Effects of Matching on Evaluation of Accuracy, Fairness, and Fairness Impossibility in AI-ML Systems

Phillip Honenberger, Omolade Ola, William Mapp, Pilhwa Lee
Center for Equitable AI & Machine Learning Systems (CEAMLS), Morgan State University
philliphonenberger@gmail.com

Abstract

"Matching" procedures in statistics involve construction of datasets with similar covariates between compared groups. Matching has recently been proposed as a means of addressing fairness impossibility (i.e. inconsistency of fairness metrics) in AI and ML systems: Beigang argues on conceptual grounds that, when matched rather than unmatched datasets are analyzed, the tradeoff between the fairness metrics *equalized odds* (EO) and *positive predictive value* (PPV) will be reduced. Here we evaluate matching as a practical rather than merely conceptual approach to reducing fairness impossibility. As a case study we conduct pre-match and post-match analyses on the well-known COMPAS dataset from Broward Co., Florida, 2013-2014. We then reflect on what these results suggest about effects of matching on (a) *accuracy* estimates, (b) *fairness* estimates, and (c) *difference between* fairness estimates – that is, the extent to which matching reduces "fairness impossibility" in practice. We conclude that matching is a promising tool for improving evaluations on all three fronts, but faces problems due to potential biases introduced by matching procedures themselves, as well as limited power under conditions extremely common to ML evaluation contexts such as non-independent variables and relevance of hidden variables.

Introduction

"Matching" procedures in statistics involve construction of datasets with similar covariates between compared groups (Stuart 2010). Matching has recently been proposed as a means of addressing the well-known problem of "fairness impossibility" (i.e. inconsistency of fairness metrics, as demonstrated in Kleinberg et al. 2016 and Choudrechova 2017) in AI and ML systems. In particular, Beigang (2023) argues on conceptual grounds that, when matched rather than unmatched datasets are analyzed, the tradeoff between the fairness metrics *equalized odds* (EO) and *positive predictive value* (PPV) will be reduced. Since the EO-PPV tradeoff is perhaps the central example of a "fairness impossibility result" (Kleinberg et al. 2016; Chouldechova 2017), this proposal stands to fundamentally alter the understanding of fairness impossibility and its implications.

Here we attempt to evaluate matching as a practical rather than merely conceptual approach to reducing fairness impossibility. This is important and potentially fruitful insofar as it stands to estimate the extent to which, in real contexts of application, matching enables (a) more precise measurement of the *accuracy* of a classifier, (b) more precise measurements of the *fairness* of a classifier, and (c) can reduce the *difference between different fairness metrics* (in particular, equalized odds [EO] and positive predictive value [PPV]) – that is, the extent to which it solves the "fairness impossibility" problem, allowing EO or PPV to serve as equivalent and mutually substitutable measurements of fairness.

Methodologically, we approach the problem in three main steps. First, we use several simple matching procedures (for review of matching methods in general, see Stuart 2010) to build matched datasets for a well-known example in the fairness literature, the COMPAS dataset from Brouward Co., Florida, 2013-2014 (Larson et al. 2016). Second, we analyze the original unmatched data set and the matched data sets for their accuracy and fairness performance. Third, we reflect on and discuss the extent to which the matched results suggest improvements on items (a), (b), and (c), keeping in mind the practical challenges we faced in performing the procedures.

Accuracy, Fairness, and Fairness Impossibility

We want AI and ML systems to be both accurate and fair. To facilitate these goals, a variety of quantitative measures of accuracy and fairness have been proposed (for overviews, see Fraenkel 2020; Barocas et al. 2023; Kearns & Roth 2019). Some common measures or "metrics" for ac-

curacy include the four squares of the classical statistical confusion matrix – *true positive rate* (TPR), *false positive rate* (FPR), *true negative rate* (TNR), and *false negative rate* (FNR) – as well as *positive predictive value* (PPV) and *negative predictive value* (NPV). The first four of these metrics track accuracy in the sense of the percentage of actual values that are correctly or incorrectly predicted. The last two measures track accuracy in the sense of the percentage of predictions that turn out to be correct or incorrect (i.e. turn out to match actual values).

Common measures of fairness tend to depend on accuracy metrics insofar as they highlight differences in accuracy estimates when compared across groups. For instance, the COMPAS recidivism prediction algorithm was argued by Larson et al. (2016) to be unfair because it exhibited higher FPR for African-American individuals than for Caucasians, as well as higher FNR for Caucasians than African-Americans. These measures are known as *false positive parity* (FPP) and *false negative parity* (FNP) respectively. However, different fairness metrics can sometimes give different estimates of the extent, or even the presence or absence, of unfairness. As representatives of the COMPAS algorithm argued in defense, for instance, COMPAS satisfies PPV approximately equally well between African-American and Caucasian groups (Dietrich et al. 2016).

Indeed, mathematical arguments in the wake of the COMPAS controversy appeared to show that, under conditions of imperfect knowledge and different base rates between compared groups, no model can simultaneously achieve perfect FPP, FNP, and PPVP (Kleinberg et al. 2016; Chouldechova 2017). As these two conditions characterize the vast majority of circumstances in which AI-powered decision- and prediction-assistive systems like COMPAS are employed or would ever be employed, this result was a bombshell for researchers on algorithmic fairness. The result, known in subsequent discussion as "fairness impossibility," has inspired many efforts at amelioration (e.g. Saleiro et al. 2020; Bell et al. 2023; DeFrance 2023), but has no generally accepted solution.

In 2023, a novel interpretation of fairness impossibility was proposed by Beigang (2023), who argued on theoretical grounds that under the assumed ideal condition of perfectly statistically matched comparison groups, the difference between (on the one hand) the aggregate difference in FPR and FNR (also known as "equalized odds") and (on the other) the difference in PPV, should reduce to 0. In short, any system that is truly unfair would fail both metrics to the same extent, and any system that is truly fair would satisfy both equally.

Beigang's proposal was intended as a theoretical untangling of the fairness incompatibility problem. The novelty and promise of Beigang's proposal, however, inspired us to ask the question, "How well could statistical matching settle worries about fairness impossibility *in practice*?" and, more generally, "How could statistical matching contribute to evaluations of accuracy and fairness?"

Table 1. Common accuracy metrics

Abbv.	Full name	Meaning
TP	True positives	Cases where model's positive prediction is accurate
FP	False negatives	Cases where model's positive prediction is inaccurate
TN	True negatives	Cases where model's negative prediction is accurate
FN	False negatives	Cases where model's negative prediction is inaccurate
TPR	True positive rate	Ratio of true positives to actual positive cases
FPR	False positive rate	Ratio of false positives to actual negative cases
TNR	True negative rate	Ratio of true negatives to actual negative cases
FNR	False negative rate	Ratio of false negatives to actual positive cases
PPV	Positive predictive value	Ratio of true positives to total positive predictions of the model
NPV	Negative predictive value	Ratio of true negatives to total negative predictions of the model

Table 2. Common fairness metrics

Abbv.	Full name	Meaning
DP	Demographic parity	Difference in % of each group predicted positive
TPP	True positive parity	Difference in TPR between groups
FPP	False positive parity	Difference in FPR between groups
TNP	True negative parity	Difference in TNR between groups
FNP	False negative parity	Difference in FNR between groups
EO	Equalized odds	Difference in TPP and FPP between groups
PPVP	PPV parity	Difference in PPV between groups

Statistical Matching

Statistical matching procedures were first developed to solve a problem faced by observational studies: namely, that the regularized conditions of controlled studies (for instance, randomized controlled trials) are usually not available in observational studies, and the reliability of results drawn from observational data are thereby reduced (Stuart 2010). Statistical matching procedures aim to solve this problem by artificially transforming observational datasets to mimic those of controlled experiments. Typically they involve splitting observational data into "treatment" and "control" groups and then employing procedures to make co-variates (that is, non-treatment and non-outcome variables) more similar, in aggregate, between the groups, thereby mimicking bias-mitigated formation of treatment and control groups in randomized controlled trials. Statistical matching is similar to techniques for correcting class imbalance, such as undersampling, oversampling, and SMOTE (Chawla et al. 2002; Wongvorachan et al. 2023) in its ambition to make observational data more like experimental data, but is directed at a different aspect of the

problem: imbalance in covariates between classes, rather than imbalance in number of instances in each class.

Statistical matching usually comprises three main steps (Stuart 2010): (1) selection of a matching procedure and application of this procedure to the original (unmatched) dataset; (2) analysis of the resulting matched dataset for "balance" (that is, the extent to which total distribution of co-variates is the same or different between the groups, as well as between the original unmatched dataset and the new matched one, with greater similarity, i.e. "balance," preferred, as this indicates greater reliability in post-match analyses); and (3) analysis of the resulting matched dataset for the variables of interest (i.e. link between treatment and outcome variables).

Common matching procedures include exact matching, k:1 nearest neighbor matching, and matching with use of propensity scores (Stuart 2010). An important decision that must be made is whether matching will be done with or without *replacement* – that is, whether duplicate records will be generated in the treatment group or control group as a means of creating the desired parallel between groups. In cases where replacement is used, attention should be given to how replacement may affect subsequent analyses (for instance, as a source of bias in results).

Regarding balance: a similar distribution of covariates between groups, as well as between matched and unmatched datasets, is desirable, but there is no standard threshold for acceptable similarity, nor any standard posthoc methods to ensure or improve balance where it is lacking (Austin 2009, Stuart 2010). Stuart (2010) recommends checking for balance and then, if balance levels are found insufficient, revising the match procedure to try to create a more balanced dataset, then running the match operation and balance checks again, and so on.

In general, we believe it is recommendable to bear in mind potential tradeoffs in use of matching procedures. Matching is a way to make observational data more like experimental data. It promises to improve the power and precision of observational studies. At the same time, it always involves some transformation of observational datasets from their original form and thereby risks introducing new biases into results. Matching procedures are not a substitute for careful and self-critical exploration of the implications of observational datasets, but rather one tool for such careful and self-critical exploration.

Methods

Our aim was to explore how statistical matching could contribute to the evaluation of real AI-ML systems for accuracy, fairness, and inconsistency of fairness metrics (i.e. "fairness impossibility"). As a case study for exploring these questions, we selected the well-known *COMPAS dataset* (Larson et al. 2016) composed of COMPAS recidivism predictions for arrests in Broward County, Florida, 2013-2014, and associated data about actual recidivism rates within a 2-year period. In addition to the original unmatched dataset, we constructed nine alternative datasets, some matched and some unmatched (construction procedures detailed in Table 3). We then tested each dataset for co-variate balance (results in Table 4) and analyzed each for accuracy, fairness, and fairness impossibility (results in Table 6 and Figures 1-2). All procedures were performed in MySQL and Excel.

Results and Discussion

Several features of the results are noteworthy. First, matching procedures tended to reduce estimates of unfairness, as evidenced by the following inequalities among datasets in Figure 1: d<(a∨c), e<(a∨c), g<(a∨f).

At the same time, signatures of unfairness are preserved in roughly the same pattern as the original (unmatched) datasets. The greater FPR for African-Americans, and greater TNR for Caucasians, noted in the original ProPublica analysis as well as our (a), appeared in *all* datasets, matched or unmatched. In other words, the original signature of unfairness of these kinds survived the matching. The combination of these trends suggests that while analysis of the original, unmatched dataset may have resulted in unduly high estimates of FPR and FNR unfairness, our confidence that the COMPAS system *is* unfair by the standards of FPP, TNP, and EO should be strengthened. In this case, post-match analyses have contributed robustness to the conclusion that the COMPAS system is unfair by the standards of FPP, TNP, and EO.

Further, matching tended to reduce the difference between fairness metrics, particularly EO and PPV, suggesting that matching can reduce fairness impossibility in practice. This is evidenced by lesser EO-PPVP difference in matched datasets by comparison with unmatched ones (Table 6). On the other hand, that h>a both for raw fairness metrics and EO-PPVP difference may signify greater unfairness, in COMPAS's performance, regarding drug-related charges.

A lingering worry is whether and how matching procedures may introduce new biases. One indicator of bias in matching procedures is *imbalance* in either of two senses: dissimilar covariate distribution between "treatment" and "control" within matched subgroups; and dissimilar covariate distribution between the original (unmatched) and constructed (matched) groups.

Regarding the first kind of balance: only to the extent that matched groups within a matched dataset (in our research design, African-American and Caucasian) are similar in covariate distribution, can we think of the matched

Table 3. Construction procedures for datasets

Name of dataset	No. of Records	Construction procedure
Total (unmatched)	8732	Download from ProPublica Github, "compass-violent-parsed-filt." Remove duplicate records. Remove records where "is_recid" (i.e. record of recidivism) is not 0 or 1. Remove records where "race" is not "Caucasian" or "African-American."
Random (unmatched)	1670	Randomly select 835 records from Total (unmatched) where race='Caucasian' and 835 records from Total (unmatched) where race ="African-American"
"battery" (unmatched)	1366	Select from Total (unmatched) all records where "c_charge_description" equals "battery."
"battery" (matched by addition)	1680	Select all records from "battery" (unmatched). Calculate differences in between-group representation of 18 subgroups defined by all possible combinations of sex, age category, and prior category (0, 1, or >1). Equalize between-group representation in each subgroup by adding randomly selected records with the subgroup's defining characteristics from the originally underrepresented group.
"battery" (matched by subtraction)	1052	Select all records from "battery" (unmatched). Calculate differences in between-group representation of 18 subgroups defined by all possible combinations of sex, age category, and prior category (0, 1, or >1). Equalize between-group representation by deleting randomly selected records with the subgroup's defining characteristics from the originally overrepresented group.
theft-related (unmatched)	1105	Copy from Total (unmatched) all cases where "c_charge_description" equals a theft-related charge, except cases with <10 records for that charge.
theft-related (matched by addition)	1458	Copy all records from Theft-related (unmatched). Calculate differences in between-group representation of 18 subgroups defined by all possible combinations of sex, age category, and prior category (0, 1, or >1). Equalize between-group representation in each subgroup by adding randomly selected records with the subgroup's defining characteristics from the originally underrepresented group.
drug-related (unmatched)	1842	Copy from Total (unmatched) all cases where "c_charge_description" equals a drug-related charge.
Random (matched)	2000	Randomly select pairs of records from Total (unmatched) where race='Caucasian' in one of each pair, and race="African-American" in the other of each pair; and sex, age, number of prior convictions, and charge description are identical between the paired records. Selection occurs with replacement; some records are duplicated up to 7 times.
Random (matched in categories)	2000	Randomly select pairs of records from Total (unmatched) where race='Caucasian' in one of each pair, and race="African-American" in the other of each pair; and sex, age category (<25, 25-45, or >45), priors category (0, 1, >1), and charge description are identical between the paired records. Selection occurs with replacement; some records are duplicated up to 10 times.

Table 4. Covariate ratios for each dataset

			Sex		Age			Priors		
		Total	**Male**	**Female**	**<25**	**25-45**	**>45**	**0**	**1**	**>1**
Total (unmat.)	*Total*	*100.0%*	79.0%	21.0%	21.0%	57.0%	22.0%	29.4%	19.4%	51.2%
	Cauc.	*41.0%*	75.9%	24.1%	15.6%	54.2%	30.3%	35.9%	20.6%	43.5%
	AA	*59.0%*	81.1%	18.9%	24.8%	58.9%	16.3%	24.9%	18.6%	56.6%
Random (unmat.)	*Total*	*100.0%*	78.9%	21.1%	20.2%	56.0%	23.8%	30.8%	19.0%	50.2%
	Cauc.	*50.0%*	74.5%	25.5%	15.6%	53.7%	30.8%	37.8%	20.8%	41.3%
	AA	*50.0%*	83.2%	16.8%	24.8%	58.4%	16.8%	23.7%	17.1%	59.2%
Battery (unmat.)	*Total*	100.0%	72.0%	28.0%	19.5%	58.2%	22.3%	48.0%	19.4%	32.6%
	Cauc.	47.7%	70.7%	29.3%	16.0%	52.1%	31.9%	52.9%	19.6%	27.5%
	AA	52.3%	73.1%	26.9%	22.7%	63.7%	13.6%	43.6%	19.2%	37.3%
Battery (mat. by add.)	*Total*	*100.0%*	72.6%	27.4%	19.3%	55.7%	25.0%	46.7%	20.5%	32.9%
	Cauc.	*50.0%*	72.6%	27.4%	19.3%	55.7%	25.0%	46.7%	20.5%	32.9%
	AA	*50.0%*	72.6%	27.4%	19.3%	55.7%	25.0%	46.7%	20.5%	32.9%
Battery (mat. by sub.)	*Total*	*100.0%*	70.9%	29.1%	19.8%	62.2%	18.1%	50.2%	17.7%	32.1%
	Cauc.	*50.0%*	70.9%	29.1%	19.8%	62.2%	18.1%	50.2%	17.7%	32.1%
	AA	*50.0%*	70.9%	29.1%	19.8%	62.2%	18.1%	50.2%	17.7%	32.1%
Theft-rel. (unmat.)	*Total*	*100.0%*	77.4%	22.6%	37.4%	44.9%	17.7%	32.9%	20.1%	47.1%
	Cauc.	*35.7%*	78.0%	22.0%	24.6%	52.4%	23.0%	34.4%	21.3%	44.3%
	AA	*64.3%*	77.0%	23.0%	44.5%	40.7%	14.8%	32.0%	19.4%	48.6%
Theft-rel. (mat. by add.)	*Total*	*100.0%*	77.7%	22.3%	43.1%	40.2%	16.8%	33.1%	19.9%	47.0%
	Cauc.	*50.0%*	77.7%	22.3%	43.1%	40.2%	16.8%	33.1%	19.9%	47.0%
	AA	*50.0%*	77.7%	22.3%	43.1%	40.2%	16.8%	33.1%	19.9%	47.0%
Drug-rel. (unmat.)	*Total*	*100.0%*	80.9%	19.1%	18.7%	56.9%	24.4%	31.5%	16.6%	51.8%
	Cauc.	*49.2%*	74.4%	25.6%	16.2%	56.2%	27.6%	40.8%	18.6%	40.6%
	AA	*50.8%*	87.3%	12.7%	21.1%	57.6%	21.3%	22.6%	14.7%	62.8%
Random (mat.)	*Total*	100.0%	84.7%	15.3%	38.1%	54.1%	7.8%	51.4%	29.0%	19.6%
	Cauc.	50.0%	84.7%	15.3%	38.1%	54.1%	7.8%	51.4%	29.0%	19.6%
	AA	50.0%	84.7%	15.3%	38.1%	54.1%	7.8%	51.4%	29.0%	19.6%
Random (mat. in cat.)	*Total*	100.0%	93.4%	6.6%	6.6%	79.8%	13.6%	15.8%	8.9%	75.3%
	Cauc.	50.0%	93.4%	6.6%	6.6%	79.8%	13.6%	15.8%	8.9%	75.3%
	AA	50.0%	93.4%	6.6%	6.6%	79.8%	13.6%	15.8%	8.9%	75.3%

Table 5. Formulae for metrics

P_r = Records w/ risk score ≥ 5	
P_n = Records w/ risk score <5	
A_r = Records that recidivized	
A_n = Records that did not recidivize	
$TP = P_r \cap A_r$	$TN = P_n \cap A_n$
$FP = P_r \cap A_n$	$FN = P_n \cap A_r$
$TPR = \frac{TP}{A_r}$	$TNR = \frac{TN}{A_n}$
$FPR = \frac{FP}{A_n}$	$FNR = \frac{FN}{A_r}$
$PPV = \frac{TP}{P_r}$	$NPV = \frac{TN}{P_n}$
R_a = Records w/ race= "African-American"	
R_c = Records w/ race="Caucasian"	
$FPR(Af.-Am.) = \frac{FP \cap R_a}{A_n \cap R_a}$	
$FNR(Cauc.) = \frac{FN \cap R_c}{A_r \cap R_c}$	
$PPV(w\text{-}avg) = \left(\frac{TP \cap R_a}{P_r \cap R_a}\right) * \frac{R_a}{R_a + R_c} + \left(\frac{TP \cap R_c}{P_r \cap R_c}\right) * \frac{R_c}{R_a + R_c}$	
$DP = \left\lvert \frac{(P_r \cap R_a)}{(R_a)} - \frac{(P_r \cap R_c)}{(R_c)} \right\rvert$	
$TPP = TPR(R_c) - TPR(R_a)$	
$FPP = FPR(R_a) - FPR(R_c)$	
$TNP = TNR(R_c) - TNR(R_a)$	
$FNP = FNP(R_c) - FNP(R_a)$	
$EO = TPP + FPP$	
$PPVP = PPV(R_c) - PPV(R_a)$	
EO-PPVP = EO – PPVP	

Table 6. Accuracy and fairness results across datasets

	Accuracy			Fairness					Fairness Impossibility
	FPR(Af-Am.)	**FNR(Cauc.)**	**PPV(w-avg)**	**DP**	**FPP**	**FNP**	**EO**	**PPVP**	**EO-PPVP**
(a) total (unmat.)	0.4881	0.4903	0.4664	0.2409	0.2237	0.1996	0.4234	0.0506	0.3727
(b) random (unmat.)	0.4795	0.5169	0.4673	0.2719	0.2391	0.2592	0.4983	0.0509	0.4474
(c) battery (unmat.)	0.3132	0.6233	0.4593	0.2023	0.1610	0.1970	0.3580	0.0816	0.2764
(d) battery (mat. by add.)	0.2916	0.6048	0.4460	0.1488	0.1106	0.1350	0.2873	0.0796	0.2077
(e) battery (mat. by sub.)	0.2893	0.5814	0.4546	0.1369	0.1105	0.1343	0.2448	0.0452	0.1997
(f) theft-rel. (unmat.)	0.5855	0.3797	0.4725	0.1508	0.1635	0.1324	0.2959	0.0349	0.2610
(g) theft-rel. (mat. by add.)	0.5861	0.3322	0.4694	0.0831	0.0931	0.0779	0.1710	0.0396	0.1314
(h) drug-rel. (unmat.)	0.5697	0.4805	0.4735	0.2995	0.2779	0.2357	0.5135	0.1217	0.3919
(i) random (mat.)	0.3451	0.4684	0.4006	0.0910	0.0843	0.0554	0.1397	0.0258	0.1139
(j) random (mat. in cat.)	0.4827	0.3968	0.4730	0.1620	0.1250	0.1704	0.2954	0.0724	0.2230

Figure 1. Fairness results for each dataset

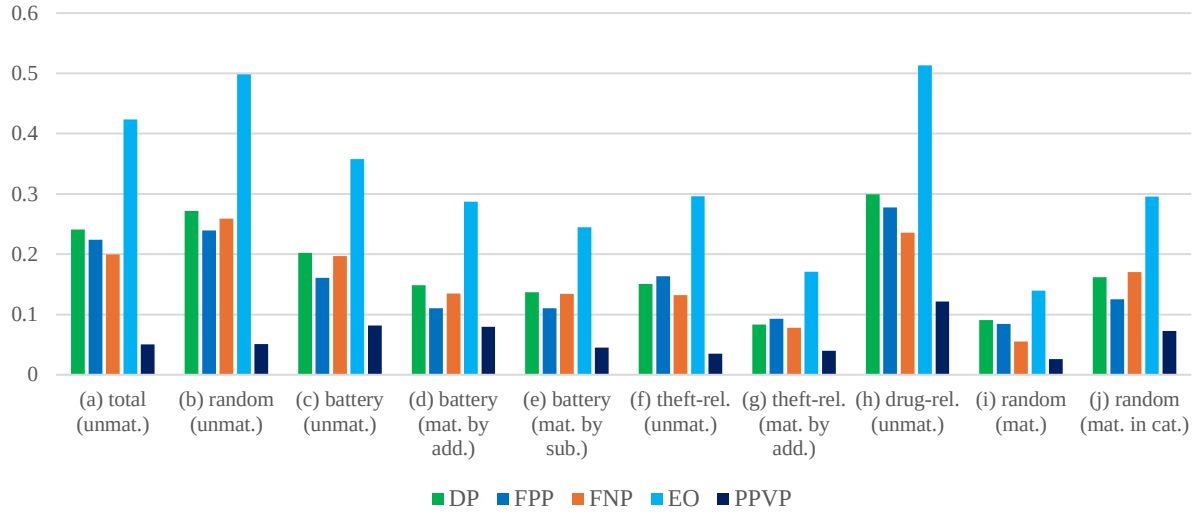


Figure 2. Accuracy results for each dataset

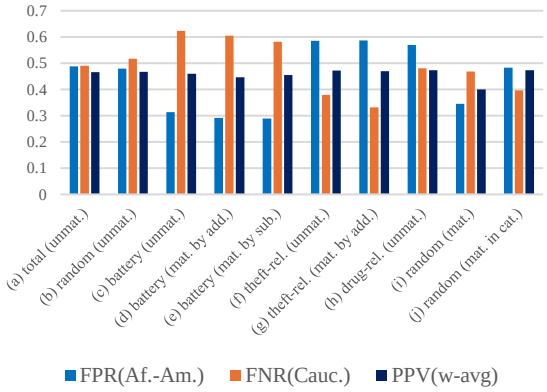


Figure 3. Sex ratio between datasets

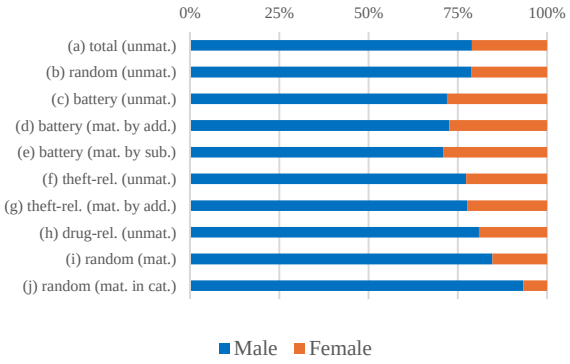


Figure 4. Age ratio between datasets

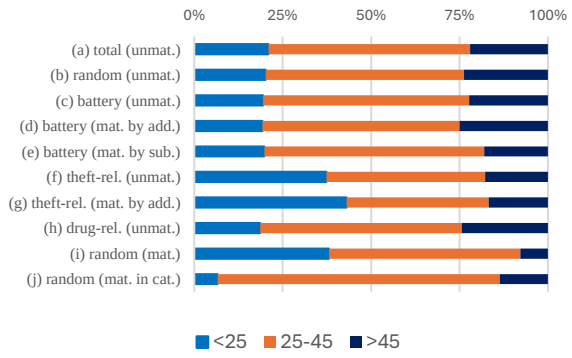
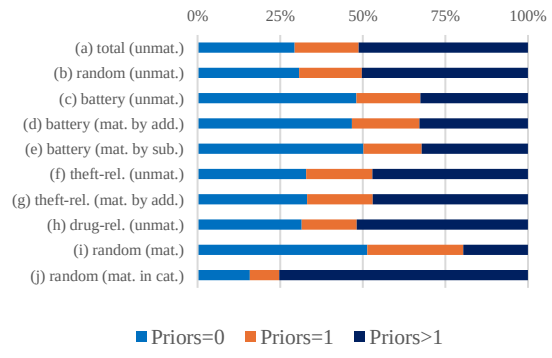


Figure 5. Priors count ratio between datasets



dataset as like a randomized controlled trial where "treatment" and "control" are selected to be similar across relevant covariates. All our matched datasets are "balanced" in this sense (Table 4). (Incidentally, this is one benefit of "exact" matching procedures in comparison with others, such as propensity scores).

Regarding the second sort of balance: When the distribution of covariates between matched datasets and the original, unmatched dataset are very different, that difference suggests that the construction procedure has introduced biases that affect post-match analyses (Stuart 2010): for instance, by leaving out individuals from either group for whom no exact co-variate match could be found; or by duplicating records from one group to achieve balance with the other. In general, construction procedure can introduce bias by changing the extent to which individuals with different covariate signatures are "counted" towards estimates of overall accuracy and fairness.

Figures 3-5 summarize tests for balance in the second sense (drawn from data in Table 4). As can be seen, covariate distribution is relevantly similar across a, b, c, d, e, f, g, and h. The difference between the subsets {a, b}, {c, d, e}, and {f, g} can be explained by the restriction to different charge types in the three subsets. However, the covariate distributions in {i, j} are clear outliers. In our interpretation, these differences in covariate distribution signal strong biases introduced by the construction procedures, and the datasets are so dissimilar to the unmatched dataset as to recommend against use.

Another concern is hidden confounding variables. Our model only tracks the link between *the considered* covariates and COMPAS's accuracy; other covariates that plausibly make a difference to that accuracy and fairness are not tracked. For instance, suppose that racial profiling by police increased likelihood of recidivism among African-Americans in this dataset – that is, that AAs in this datatset recidivized at higher base rates not because they committed more crimes, but because they were more frequently arrested and charged. That this difference contributes to differential accuracy of COMPAS is another source of unfairness, but not one that our model discerns or measures. (The same is true of the original ProPublica analysis.) Similarly, where the treatment variable (here "race") is correlated to hidden variables that are not included in the model, but plausibly have causal influence on recidivism – such as socio-economic status (SES) – then estimates of accuracy and unfairness may be distorted. For instance, if SES is a causal contributor to recidivism rates, and SES is correlated to race, then the matched datasets constructed here cannot decide whether the unfairness is due to discrimination by race, or rather by SES.

Yet another concern is non-independent variables. Our model treats the covariates *age*, *sex*, *priors count*, *charge*, and *race* as causally independent of one another. But suppose that race, via the influence of racial profiling, makes a difference to how likely an individual is to have been charged previously (priors count), or the type of charge brought against them (charge description). Then analyses on matched groups will underestimate the extent of unfairness in the systems in question, since the lower fairness scores when matching on priors fails to consider unfairly different pathways that lead to different prior counts in the two groups. (For some insight on hidden and non-independent variables and the difficulty of estimating their influence, see Wang et al. 2023 and Tolbert 2024.)

Conclusions

Regarding our motivating questions, our results suggest the following:

(1) Matching tends to reduce estimates of unfairness. The reasons for this reduction are unclear; two possible sources are (a) increased precision due to greater covariate balance and (b) increased bias due to matching procedure.

(2) Matching tends to reduce the difference between some fairness metrics (e.g. EO and PPVP), thereby partly resolving the problem of fairness impossibility in practice.

(3) Matching can contribute robustness to the conclusion that a system exhibits unfairness, and (to some extent) the type, pattern, and extent of this unfairness.

(4) Matching can introduce new biases, and its power is limited when hidden and non-independent variables are possible and (especially) likely.

The authors thank Olusola Olabanjo, Gabriella Waters, and three anonymous reviewers for comments on the manuscript.

References

Austin, P. 2009. "Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research." *Communications in Statistics – Simulation and Computation* 38 (6): 1228-1234.

Barocas, S., M. Hardt, & A. Narayan. 2019. *Fairness & Machine Learning*. https://fairmlbook.org/

Beigang, Fabian. 2023. "Reconciling Algorithmic Fairness Criteria." *Philosophy & Public Affairs* 51 (2): 166-190.

Bell, A., L. Bynum, N. Drushchak, L. Rosenblatt, T. Zakharchenko, and J. Stoyanovic. 2023. "The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice." *FAccT '23, June 12-15, 2023. Chicago IL*. (accessed 8-7-2023)

Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Oversampling Technique." *Journal of Artificial Intelligence Research* 16: 321-357.

Chouldechova, A. 2017. "Fair prediction with disparate impact." *Big Data* 5 (2): 153-163.

Defrance, M. & De Bie, T. 2023. "Maximal Fairness." *FAccT '23, June 12-15, 2023. Chicago IL*. (accessed 8-7-2023)

Dietrich, W., C. Mendoza, and T. Brennan. 2016. “COMPAS risk scales: Demonstrating accuracy equity and predictive parity.” Technical report, Northpointe: http://www.northpointeinc.com/northpointe-analysis

Fraenkel, A. 2020. *Fairness and Algorithmic Decision-Making*. https://afraenkel.github.io/fairness-book/intro.html

Kearns, D. & Roth, A. 2020. *The Ethical Algorithm*. Oxford University Press.

Larson, J., M. Surya, L. Kirchner, and J. Angwin. 2016. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, accessed 11-1-2023.

Saleiro, P., Rodolfa, K., and Ghani, R. 2020. “Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial.” *KDD* ’20.

Stuart, E. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25 (1): 1-21.

Tolbert, A.W. 2024. “Causal Agnosticism About Race: Variable Selection Problems in Causal Inference.” *Philosophy of Science* 2024: 1-11.

Wang et al. 2023. “Against predictive optimization.” SSRN 2023 (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4238015)

Wongvorachan, R., He, S., and Bulut, O. 2023. “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining.” *Information* 14 (1): 54.