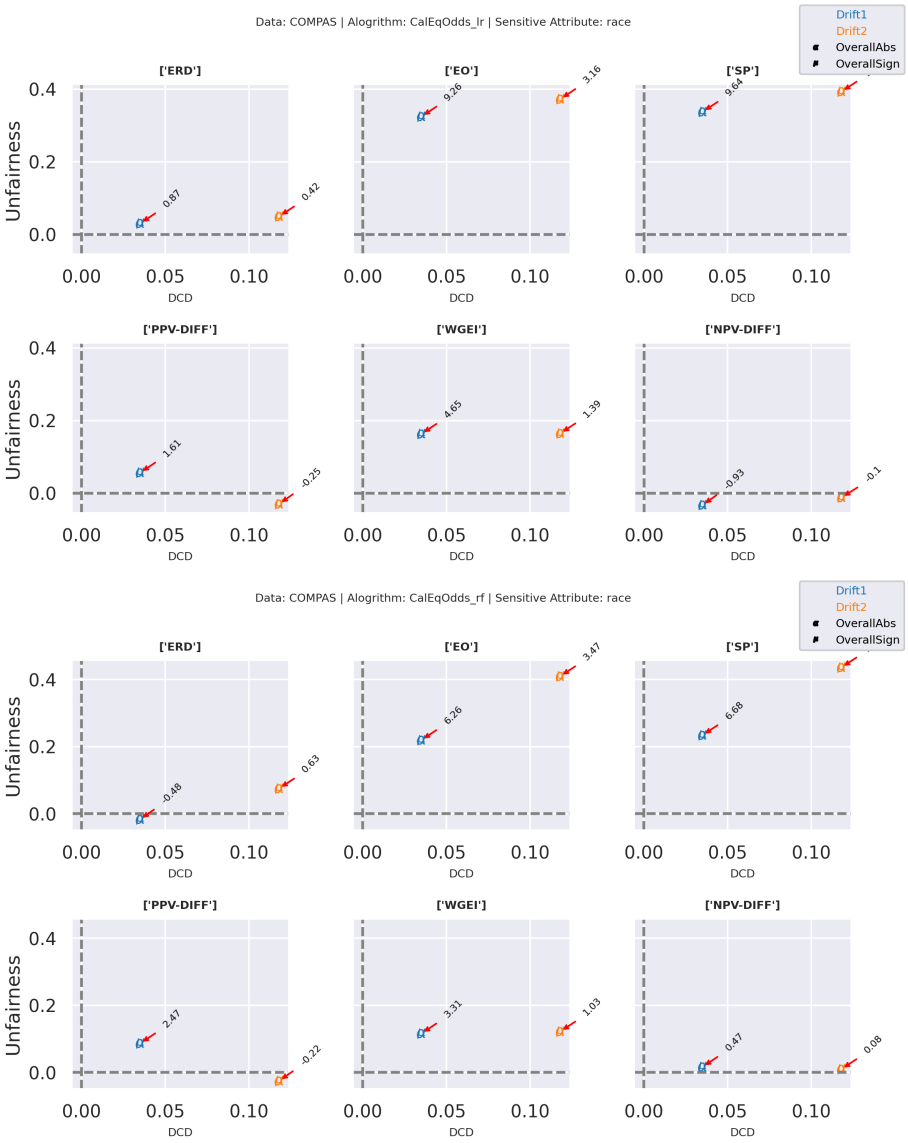

Figure 40: DCD vs unfairness for CalEqOdds for COMPAS dataset.

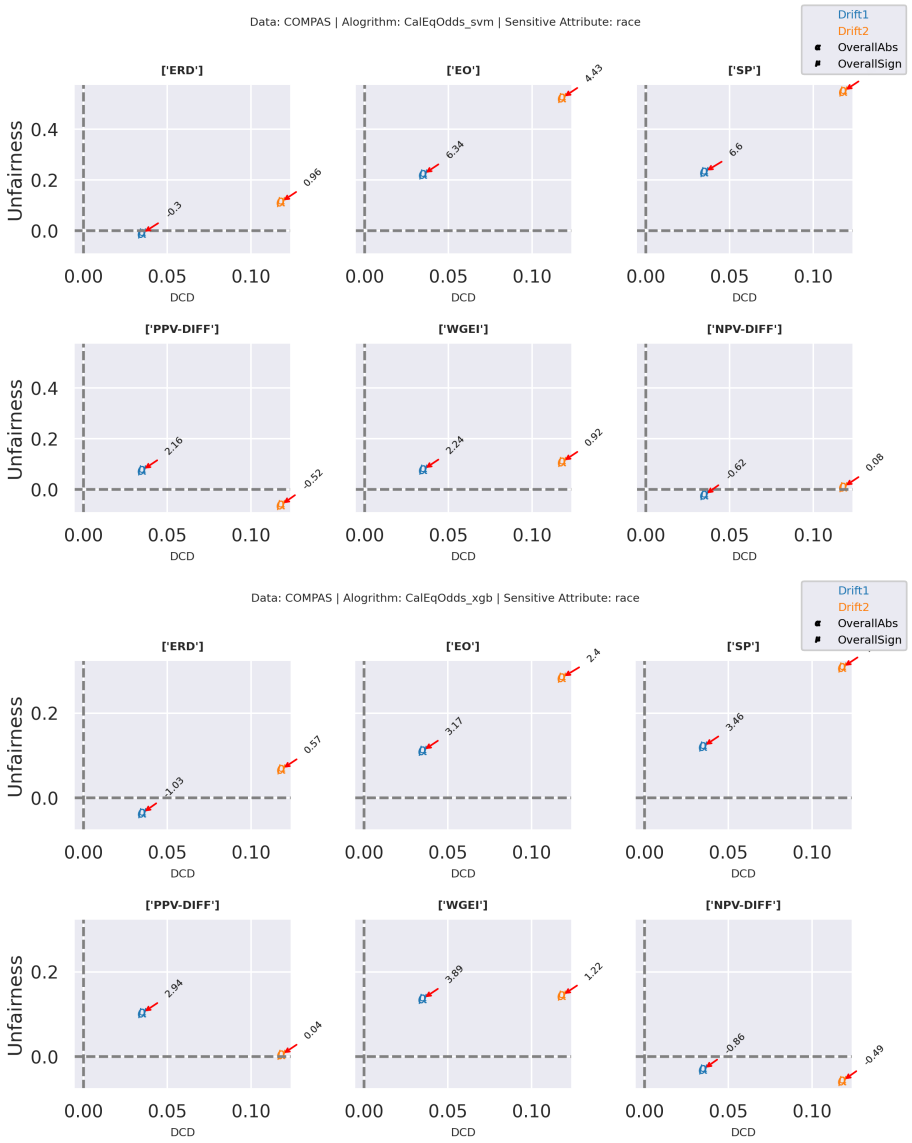

Figure 41: DCD vs unfairness for CalEqOdds for COMPAS dataset.

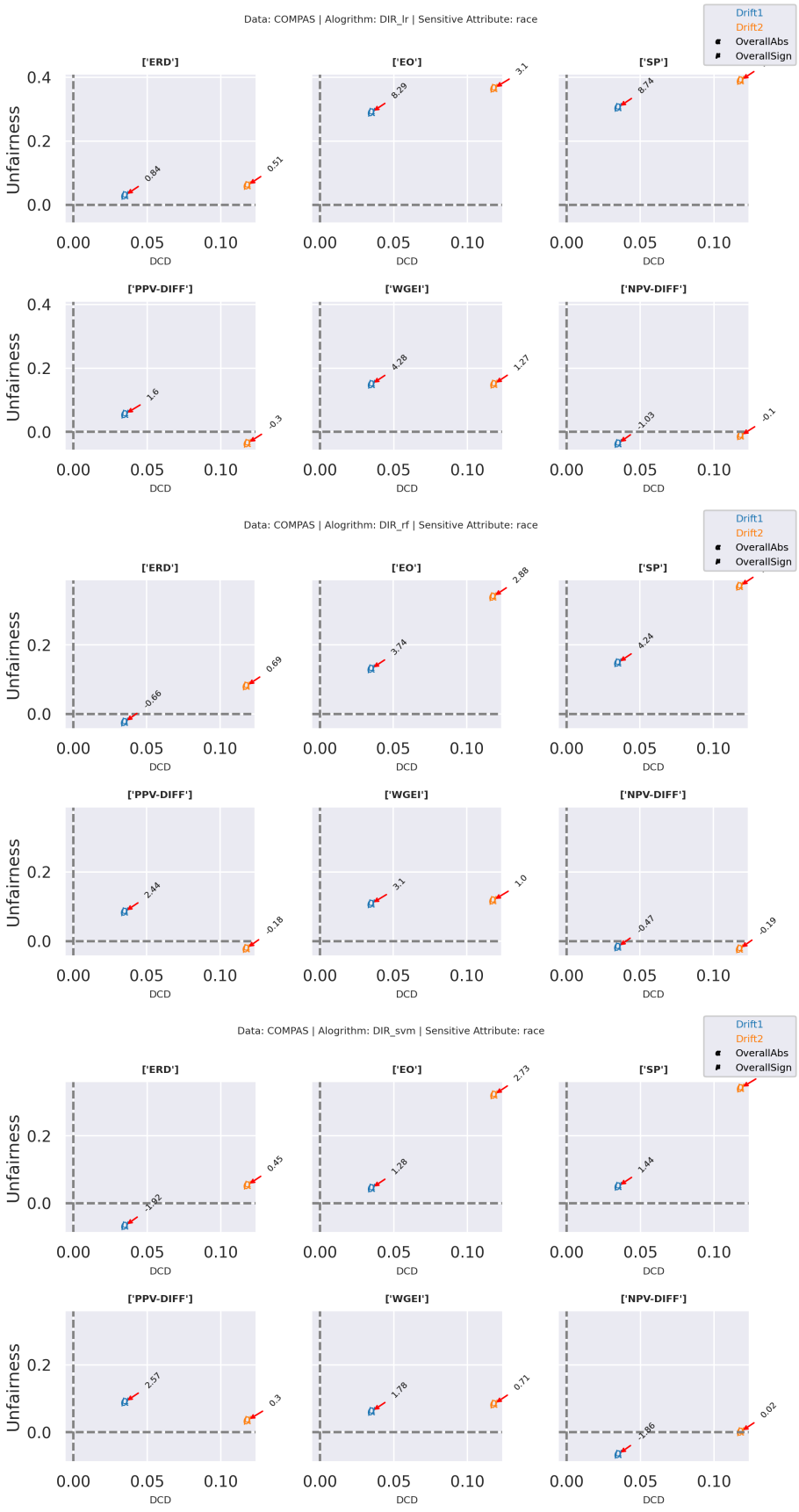

Figure 42: DCD vs unfairness for DIR for COMPAS dataset.

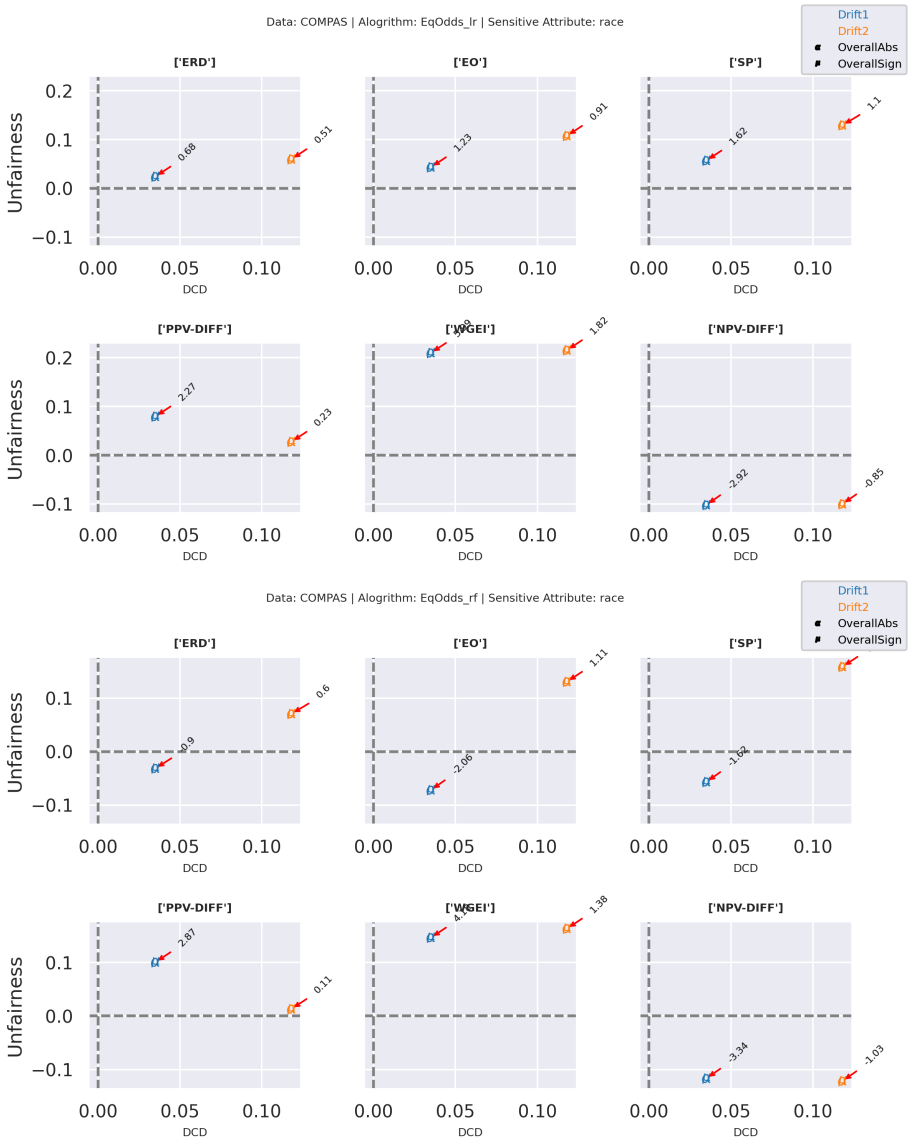

Figure 43: DCD vs unfairness for EqOdds for COMPAS dataset.

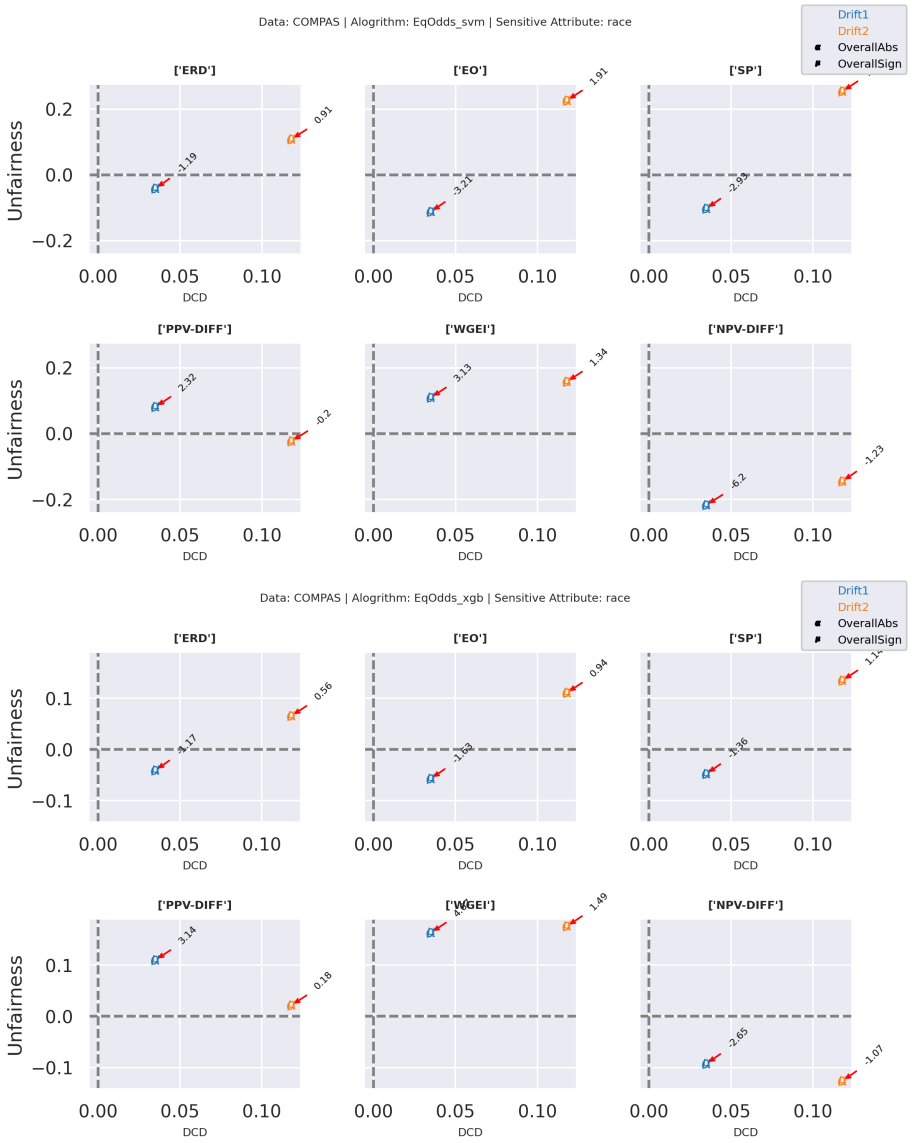

Figure 44: DCD vs unfairness for EqOdds for COMPAS dataset.

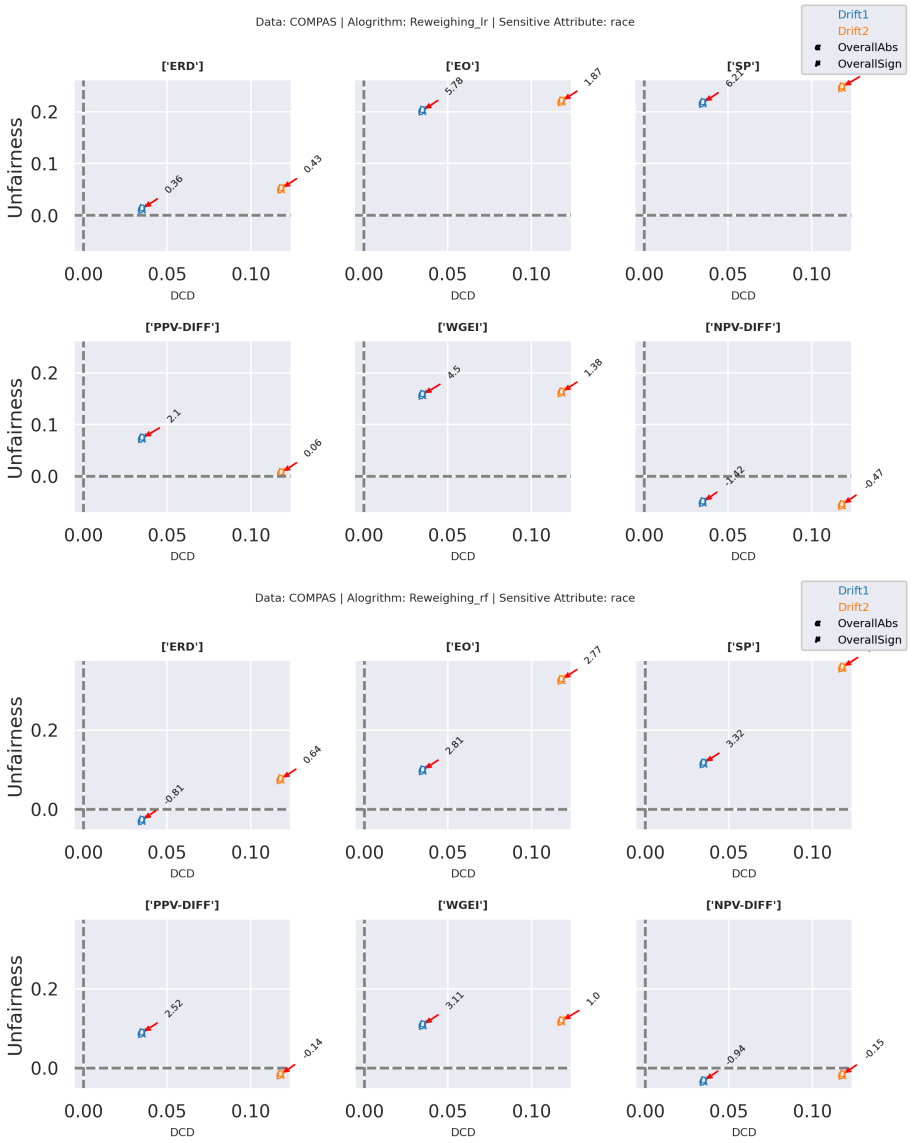

Figure 45: DCD vs unfairness for Reweighing for COMPAS dataset.

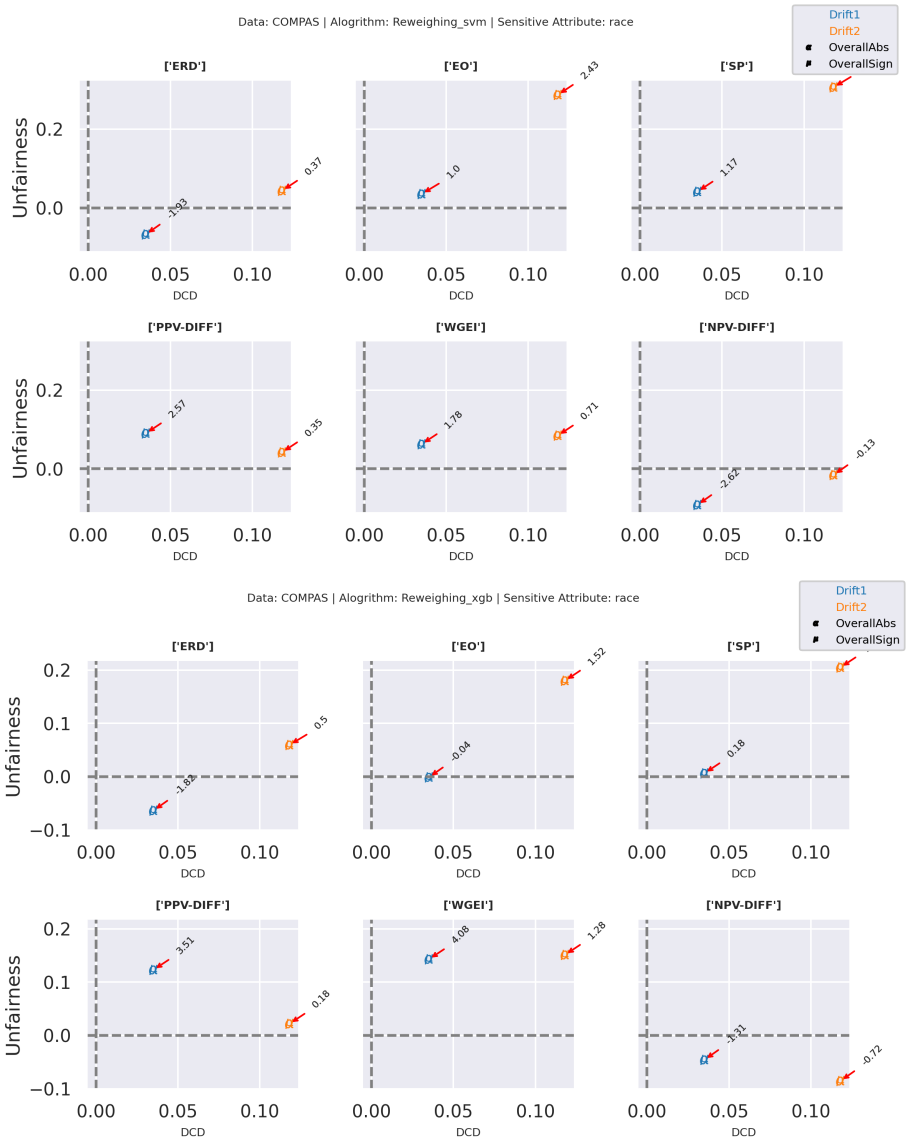

Figure 46: DCD vs unfairness for Reweighing for COMPAS dataset.