This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making

Shubha Guha[1] Falaah Arif Khan[2] Julia Stoyanovich[2] Sebastian Schelter[1]
[1]University of Amsterdam [2]New York University
s.guha@uva.nl fa2161@nyu.edu stoyanovich@nyu.edu s.schelter@uva.nl

***Abstract*—In this paper, we interrogate whether data quality issues track demographic group membership (based on sex, race and age) and whether automated data cleaning — of the kind commonly used in production ML systems — impacts the fairness of predictions made by these systems. To the best of our knowledge, the impact of data cleaning on fairness in downstream tasks has not been investigated in the literature.**

We first analyse the tuples flagged by common error detection strategies in five research datasets. We find that, while specific data quality issues, such as higher rates of missing values, are associated with membership in historically disadvantaged groups, poor data quality does not generally track demographic group membership. As a follow-up, we conduct a large-scale empirical study on the impact of automated data cleaning on fairness, involving more than 26,000 model evaluations. We observe that, while automated data cleaning is unlikely to worsen accuracy, it is more likely to worsen fairness than to improve it, especially when the cleaning techniques are not carefully chosen. Furthermore, we find that the positive or negative impact of a particular cleaning technique often depends on the choice of fairness metric and group definition (single-attribute or intersectional). We make our code and experimental results publicly available.

The analysis we conducted in this paper is difficult, primarily because it requires that we think holistically about disparities in data quality, disparities in the effectiveness of data cleaning methods, and impacts of such disparities on ML model performance for different demographic groups. Such holistic analysis can and should be supported by data engineering tools, and requires substantial data engineering research. Towards this goal, we discuss open research questions, envision the development of fairness-aware data cleaning methods, and their integration into complex pipelines for ML-based decision making.

***Index Terms*—responsible data management; data cleaning; data preparation; fairness in machine learning**

I. Introduction

Software systems that learn from user data with machine learning (ML) are in ubiquitous use in critical decision-making processes such as loan approvals, hiring, and prioritizing access to medical interventions. Unfortunately, if left unchecked, such applications often reproduce or even amplify pre-existing bias in the data, and may lead to unlawful discrimination [1].

Most ML applications in production are data-intensive, and require data cleaning [2]. Such applications regularly acquire new training data in short intervals (e.g., nightly from log files), and subsequently retrain and redeploy models, which then make predictions on previously unseen data. Real-world data — processed by production ML systems — inevitably contains data errors [3], [4], [5], [6]. Due to large data volumes and short redeployment intervals, data quality issues are often addressed with automated cleaning techniques (e.g., to impute missing values, which many ML models cannot handle directly).

What is the impact of data errors and automated cleaning on model performance, both overall and for subsets of the data corresponding to different demographic groups? This question is both crucial and understudied, with very real implications for production ML systems currently used for critical decision-making. There are indications that data from historically disadvantaged groups may be more likely to suffer from poor quality, such as higher occurrence of missing values [7]. Such "heteroskedastic noise" in the data, in turn, has the potential to negatively impact ML model fairness [8]. Yet, while there is plenty of evidence that data quality issues hurt the predictive accuracy of ML models [5], it is unclear whether (1) poor data quality tracks membership in disadvantaged groups, and (2) attempts to improve data quality through automated cleaning impact the *fairness* of ML models (e.g., by amplifying disparities in prediction quality among groups).

Related work and research gap. To the best of our knowledge, these questions have not been investigated in prior work. On the one hand, the growing body of work on joint cleaning and learning [9], [10], [11], [5] focuses on predictive accuracy but not on fairness. On the other hand, research on fairness in ML usually ignores data quality issues; it is common, for example, to simply remove tuples with missing values from the data before experimentation [12], [13]. Moreover, existing data-centric work on fairness either focuses on coverage (e.g., under-representation) at training time [14], [8], [15] (and not on repairing erroneous tuples), or it introduces synthetically-generated errors only [16], [17], [18], making it difficult to judge how representative the results are of real world settings.

Why should we care about fairness at the data cleaning stage? A data error — and it's subsequent repair — is a purely technical conception. However, the mechanisms that lead to data errors, specifically in social domains where the data is of and about people, are not purely technical in nature. There exist powerful social, political and legal systems that affect people's outcomes, as well as how data about these outcomes is collected, curated and shared. An accessible way to think about this is through the data-mirror metaphor [1]: data is a reflection of the world. A reflection cannot by itself tell us whether and why it is distorted. We must instead make assertions about the world, and then examine whether the data is a faithful reflection of the world, or whether it is

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

distorted. Hence, data cleaning approaches that fail to take into account broader normative thinking and that narrowly focus on the statistical properties of the data itself will be sub-optimal, in that they will fail to detect and correct for real-world mechanisms that cause data errors.

Further, fairness is swiftly becoming a major desideratum in ML systems alongside accuracy [19], [20], [21]. In this work we take a group fairness perspective, namely, that unfairness is quantified by the disparity in model performance aggregated over socially privileged and disadvantaged groups, respectively. From a purely technical perspective, model unfairness is an indication that the model does not perform "equally well" on all parts of the input space. Hence, even absent any ethical/moral justification, a data intervention (such as cleaning) that results in disparate model performance (unfairness) is problematic.

Why is it difficult to study the effect of data cleaning on fairness? A major challenge in studying the impact of automated data cleaning on model fairness is that there is no "clean" ground truth available for datasets that are commonly used for ML fairness research. This means that there is no "correct" benchmark against which we can evaluate the performance of error detection and data repair/cleaning techniques.

Furthermore, such datasets are hard to clean manually, in part because validating data errors would require access and corroboration through secondary data sources (such as bank records or medical files), which is expensive and time-consuming. More importantly, datasets for fairness-related research are by definition at the person-level — this is data of people, from critical domains such as finance or healthcare. There are significant legal and ethical privacy challenges to collecting, maintaining, and sharing such data. For example, requirements for data deletion in the case of withdrawn consent in the European Union is dictated by the General Data Protection Regulation (GDPR).

Even for popular datasets from the research ecosystem — such as the ones used in this study — it is infeasible to obtain "ground truth" information regarding data errors. This is because samples are usually de-identified before release (to maintain privacy) and so there is no way to map a tuple in the dataset to a person in the real world. Further, as a systemic issue, the origin and provenance of fairness datasets is poorly understood [22], [23].

For these reasons, rather than attempting to directly quantify data quality, we focus on automated data cleaning in this work.

Research questions. We tailor our research questions to address two common stages of automated data cleaning: (1) error detection, which flags potentially erroneous tuples, and (2) data repair, which attempts to correct the erroneous tuples:

- *RQ1. Does the incidence of data errors track demographic group membership in ML fairness datasets?*
- *RQ2. Do common automated data cleaning techniques impact the fairness of ML models trained on the cleaned datasets?*

To address *RQ1*, we analyze the tuples flagged by common error detection strategies in five widely used fairness benchmark datasets, with respect to groups based on sex, race, and age (Section III). To address *RQ2*, we conduct an empirical study of the impact of data cleaning on model fairness (Section V), by applying common automated data cleaning techniques to the potentially erroneous tuples detected in *RQ1*. We consider single-attribute group definitions as well as intersectional definitions with multiple sensitive attributes. Our study involves training and evaluating more than 26,000 models and, in contrast to existing work, does not inject synthetic noise but works with the raw data as provided.

Key findings. We summarize our key findings in the following.

- We find that higher rates of missing values are associated with membership in historically disadvantaged groups. However, for other types of data errors, we do not find sufficient evidence that poor data quality tracks demographic group membership, both with single-attribute and intersectional group definitions (Section III).
- We find that cleaning missing values is unlikely to have a negative impact on accuracy, but is likely to cause unfairness at the single-attribute level. Interestingly, however, we also find that cleaning missing values improves fairness in outcomes for intersectional groups, underscoring the need to select fairness definitions — here, how we define protected groups — carefully, based on the specific context of decision-making (Section V).
- We find that automated cleaning of outliers is very likely to worsen accuracy and have an insignificant impact on fairness. Further, when it does impact fairness, it is more likely to worsen fairness than to improve it — at the group level for single-attributes and by causing in-group unfairness for intersectional groups (Section V).
- We find that repairing label errors is very likely to have a strong effect on both accuracy and fairness. Accuracy is improved in most cases, while the direction of impact on fairness (positive or negative) is highly metric specific. For single-attribute groups, cleaning label errors is very likely to improve fairness according to the equal opportunity measure, but worsen fairness according to the predictive parity measure, and these effects are even stronger for intersectional groups (Section V).

Our findings are significant because they potentially implicate many production ML systems. The observed effect varies based on dataset, fairness metric, group definition, and type of error being repaired. Notably, in many cases, we do not encounter a configuration that simultaneously improves both fairness and accuracy (Section V). In Section VI we outline which cleaning techniques, error detection strategies and ML models turned out to be the best performing with respect to fairness and accuracy in our study.

We discuss the implications of our findings, and outline research challenges and directions for follow-up work in Section VII. We provide the code and results for our study, and experiments for reproducibility and follow-up research.[1]

[1] https://github.com/amsterdata/demodq

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

II. Preliminaries

Benchmark datasets. We use five publicly available datasets listed in Table I from three source domains: census, finance, and healthcare. These datasets are commonly used in research on responsible machine learning and data management [7], [8], [12], [23]. Each dataset is associated with a binary classification task. In our setup, the positive class always corresponds to the desirable outcome for the individuals in the dataset, such as being considered creditworthy or being prioritized for access to healthcare resources. Note that the choice of sensitive attribute(s) is taken from existing research on these datasets [7], [8], [12], [23].

TABLE I
Datasets for our experimental study.

name	source	number of tuples	sensitive attributes
adult	census	48,844	sex, race
folk	census	378,817	sex, race
credit	finance	150,000	age
german	finance	1,000	age, sex
heart	healthcare	70,000	sex, age

The adult[2] dataset contains demographic and financial data, and the target variable denotes whether a person earns more than 50,000 dollars per year. This dataset has been used extensively to evaluate fairness in predictions of creditworthiness. Recent work proposes to "retire" this dataset due to both unclear data origins and the apparent — and unrepresentative — class-label imbalance, which renders the prediction task unrealistic [23]. We include this dataset in our study to complement these concerns from the data management perspective, exposing additional data quality issues.

The folk[3] dataset is based on US census data and has been proposed as a replacement for the problematic adult [23] dataset, to be used for financial decisions. We use the subset of the data from the census in California in 2018, and replicate the prediction task from adult.

The credit[4] and german[5] datasets contain financial information, and the target variable denotes whether a person has a good credit score. We remove the foreign_worker attribute from the german dataset, due to unclear semantics. According to the documentation of the data, more than 96% of the records would belong to foreign workers. This interpretation is likely due to an error, and the attribute is handled differently in other derived versions of the dataset. We derive the sex attribute for the german credit dataset from the personal_status attribute, which encodes each unique combination of marital status and sex.

The heart[6] dataset consists of patient measurements with respect to cardiovascular diseases, and the target variable denotes the presence of a heart disease. This dataset has been used to evaluate fairness of predictive tasks that allocate access to priority medical care for individuals.

[2]https://archive.ics.uci.edu/ml/datasets/adult
[3]https://github.com/zykls/folktables
[4]https://www.kaggle.com/c/GiveMeSomeCredit
[5]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[6]https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Protected groups. We investigate disparities with respect to sensitive attributes based on which unlawful discrimination in decision-making has been observed [1], e.g., violating US labor law [24] or European non-discrimination law [25].

Single-attribute groups. Given a sensitive attribute, we partition the data into tuples belonging to a *privileged group* and all other tuples as belonging to a *disadvantaged group*. Following prior work [7], [8], [12], [23], [26], we consider sex (with 'male' as the privileged group), race (with 'white' as the privileged group) and age (with people older than 30, 25 and 45 years old as the privileged group in the credit, german and cardio datasets, respectively). Note that which demographic group is considered privileged vs. disadvantaged is task-specific, and is designated as appropriate for the benchmark datasets and tasks described here. For example, older age is considered privileged in the context of lending, but disadvantaged in the context of hiring.

Intersectional groups. Intersectionality [27] is the idea that interlocuting axes of discrimination give rise to a social experience that cannot be understood in terms of single-axis effects. Crenshaw [27] writes: "Focusing on the most privileged group members marginalizes those who are multiply-burdened and obscures claims that cannot be understood as resulting from discrete sources of discrimination." A famous example is the *Gender Shades* project [28], which showed that facial recognition software performs significantly worse on Black women than on other social groups.

We consider the intersection of sex and race in the adult and folk dataset (with 'white, male' as the intersectionally privileged group and 'black, female' as the disadvantaged group), the intersection of sex and age in the german and heart datasets (with 'male, over 25' and 'male, over 45' as the intersectionally privileged group, and 'female, under 25' and 'female, under 45' as the intersectionally disadvantaged group, respectively). The credit dataset does not include an additional demographic attribute and is therefore left out of this analysis. Note that unlike the single-attribute groups, our chosen intersectional group definitions do not induce a partition over the full dataset, i.e., we exclude tuples that are privileged along one axis and disadvantaged along another.

Fairness metrics. In our experimental study, we report the following group fairness metrics:

- *Predictive Parity (PP)* is satisfied if a classifier has equal precision for the subjects in the privileged and disadvantaged groups. This metric is computed as $\frac{TP_{\text{priv}}}{TP_{\text{priv}}+FP_{\text{priv}}} - \frac{TP_{\text{dis}}}{TP_{\text{dis}}+FP_{\text{dis}}}$, and denotes equal probability of a correct positive prediction for the groups.
- *Equal Opportunity (EO)* is satisfied if a classifier has equal recall for the subjects in the privileged and disadvantaged groups. This metric is computed as $\frac{TP_{\text{priv}}}{TP_{\text{priv}}+FN_{\text{priv}}} - \frac{TP_{\text{dis}}}{TP_{\text{dis}}+FN_{\text{dis}}}$.

The choice of a particular group fairness metric always involves a value-based decision [29]. In line with existing research [13], we choose these metrics from dozens of existing fairness metrics because they intuitively represent the opposing interests of two key stakeholders in many decision making processes — individuals who seek access to resources, and

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

vendors who grant access. For example, in lending, the bank, on the one hand, wants high precision (to avoid giving loans to creditors who might not have the means to repay them), while customers, on the other hand, want high recall (to avoid being denied a loan that they would have been able to repay).

In addition, we report results for *Demographic Parity (DP)*, which aims for the same positive prediction rates between groups (independent of the historical outcome distribution captured in the training data). This metric assumes no intrinsic differences between the groups, which may not always be realistic, e.g., when making decisions about medical prioritisation for different age groups. Demographic parity is computed as $\frac{TP_{\text{priv}}+FP_{\text{priv}}}{TP_{\text{priv}}+FP_{\text{priv}}+TN_{\text{priv}}+FN_{\text{priv}}} - \frac{TP_{\text{dis}}+FP_{\text{dis}}}{TP_{\text{dis}}+FP_{\text{dis}}+TN_{\text{dis}}+FN_{\text{dis}}}$.

Error detection strategies. We apply common error detection strategies that have been proposed in the data cleaning literature [3], [30], [31] and are also used in studies about the impact of data cleaning on machine learning tasks [5].

Missing values. We identify tuples with `missing values` by detecting NULL and NaN values in the datasets.

Outliers. We detect numerical outliers with the following techniques: (i) `outliers-sd` – we consider a value of a column to be an outlier if it is more than n standard deviations away from the mean of the column (with $n = 3$); (ii) `outliers-iqr` – we consider a value of a column to be an outlier if it lies outside of the interval $[p_{25} - k \cdot \text{iqr}, p_{75} + k \cdot \text{iqr}]$ with $k = 1.5$. Note that iqr refers to the interquartile range defined as the difference between the 75th and 25th percentile of the column distribution: $\text{iqr} = p_{75} - p_{25}$; (iii) `outliers-if` – a tuple is considered to be an outlier if it is identified as such by an isolation forest trained on the data with a contamination parameter of 0.01. Note that `outliers-sd` and `outliers-iqr` are univariate techniques that inspect individual attributes, while the multivariate approach `outliers-if` inspects whole tuples.

Label errors. An ML-specific type of error is mislabeled examples: tuples with the wrong prediction label assigned to them. Such errors have recently received a lot of attention, due to the fact that they are pervasive in widely used benchmarking datasets for ML [32]. We detect tuples with potential `label errors` with the cleanlab [33] library, using a logistic regression model as the base classifier. Cleanlab identifies label errors in datasets by estimating the joint distribution between noisy (given) labels and uncorrupted (unknown) labels.

Limitations. Unfortunately, there are no known integrity constraints available for the datasets (e.g., in the form of functional dependencies or denial constraints [34]) and no verified sets of clean records, which prevents us from applying more advanced cleaning and error detection techniques such as HoloClean [35], HoloDetect [36] or kNN-Shapley [37]. We consider it an interesting avenue for future work to include these approaches on appropriate datasets and tasks.

Automated repair methods. We apply standard techniques for repairing erroneous tuples, which are implemented in popular data science packages such as scikit-learn[7] or pandas, and used in existing studies on joint cleaning and learning [5]. We apply several methods to impute missing values, namely, via the column mean, median or mode for numerical columns, and via the mode or a constant "dummy" value for categorical columns. We repair outlier values in numerical columns by replacing them with the mean, median or mode of the column. We repair label errors by flipping the labels of flagged tuples.

III. Indications of Demographically Disparate Data Quality Issues

To address *RQ1*, we search for cases in which our error detection strategies flag significantly different fractions of the privileged and disadvantaged groups, based on sex, race and age. For a dataset D, let the Boolean predicate $\text{priv}(t)$ evaluate if tuple $t \in D$ belongs to the privileged group. Further, let the Boolean error function $\sigma(t)$ evaluate if t is considered erroneous by detection strategy σ.

To identify statistically significant disparities, we compute the number of erroneous tuples $|\{t \in D \,|\, \text{priv}(t) \wedge \sigma(t)\}|$ from the privileged group, the number of erroneous tuples $|\{t \in D \,|\, \neg\text{priv}(t) \wedge \sigma(t)\}|$ from the disadvantaged group, and conduct a G^2 significance test with a threshold of $p = .05$. We report only cases that pass this test. We run the error detectors to identify data quality disparities among the protected groups described in Section II, and report the results for single-attribute and intersectional group definitions in Figures 1 and 2, respectively.

Results. We find that all three data error types (missing values, outliers and label noise) are frequently detected in the research datasets.[8] These errors are flagged in disparate proportions for different datasets and protected group definitions, and, strikingly, error detection strategies often identify large fractions of erroneous tuples (e.g., up to 51% of the tuples of a particular group). Notably, `adult` — one of the most widely used datasets in fair ML — is the only dataset for which all five error detectors flag tuples with significant disparities, for both single-attribute and intersectional group definitions. We interpret this as additional evidence that it is time to "retire" `adult` [23].

Disparities in missing values. We find that tuples from the disadvantaged group are subject to missing data more frequently: in 4 out of 6 cases (dataset/sensitive attribute pairs) with single-attribute group definitions, and in 2 out of 3 cases with intersectional group definitions.

Disparities in outliers. We see a mixed picture with respect to outliers, where the trends vary strongly based on detection technique and group definition (single-attribute or intersectional). There are several cases where we encounter disparate proportions of outliers with only a particular detection technique but not with others. Additionally, we find that the number of outliers detected heavily varies based on the applied detection strategy.

Disparities in predicted label errors. For label errors, we find that, in the majority of cases (for both single-attribute and intersectional group definitions), the fraction of tuples

[7]https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

[8]Note that the `heart` dataset has no missing values at all.

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

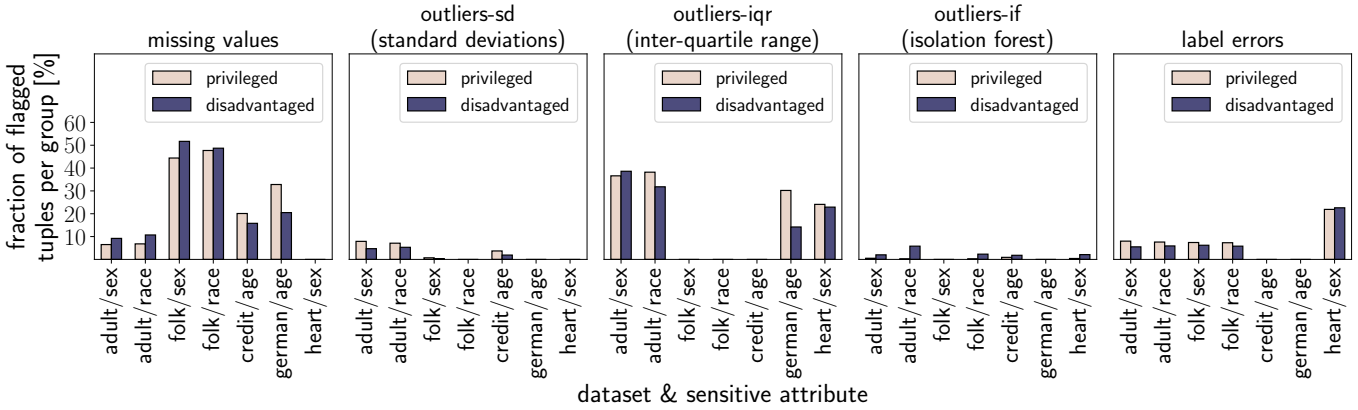

Fig. 1. Single-attribute analysis: Proportions of tuples flagged by common error detection strategies for the privileged and disadvantaged groups.

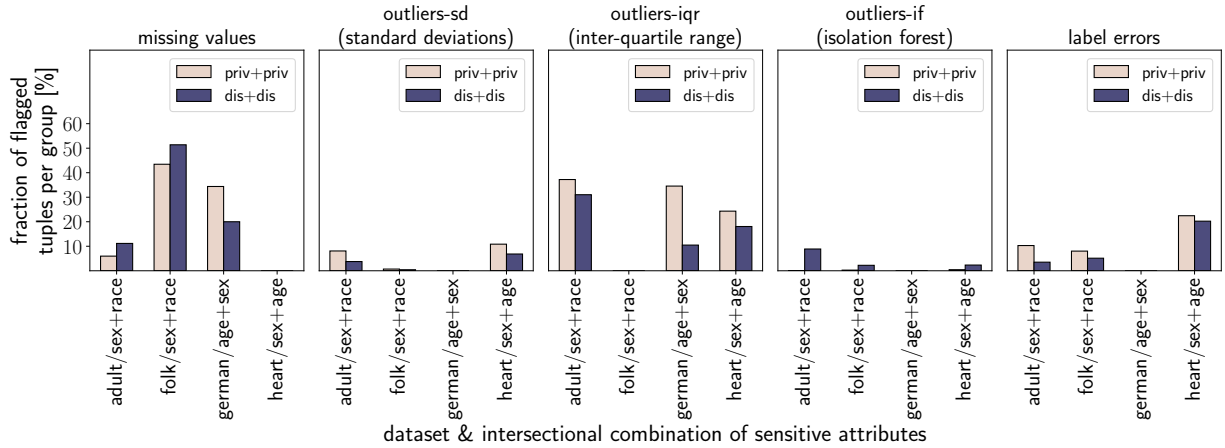

Fig. 2. Intersectional analysis: Disparate proportions of tuples flagged by common error detection strategies for the intersectionally privileged and intersectionally disadvantaged groups.

from the privileged group in the mislabeled data is higher than the fraction of tuples from the disadvantaged group. (Recall that these labeling errors are predicted, and that we do not have access to the ground truth.) We drill in on the type of label error — false positive or false negative — and find no significant differences between the privileged and the disadvantaged groups in most cases. However, in one case (single-attribute groups in the `heart` dataset) the fraction of false positives was significantly higher for the privileged group than for the disadvantaged (57.7% vs. 52.2%, respectively), and the trend was reversed for the false negatives (42% vs. 47.8%, respectively). This is potentially problematic, because false positives can amplify the advantage, while false negatives can exacerbate the disadvantage for the respective groups.

Discussion. Overall, while we do find strong indication of a large number of data quality issues in benchmark datasets, we do not find sufficient evidence that these potential data errors track demographic group membership with respect to sex, race and age. In the `folk` and `heart` datasets, overall, errors are detected more frequently in the disadvantaged group, but the disparity in errors between groups is small. In the `credit` and `german` datasets, where the disparity in the incidence of errors across groups is large, errors do not systematically occur more frequently for the disadvantaged group. Interestingly, the fraction of detected errors is comparable across groups, irrespective of whether groups are defined based on a single-attribute (Figure 1) or intersectionally (Figure 2). This finding is consistent with one of two possibilities. The first is that data errors are, in fact, uniformly distributed across social groups, in which case the answer to *RQ1* is: no, data quality does not track group membership. The second possibility is that data errors do, in fact, occur more frequently for some groups than for others, *but* the effectiveness of error detection methods is also non-uniform across groups.

Explicitly missing values are the only error type in our study where detection is straightforward: a tuple either contains a NULL or it does not. For outliers and label errors, however, we cannot tell what fraction of errors have been discovered/missed because (as discussed in Section I) we do not have access to the ground truth. Recall that missing values were the only error type for which we did detect a demographic disparity in data quality. We posit that the outlier and mislabel detection techniques we use could be incurring a high number of false negatives, i.e., current techniques are only capable of identifying errors caused by mechanisms that affect the majority

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

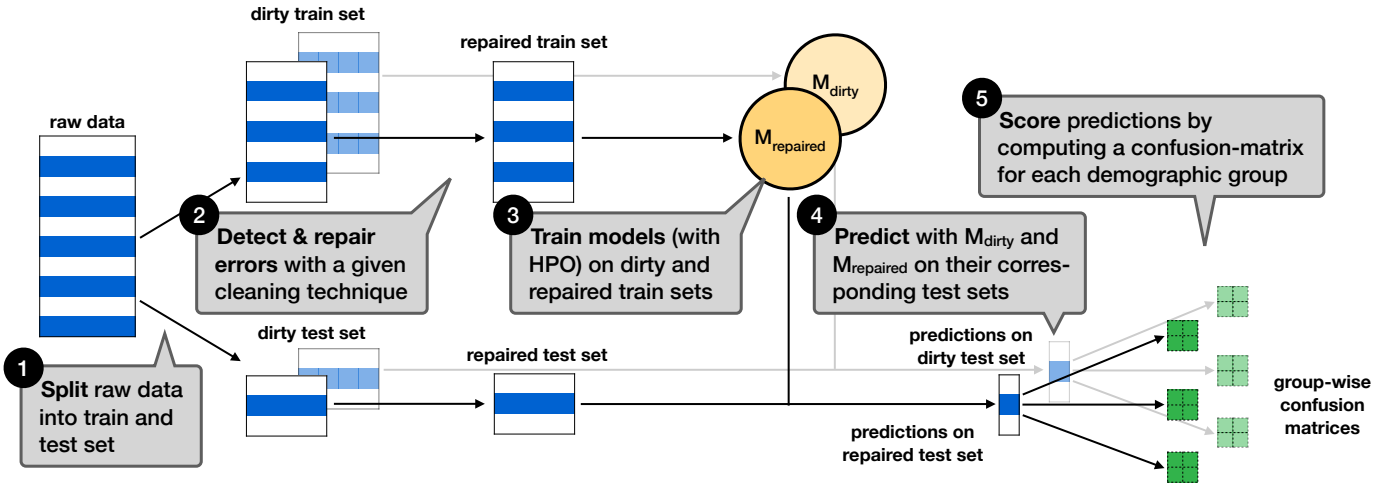

Fig. 3. Overview of our experimentation framework. For each experimental configuration (dataset/model/error/repair), we ❶ split the dataset into train/test sets; ❷ save the original raw data as a dirty version and apply the repair strategy to the raw data to generate a repaired version; ❸ train a classifier on the dirty train data and another classifier on the repaired train data; ❹ generate predictions on the dirty test set using the classifier trained on dirty data and predictions on the repaired test set using the classifier trained on the repaired train data; and ❺ score each model on accuracy and fairness and compare the scores computed from repaired data with the scores computed from dirty data to assess the impact of auto-cleaning for this configuration.

(here, privileged) group. This hypothesis is further supported by the fact the general trend for detected errors stays the same whether we look at disparities between single-attribute groups or intersectional groups.

In summary, these results neither support nor conclusively refute the hypothesis on which we focused under *RQ1*, namely, that data from historically marginalized groups is more likely to be erroneous. This motivates our large-scale empirical study for a principled answer to *RQ2*, where we look at the downstream effect of repairing these errors on model accuracy and fairness instead.

IV. Experimentation Framework

We introduce our experimentation framework, before we detail our empirical study. We extend the existing CleanML benchmark [5] for joint data cleaning and model training, to additionally compute fairness metrics for the cleaning impact. Our goal is to enable fairness-related experimentation with minimal effort. In order to achieve that, our extension enables a declarative definition of sensitive attributes per dataset, after which the benchmarking framework will automatically compute the corresponding fairness metrics. In this section, we give an overview of our implementation.

Declarative definition of datasets. CleanML already allows users to declaratively define datasets to experiment on, by specifying the data location (`data_dir`), the `error_types` to clean, the `label` to predict and the attributes to hide from the classifier (`drop_variables`). We extend CleanML with three additional datasets: `folk`, `heart`, and `german`. We also add to the dataset definitions the ability to specify `privileged_groups` present in a dataset. Membership in a privileged group is defined by a binary predicate on the sensitive attribute, e.g., that the age of a person is higher than 25. The declarative definition of the `german` credit dataset looks as follows for example:

```
German = {
  "data_dir": "German",
  "error_types": [
    "missing_values",
    "outliers",
    "mislabels"
  ],
  "drop_variables": [
    "age",
    "personal_status",
    "sex",
    "foreign_worker"
  ],
  "label": "credit",
  "ml_task": "classification",
  "privileged_groups": [
    ("age", operator.gt, 25),
    ("sex", operator.eq, "male")
  ],
}
```

Evaluation process. Figure 3 shows how we use the CleanML framework: for each experimental configuration (dataset/model/error/repair), we ❶ split the dataset into train/test sets; ❷ save the original raw data as a dirty version and apply the repair strategy to the raw data to generate a repaired version; ❸ train a classifier on the dirty train data and another classifier on the repaired train data; ❹ generate predictions on the dirty test set using the classifier trained on dirty data and predictions on the repaired test set using the classifier trained on the repaired train data; and ❺ score each model on accuracy and fairness, and compare the scores computed from repaired data with the scores computed from dirty data to assess the impact of auto-cleaning.

Automated computation of group-wise confusion matrices per cleaning technique. During benchmark execution, we automatically compute confusion matrices (counting the number of true negative, false positive, false negative, and true positive predictions) for the privileged and disadvantaged groups, per cleaning technique. To get insights into intersectional groups, we additionally compute the confusion matrices for combinations of two groups. Our design decision to compute the "raw" confusion matrices gives us the flexibility to use a broad range

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

of fairness metrics during analysis, including the commonly-reported group fairness metrics for binary classification [29].

For example, the following JSON snippet represents the results for training a logistic regression classifier on the `german` dataset with missing values from the training data in numerical columns imputed with their mean and in categorical columns with a "dummy" indicator. CleanML already computes general metrics for this experiment, such as the accuracy on the training set (`train_acc`) and the accuracy of a given cleaning technique on the test set, e.g., imputing missing values in numerical columns with their mean and inserting a "dummy" indicator for missing values in categorical attributes (`impute_mean_dummy_test_acc`). For each such cleaning technique, we automatically compute the confusion matrices for all definitions of privileged and disadvantaged groups (and their intersectional combinations) and record the resulting counts. For example, the key `impute_mean_dummy__sex_priv__age_priv__fp` denotes the number of false positive predictions on the test set for the privileged intersectional group with respect to sex and age (e.g., male persons older than 25 years) when applying the `impute_mean_dummy` cleaning technique.


```
"German/v235/missing_values/impute_mean_dummy/logreg/6130": {
  "best_params": { "C": 0.370200059179964 },
  "train_acc": 0.8223048761489329,
  "val_acc": 0.747454133168419,
  ...
  "impute_mean_dummy_test_acc": 0.7133333333333334,
  "impute_mean_dummy_test_f1": 0.46913580246913583,
  ...
  "impute_mean_dummy__age_priv__tn": 145,
  "impute_mean_dummy__age_priv__fp": 22,
  "impute_mean_dummy__age_priv__fn": 39,
  "impute_mean_dummy__age_priv__tp": 24,
  ...
  "impute_mean_dummy__age_dis__tn": 31,
  "impute_mean_dummy__age_dis__fp": 16,
  "impute_mean_dummy__age_dis__fn": 9,
  "impute_mean_dummy__age_dis__tp": 14,
  ...
  "impute_mean_dummy__sex_priv__age_priv__tn": 114,
  "impute_mean_dummy__sex_priv__age_priv__fp": 13,
  "impute_mean_dummy__sex_priv__age_priv__fn": 31,
  "impute_mean_dummy__sex_priv__age_priv__tp": 19,
  ...
  "impute_mean_dummy__sex_dis__age_dis__tn": 17,
  "impute_mean_dummy__sex_dis__age_dis__fp": 9,
  "impute_mean_dummy__sex_dis__age_dis__fn": 5,
  "impute_mean_dummy__sex_dis__age_dis__tp": 10,
  ...
}
```


Reproducibility. A crucial point of experimental work is to ensure the reproducibility of the results. CleanML already has a solid foundation for this by making all randomised decisions like dataset splits depend on globally specifiable random seeds. Furthermore, the framework supports stopping and resuming computations, such that it will make sure not to repeat previously conducted experiments. We implement our extensions to be compatible with the existing design for reproducibility in CleanML (e.g., we re-use existing splits for computing fairness metrics).

Alarmingly, while conducting our experimental study, we identified a severe reproducibility issue in CleanML: the key-value mapping between the names of the cleaning techniques and the resulting metric values is randomly reshuffled in some cases due to a software bug, which leads to unreliable and non-reproducible results. We fixed this issue in our codebase, and also contacted the CleanML authors via a bug report in their repository,[9] which led to them also addressing the issue. To additionally verify the reproducibility of our results, we ran our experimental study with 26,000 model evaluations twice on identical machines with the same operating system and software packages, and validated that we obtain the same results from both runs.

V. Impact of Automated Data Cleaning on Fairness

In the following, we discuss the setup and results of our empirical study to address *RQ2*.

Classification models and training procedure. We use three ML model types, each of which we tune using 5-fold cross-validation: logistic-regression (`log-reg`) with a tuned learning rate, nearest neighbors (`knn`) with a tuned number of neighbors, and gradient-boosted decision trees (`xgboost`) with a tuned maximum tree depth. During each run, we sample 15,000 records from a given dataset, randomly split these into train and test set, and evaluate five different model instances (with different random seeds for the hyperparameter search) per split. We repeat this 20 times per configuration (dataset/model/error/repair), resulting in the training and evaluation of 100 models per configuration.

Evaluation. For each run, we evaluate the predictions of the corresponding model (learned on the repaired training set) on an equivalently repaired test set. We compare these predictions to the "dirty" baseline predictions of a model, trained and evaluated on the "dirty" version of the data, as described in Section IV. We aggregate confusion-matrix values over the samples from the privileged and disadvantaged groups to compute the fairness metrics described in Section II.

Error detection and repairs. We detect errors and repair flagged tuples as outlined in Section II.

Missing values. We apply different variants of missing value imputation. Note that most classifiers cannot naturally handle missing values, which requires us to define a modified version of the data as the 'dirty' version. For the 'dirty' setup, we remove tuples with missing values from the training data and impute them with the mean for numerical columns and dummy for categorical columns on the test data. Note that one cannot simply remove tuples with missing values from the data during prediction in a real-world setup, therefore we have to impute on the test set as well for consistency. For other types of errors, missing values have to be removed from the data beforehand.

Outliers. We detect outliers and impute them as outlined earlier in Section II. In the "dirty" setup, we simply retain the outliers in both the train set and the test set.

Mislabels. For labeling errors, we run cleanlab for detection and flip the labels of identified tuples as a repair technique. For the "dirty" setup, we leave the labels as is in both train and test set. Note that we never flip labels on the test set, as this would make the prediction results incomparable with the other experiments.

[9] https://github.com/chu-data-lab/CleanML/issues/3

Results. We evaluate 26,400 models in total, and compute a result table from our experiments, where each row contains the result of a particular configuration with respect to a dataset, sensitive attribute, fairness metric, model, error type, detection method, repair method, and indicators for the impact on fairness and accuracy. The impact on fairness as well as the impact on accuracy of a configuration can be positive, negative or insignificant. We determine this by comparing the resulting 100 fairness and accuracy scores from the "dirty" baseline (with no cleaning) to the scores from a cleaning configuration (dataset, sensitive attribute, fairness metric, error, detection, repair). We leverage a sequence of paired sample t-tests as proposed by CleanML [5] with a threshold for the p-value of .05 adjusted by Bonferroni correction to account for multiple hypothesis tests.

TABLE II
IMPACT OF AUTO-CLEANING MISSING VALUES FOR SINGLE-ATTRIBUTE GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	3.7% (4)	1.9% (2)	16.7% (18)	22.2% (24)
	insign.	5.6% (6)	34.3% (37)	7.4% (8)	47.2% (51)
	better	3.7% (4)	7.4% (8)	19.4% (21)	30.6% (33)
		13.0% (14)	43.5% (47)	43.5% (47)	

TABLE III
IMPACT OF AUTO-CLEANING MISSING VALUES FOR SINGLE-ATTRIBUTE GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	1.9% (2)	15.7% (17)	19.4% (21)	37.0% (40)
	insign.	9.3% (10)	25.9% (28)	13.0% (14)	48.1% (52)
	better	1.9% (2)	1.9% (2)	11.1% (12)	14.8% (16)
		13.0% (14)	43.5% (47)	43.5% (47)	

TABLE IV
IMPACT OF AUTO-CLEANING MISSING VALUES FOR SINGLE-ATTRIBUTE GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	3.7% (4)	13.0% (14)	19.4% (21)	36.1% (39)
	insign.	9.3% (10)	12.0% (13)	18.5% (20)	39.8% (43)
	better	0.0% (0)	18.5% (20)	5.6% (6)	24.1% (26)
		13.0% (14)	43.5% (47)	43.5% (47)	

Impact of repairing missing values. The effect of the automated repair of missing values on predictive parity (PP), equal opportunity (EO), and demographic parity (DP) for single-attribute group definitions are reported in Tables II to IV, and for intersectional groups in Tables V to VII, respectively.

Single-attribute groups. Repairing missing values is very unlikely to worsen accuracy (only 13% of the cases), and in most cases (approximately 50% of the time) has an insignificant impact on fairness (based on single-attribute group definitions). However, when cleaning does have an effect on fairness, the direction of the effect (positive/negative) is highly metric specific: cleaning missing values is more likely to worsen DP (36.1%) than to improve it (24.1%) and is even more than twice as likely to worsen EO (37%) than to improve it (14.8%), but is more likely to improve PP (30.6%) than worsen it (22.2%).

Recall from Section II that PP measures the disparity in group-specific precision, whereas EO measures disparity in group-specific recall. An intervention that marginally improves PP and worsens EO increases the true positive rate parity, but worsens the false negative rate parity. This means that the model ends up allocating fewer positive outcomes (more false negatives) to the disadvantaged group than the privileged group, as is seen in the results for DP, where cleaning is more likely to worsen positive rate parity than improve it.

In summary, for single-attribute group definitions, cleaning missing values only has an insignificant effect on fairness approximately half of the times, and when it does has an effect on fairness, it is likely to worsen it.

TABLE V
IMPACT OF AUTO-CLEANING MISSING VALUES FOR INTERSECTIONAL GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	0.0% (0)	0.0% (0)	5.6% (3)	5.6% (3)
	insign.	3.7% (2)	27.8% (15)	11.1% (6)	42.6% (23)
	better	3.7% (2)	14.8% (8)	33.3% (18)	51.9% (28)
		7.4% (4)	42.6% (23)	50.0% (27)	

TABLE VI
IMPACT OF AUTO-CLEANING MISSING VALUES FOR INTERSECTIONAL GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	0.0% (0)	11.1% (6)	11.1% (6)	22.2% (12)
	insign.	7.4% (4)	20.4% (11)	22.2% (12)	50.0% (27)
	better	0.0% (0)	11.1% (6)	16.7% (9)	27.8% (15)
		7.4% (4)	42.6% (23)	50.0% (27)	

TABLE VII
IMPACT OF AUTO-CLEANING MISSING VALUES FOR INTERSECTIONAL GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	7.4% (4)	3.7% (2)	27.8% (15)	38.9% (21)
	insign.	0.0% (0)	27.8% (15)	11.1% (6)	38.9% (21)
	better	0.0% (0)	11.1% (6)	11.1% (6)	22.2% (12)
		7.4% (4)	42.6% (23)	50.0% (27)	

Intersectional groups. The trends for PP and EO flip when we consider intersectional group definitions instead of single-attribute groups. Cleaning now affects both metrics in the same way: it is nearly 10 times more likely to improve PP (51.9% of the time) than worsen to it (5.6% of the time), and is also marginally more likely to improve EO (27.8%) than to worsen it (22.2%). However, it is still nearly twice as likely to worsen DP (38.9%) than to improve it (22.2%).

Together with the results from binary group definitions, this is a very interesting finding: cleaning missing values is likely to worsen fairness at the single-attribute level but to improve fairness at the intersectional level, according to PP

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

and EO. However, it fails to improve DP for both single-attribute and intersectional groups, because the base rates for these subgroups might be different, thereby making it more difficult to equalize positive rates, if desired at all.

TABLE VIII
IMPACT OF AUTO-CLEANING OUTLIERS FOR SINGLE-ATTRIBUTE GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	21.2% (40)	1.1% (2)	1.6% (3)	23.8% (45)
	insign.	21.2% (40)	25.9% (49)	14.3% (27)	61.4% (116)
	better	5.3% (10)	3.2% (6)	6.3% (12)	14.8% (28)
		47.6% (90)	30.2% (57)	22.2% (42)	

TABLE IX
IMPACT OF AUTO-CLEANING OUTLIERS FOR SINGLE-ATTRIBUTE GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	28.0% (53)	5.8% (11)	14.8% (28)	48.7% (92)
	insign.	15.9% (30)	24.3% (46)	7.4% (14)	47.6% (90)
	better	3.7% (7)	0.0% (0)	0.0% (0)	3.7% (7)
		47.6% (90)	30.2% (57)	22.2% (42)	

TABLE X
IMPACT OF AUTO-CLEANING OUTLIERS FOR SINGLE-ATTRIBUTE GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	16.9% (32)	5.3% (10)	12.2% (23)	34.4% (65)
	insign.	14.3% (27)	24.3% (46)	10.1% (19)	48.7% (92)
	better	16.4% (31)	0.5% (1)	0.0% (0)	16.9% (32)
		47.6% (90)	30.2% (57)	22.2% (42)	

TABLE XI
IMPACT OF AUTO-CLEANING OUTLIERS FOR INTERSECTIONAL GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	14.8% (16)	0.9% (1)	0.9% (1)	16.7% (18)
	insign.	28.7% (31)	25.0% (27)	8.3% (9)	62.0% (67)
	better	4.6% (5)	2.8% (3)	13.9% (15)	21.3% (23)
		48.1% (52)	28.7% (31)	23.1% (25)	

TABLE XII
IMPACT OF AUTO-CLEANING OUTLIERS FOR INTERSECTIONAL GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	15.7% (17)	0.9% (1)	16.7% (18)	33.3% (36)
	insign.	32.4% (35)	26.9% (29)	6.5% (7)	65.7% (71)
	better	0.0% (0)	0.9% (1)	0.0% (0)	0.9% (1)
		48.1% (52)	28.7% (31)	23.1% (25)	

Impact of repairing outliers. The effect of the automated cleaning of outliers on predictive parity (PP), equal opportunity (EO), and demographic parity (DP) for single-attribute group definitions are reported in Tables VIII to X, and for intersectional groups in Tables XI to XIII, respectively.

TABLE XIII
IMPACT OF AUTO-CLEANING OUTLIERS FOR INTERSECTIONAL GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	6.5% (7)	3.7% (4)	13.9% (15)	24.1% (26)
	insign.	20.4% (22)	24.1% (26)	9.3% (10)	53.7% (58)
	better	21.3% (23)	0.9% (1)	0.0% (0)	22.2% (24)
		48.1% (52)	28.7% (31)	23.1% (25)	

Outlier cleaning is in general not very helpful: it worsens accuracy in nearly half of the cases. Interestingly, independent of how groups are constructed (using a single-attribute or intersectionally), automated outlier repair has an insignificant effect on fairness (close to 60% of the time, for 3 out of 6 metric-group pairs, and close to 50% otherwise). However, when it does have an effect, it is far more likely to worsen fairness than to improve it. For example, for EO on single-attribute groups, cleaning outliers worsens fairness 48.7% of the time, and only improves it 3.7% of the time. The exception to this trend is the PP measure, on intersectional groups, where cleaning is marginally more likely to improve fairness (21.3% of the time) than worsen it (16.7% of the time). Recall from our discussion on missing value repair, that improving PP while worsening EO and DP results in group unfairness. We observe a similar trend here, for the effect of outlier repair on intersectional groups.

In summary, auto-cleaning outliers is most likely to worsen accuracy. We attribute this to the high fraction of records wrongly flagged as outliers, which we already encountered in Figure 1 of Section III. Furthermore, outlier cleaning has an insignificant impact on fairness in the majority of cases. However, when it does impact fairness, it is more likely to worsen fairness than to improve it for both single-attribute and intersectional groups.

TABLE XIV
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR SINGLE-ATTRIBUTE GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	14.3% (3)	14.3% (3)	19.0% (4)	47.6% (10)
	insign.	9.5% (2)	0.0% (0)	9.5% (2)	19.0% (4)
	better	0.0% (0)	0.0% (0)	33.3% (7)	33.3% (7)
		23.8% (5)	14.3% (3)	61.9% (13)	

TABLE XV
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR SINGLE-ATTRIBUTE GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	0.0% (0)	4.8% (1)	0.0% (0)	4.8% (1)
	insign.	0.0% (0)	0.0% (0)	14.3% (3)	14.3% (3)
	better	23.8% (5)	9.5% (2)	47.6% (10)	81.0% (17)
		23.8% (5)	14.3% (3)	61.9% (13)	

Impact of repairing predicted label errors. The effect of the automated cleaning of label errors on predictive parity (PP), equal opportunity (EO), and demographic parity (DP) for single-attribute group definitions are reported in Tables XIV

This article has been accepted for publication in IEEE Transactions on Knowledge and Data Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2024.3365524

TABLE XVI
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR SINGLE-ATTRIBUTE GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	19.0% (4)	14.3% (3)	42.9% (9)	76.2% (16)
	insign.	4.8% (1)	0.0% (0)	14.3% (3)	19.0% (4)
	better	0.0% (0)	0.0% (0)	4.8% (1)	4.8% (1)
		23.8% (5)	14.3% (3)	61.9% (13)	

TABLE XVII
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR INTERSECTIONAL GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	25.0% (3)	8.3% (1)	33.3% (4)	66.7% (8)
	insign.	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
	better	0.0% (0)	0.0% (0)	33.3% (4)	33.3% (4)
		25.0% (3)	8.3% (1)	66.7% (8)	

TABLE XVIII
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR INTERSECTIONAL GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
	insign.	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
	better	25.0% (3)	8.3% (1)	66.7% (8)	100.0% (12)
		25.0% (3)	8.3% (1)	66.7% (8)	

TABLE XIX
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR INTERSECTIONAL GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	25.0% (3)	8.3% (1)	41.7% (5)	75.0% (9)
	insign.	0.0% (0)	0.0% (0)	25.0% (3)	25.0% (3)
	better	0.0% (0)	0.0% (0)	0.0% (0)	0.0% (0)
		25.0% (3)	8.3% (1)	66.7% (8)	

to XVI, and for intersectional groups in Tables XVII to XIX, respectively. Repairing label errors is very likely to have a strong effect on both accuracy and fairness: Auto-repairing mislabels improves accuracy in over 60% of the cases, and has an insignificant effect on fairness in no more than 25% of the cases, irrespective of fairness metric and group definition. As expected, fairness according to DP is very sensitive to mislabel repair, since this intervention affects the base rates in the dataset. For both single-attribute and intersectional groups, auto-cleaning label errors worsens DP over 75% of the time.

The direction of impact (positive or negative) for PP and EO is highly metric specific. For single-attribute groups, cleaning label errors is very likely to improve EO (81% of the times), whereas for PP, cleaning is more likely to worsen fairness (47.6%) than to improve it (33.3%). These effects are even more pronounced for intersectional groups: EO improves in 100% of the cases, whereas PP is twice as likely to worsen (66.7%) than to improve (33.3%).

This is the opposite trend to what we observed with missing value repair: here, cleaning is very likely to improve EO (recall parity) and worsens PP (precision parity), while improving accuracy and worsening DP (positive rate parity). This means that the model trained on clean data becomes less conservative: false negative rate parity is improved, but false positive rate disparity increases.

From a moral standpoint this is the opposite concern to the case of unfairness due to missing value repair: for missing values, the model trained on clean data was more likely to withhold positive outcomes from deserving candidates in the disadvantaged group (worsens EO), thereby creating a positive rate disparity (worsens DP), whereas for mislabel repair, the model is more likely to incorrectly distribute positive outcomes to undeserving candidates in the privileged group (worsens PP), thereby once again creating a positive rate disparity (worsens DP).

VI. DEEP DIVE

These results motivate us to look at the impact of automated cleaning on a more granular level.

For which cases (dataset, error and fairness metric) is cleaning potentially beneficial at all? In order to assess whether it would be possible to carefully choose a beneficial cleaning technique for a given setting, we analyse for which of the cases in our study we encounter a beneficial auto-cleaning technique at all. We define a case as a combination of a fairness metric — predictive parity (PP), equal opportunity (EO) or demographic parity (DP), a dataset with a single sensitive attribute, and an error type (missing values, outliers or label errors), resulting in 60 different cases in total. A promising finding is that for most cases (53 out of 60), we encounter at least one cleaning technique which does not worsen fairness. In half of the cases (30 out of 60), there exists a cleaning technique which improves fairness, while we can improve both fairness and accuracy simultaneously only in 17 out of 60 cases.

Which repair and detection techniques produce the most gains? Next, we focus on configurations with a positive impact on fairness, and analyze the applied detection and repair techniques in such cases.

For missing values, we do not encounter a dominating imputation approach for numerical columns. However, for categorical columns, "dummy" imputation with a constant value turns out to be most beneficial for fairness (with fairness improvements in 35 cases, compared to 18 cases with a different imputation technique). We attribute this to the fact that dummy imputation allows the model to identify tuples with missing values and learn extra parameters for them (which is not the case for mode and mean imputation). For example, in the `folk` dataset, the accompanying datasheet makes it clear that missing values are typically 'Not Applicable (N/A)', based on values in another column; e.g., Occupation (OCCP) and Class of Worker (COW) are missing for people with Age (AGEP) less than 18. In this case, the missing value is actually a special N/A value, and dummy imputation allows the model to learn such a dependency.

For outlier-repair, which has the worst impact on both fairness and accuracy, we observe no noticeable differences between the repair techniques. However, we find slight differences when analyzing the detection techniques. Cleaning

TABLE XX
SINGLE-ATTRIBUTE ANALYSIS: IMPACT OF AUTO-CLEANING ON ACCURACY AND FAIRNESS FOR DIFFERENT ML MODELS ON 318 CONFIGURATIONS IN TOTAL. WE LIST CASES WHERE FAIRNESS GETS WORSE, FAIRNESS GETS BETTER, AND WHERE BOTH FAIRNESS AND ACCURACY GET BETTER. AUTO-CLEANING IS MORE LIKELY TO WORSEN THAN TO IMPROVE FAIRNESS ACROSS ALL MODELS.

	auto-cleaning makes		
model	fairness worse	fairness better	fairness & accuracy better
xgboost	24.8% (79)	13.8% (44)	4.7% (15)
knn	32.1% (102)	13.5% (43)	7.9% (25)
log-reg	24.5% (78)	15.7% (50)	6.9% (22)

outliers detected via the interquartile rule (outliers-iqr) has a negative impact on fairness in 41.66% of the cases (compared to 28.33% for detection with the standard deviation rule (outliers-sd) and 35% for detection with an isolation forest (outliers-if)). This likely due to the high fraction of records wrongly flagged as outliers, which we already encountered in Figure 1 of Section III.

Impact on single-attribute groups compared to intersectional groups. We summarise our findings from the previous sections, regarding the impact of automated data cleaning on fairness for single-attribute and intersectional group definitions. As discussed in Section III, we observe a comparable fraction of detected errors for single-attribute and intersectional groups, but could not conclusively support or refute the hypothesis that data from historically marginalised groups is more likely to be erroneous. When zooming in on different automated repair methods and fairness metrics (Section V), we find varying effects between the group definitions. For missing value imputation, the direction of the impact on fairness is different per metric for single-attribute groups, but the same for intersectional groups. Automated outlier repair has an insignificant effect on fairness and a negative effect on accuracy, independent of the group definition. Label repairs have a strong effect on both accuracy and fairness, and the direction of the impact on fairness is dependent on the fairness metric, but independent of the group definition, even though we observe a stronger effect for intersectional groups. We conclude that it is important to consider both single-attribute and intersectional group definitions when analysing the impact of data cleaning on accuracy, fairness, and the trade-offs between them.

Model choice. We also investigate the influence of the choice of ML model on the impact on fairness and accuracy. The highest accuracy over all tasks is provided by the logistic regression model (log-reg). It is only outperformed by gradient-boosted decision trees (xgboost) for outliers on folk and heart, as well as for missing values on adult and folk. Apart from that, we find that all models perform comparably with respect to the impact of auto-cleaning on the fairness of their predictions (Table XX). In the majority of cases, this impact is insignificant, however, if there is an impact, auto-cleaning is more likely to worsen (between 24.8% to 32.1% of the cases) than to improve fairness (between 13.5% to 15.7% of the cases).

Logistic regression (log-reg) turns out to benefit most from cleaning in our study, with the largest benefit in fairness (15.7%) and a competitive gain in fairness & accuracy (6.9%), while xgboost benefits least from cleaning in the most desirable setting (fairness and accuracy improve in only 4.7% of cases). knn benefits most from auto-cleaning but does not outperform the other models in terms of accuracy in any configuration.

VII. VISION: FAIRNESS-AWARE DATA CLEANING

The analysis we conducted in this paper is difficult, primarily because it requires that we think holistically about disparities in data quality, disparities in the effectiveness of data cleaning methods, and impacts of such disparities on ML model performance for different demographic groups. Such holistic analysis can and should be supported by data engineering tools, but it requires substantial future research. To detect disparities in data quality, and mitigate the impact of such disparities on the performance of ML models downstream, we envision the development of fairness-aware data cleaning methods and their integration into complex data-intensive pipelines.

Implications for ML in production. While we did notice that historically disadvantaged groups are subject to higher rates of missing values in the majority of cases, we did not find sufficient evidence of a demographic dependency in data errors. This is counter-intuitive to a socio-technical framing, which posits that marginalised groups also appear noisier in the data (have more data errors), and could embolden data scientists to not worry about disparate effects along demographic lines when applying automated cleaning procedures.

However, our second result about the downstream effect of automated cleaning demonstrates that repairing data errors does, in fact, distribute gains disparately across demographic groups! In Section III, we found that automated data cleaning can have a negative impact on fairness, and was, in our study, more likely to worsen fairness than to improve it. Furthermore, we showed that the positive or negative impact of a particular cleaning technique depends on the choice of fairness metric and group definition. These findings are extremely worrying, due to the potential negative impact on the fairness of decisions made by many ML systems that are already in production.

The good news is, however, that we encountered at least one configuration for almost every case (dataset, error type, cleaning method, fairness metric) that did not negatively impact the fairness of model predictions. This indicates that we can — and should — mitigate any potential negative impact of automated cleaning with the help of a principled methodology for selecting an appropriate cleaning procedure. Our results underscore the importance of such a methodology, and motivate its development.

Open questions and research directions. Our findings indicate that we are either unable to detect demographically-salient data errors with current approaches, or that current cleaning procedures are not equally 'effective' for different demographic groups, or — most disturbingly — we are seeing failure modes in both detection and repair. In order to confirm

whether the disparate proportions of tuples flagged by the error detection strategies in Section III correspond to actual errors, one would need to repeat this analysis on a dirty fairness-critical dataset where the clean ground truth is available. Thus future work on fairness-aware data cleaning must include additional empirical evaluation.

Our findings from the study in Section V impose the immediate question of how to choose a cleaning technique that does not negatively impact fairness. Additionally, we consider it important future work to analyse whether more advanced cleaning/error detection techniques [11], [35], [36], which leverage additional metadata, impact fairness in a similar way (note that we had to exclude them from our study due to a lack of clean example tuples and integrity constraints). Connected to this, an important long-term research question in fairness-aware data cleaning is whether it will be sufficient to appropriately choose from existing cleaning techniques or whether we would need new fairness-aware cleaning procedures. The selection of cleaning techniques and model hyperparameters is typically steered by cross-validation techniques which aim for the highest accuracy. A promising direction here might be to extend existing techniques and implementations to adhere to fairness constraints during the selection procedure. A starting point for designing new cleaning techniques is the identification of input tuples with negative impact on fairness, which would then need to be cleaned in a fairness-enhancing manner. Several techniques for identifying such tuples have recently been proposed, e.g., by computing Shapley values with respect to a given fairness metric [38] or via causal explanations [13].

A further direction is to focus on the impact of label errors [32] on fairness, for example by adding instance-dependent synthetic label noise [39] and evaluating the corresponding cleaning quality of various label repair techniques. Moreover, it would be interesting to explore whether our observed results also hold for other notions of fairness, such as individual fairness [40], where the goal is to have similar individuals treated similarly, and causal fairness [41], where one of the questions is whether individuals are being treated differently on the basis of their membership in a historically-disadvantaged group.

Finally, a limitation of our study is that we mainly worked with US-centric datasets (which are common in fairness research). This limitation should be overcome in future work on fairness-aware data cleaning.

Acknowledgements. *This work was supported by the NSF Awards No. 1916505, 1922658 and 2312930, and by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.*

REFERENCES

[1] J. Stoyanovich, S. Abiteboul, B. Howe, H. V. Jagadish, and S. Schelter, "Responsible data management," *Commun. ACM*, vol. 65, no. 6, pp. 64–74, 2022. [Online]. Available: https://doi.org/10.1145/3488717

[2] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: a survey," *SIGMOD Record*, vol. 47, no. 2, 2018.

[3] J. M. Hellerstein, "Quantitative data cleaning for large databases," *UNECE*, vol. 25, 2008.

[4] I. F. Ilyas and F. Naumann, "Data errors: Symptoms, causes and origins," *IEEE Data Eng. Bull.*, p. 4, 2022.

[5] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "Cleanml: A benchmark for joint data cleaning and machine learning," *ICDE*, 2019.

[6] I. F. Ilyas and T. Rekatsinas, "Machine learning and data cleaning: Which serves the other?" *JDIQ*, 2020.

[7] S. Schelter, Y. He, J. Khilnani, and J. Stoyanovich, "Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions," *EDBT*, 2019.

[8] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" *NeurIPS*, 2018.

[9] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu, "From cleaning before ml to cleaning for ml." *IEEE Data Eng. Bull.*, vol. 44, no. 1, 2021.

[10] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "Activeclean: Interactive data cleaning for statistical modeling," *PVLDB*, vol. 9, no. 12, 2016.

[11] B. Karlaš, P. Li, R. Wu, N. M. Gürel, X. Chu, W. Wu, and C. Zhang, "Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions," *VLDB*, 2020.

[12] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *FACT**, 2019.

[13] R. Pradhan, J. Zhu, B. Glavic, and B. Salimi, "Interpretable data-based explanations for fairness debugging," *SIGMOD*, 2022.

[14] A. Asudeh, Z. Jin, and H. Jagadish, "Assessing and remedying coverage for a given dataset," *ICDE*, 2019.

[15] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," *NeurIPS*, 2021.

[16] M. T. Islam, A. Fariha, A. Meliou, and B. Salimi, "Through the data management lens: Experimental analysis and evaluation of fair classification," *SIGMOD*, 2022.

[17] Z. Liu, Z. Zhou, and T. Rekatsinas, "Picket: Self-supervised data diagnostics for ml pipelines," *VLDBJ*, 2020.

[18] S. Caton, S. Malisetty, and C. Haas, "Impact of imputation strategies on fairness in machine learning," *Journal of Artificial Intelligence Research*, vol. 74, 2022.

[19] California Privacy Protection Agency, "California consumer privacy act - frequently asked questions." [Online]. Available: https://cppa.ca.gov/faq.html#faq_res_1

[20] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[21] House of Commons of Canada, "Digital charter implementation act: An act to enact the consumer privacy protection act, the personal information and data protection tribunal act and the artificial intelligence and data act and to make consequential and related amendments to other acts," 2022. [Online]. Available: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading

[22] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Tackling documentation debt: A survey on algorithmic fairness datasets," ser. EAAMO '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3551624.3555286

[23] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *NeurIPS*, 2021.

[24] Federal Trade Commission, "Protections Against Discrimination and Other Prohibited Practices," https://www.ftc.gov/policy-notices/no-fear-act/protections-against-discrimination.

[25] European Commission, "What the European Commission is doing to protect your rights," https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en#what-the-european-commission-is-doing-to-protect-your-rights.

[26] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[27] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," pp. 139–167, 1989.

[28] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[29] A. Narayanan, "Fairness definitions and their politics," *ACM FaccT*, 2018.

[30] Z. Abedjan, X. Chu, D. Deng, R. C. Fernandez, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang, "Detecting data errors: Where are we and what needs to be done?" *PVLDB*, vol. 9, no. 12, 2016.

[31] M. Mahdavi, Z. Abedjan, R. Castro Fernandez, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang, "Raha: A configuration-free error detection system," *SIGMOD*, 2019.

[32] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," *NeurIPS*, 2021.

[33] C. G. Northcutt, T. Wu, and I. L. Chuang, "Learning with confident examples: Rank pruning for robust classification with noisy labels," *UAI*, 2017.

[34] X. Chu, I. F. Ilyas, and P. Papotti, "Discovering denial constraints," *PVLDB*, vol. 6, no. 13, 2013.

[35] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: Holistic data repairs with probabilistic inference," *VLDB*, 2017.

[36] A. Heidari, J. McGrath, I. F. Ilyas, and T. Rekatsinas, "Holodetect: Few-shot learning for error detection," *SIGMOD*, 2019.

[37] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. J. Spanos, and D. Song, "Efficient task-specific data valuation for nearest neighbor algorithms," *VLDB*, 2019.

[38] B. Karlaš, D. Dao, M. Interlandi, B. Li, S. Schelter, W. Wu, and C. Zhang, "Data debugging with shapley importance over end-to-end machine learning pipelines," 2022. [Online]. Available: https://arxiv.org/abs/2204.11131

[39] S. Wu, M. Gong, B. Han, Y. Liu, and T. Liu, "Fair classification with instance-dependent label noise," in *Proceedings of the First Conference on Causal Learning and Reasoning*, ser. Proceedings of Machine Learning Research, B. Schölkopf, C. Uhler, and K. Zhang, Eds., vol. 177. PMLR, 11–13 Apr 2022, pp. 927–943. [Online]. Available: https://proceedings.mlr.press/v177/wu22b.html

[40] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[41] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *The World Wide Web Conference*, 2019, pp. 2907–2914.