JUSTFAIR
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

This dataset was created to increase the ease of access to data on public criminal trials. Specifically, it combines information from the US sentencing commission, Federal judicial center, PACER, wikipedia, and Federal Judicial Center biographies to create a dataset of defendants and their demographic characteristics with information about their crimes, their sentences, and the identity of the sentencing judge [1].

I-B Who created the dataset?

Is it an official law enforcement or government body? An academic research team? Other?

The dataset was created by researchers at the Institute for the Quantitative Study of Inclusion, Diversity, and Equity (QSIDE): Maria-Veronica Ciocanel, Chad Topaz, Rebecca Santorella, Shilad Sen, Christian Smith, and Adam Hufstetler.

I-C Was there a specific task in mind, or gap that needed to be filled?

The authors wished to determine whether judges with significant sentencing outcome disparities across the race of defendants were due to racial bias by creating a dataset with sufficient controls, such as defendant's education level or age.

II. Composition

II-A What do the instances that comprise the dataset represent?

For example: crimes, offenders, court cases, police officers

Each instance in this dataset corresponds to a single criminal trial sentencing.

II-B Are there multiple types of instances?

For example: offenders, victims, and the relationship between them.

No.

II-C How many instances are there in total?

Of each type, if appropriate.

There are a total of 595,850 sentences in the dataset.

II-D Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

For example, if it is traffic stops from a territory, is it all traffic stops conducted within that territory within a specific time? If not, is it a representative sample of all stops? Describe how representativeness was validated/verified. If it is not representative, please describe why.

The United States Sentencing Commission maintains publicly accessible data sets, including files which provide information about sentences given to individuals in federal district courts. This dataset contained a subset of these which have been successfully linked with the data from other datasets.

II-E What data does each instance consist of?

If there is a large number of variables, please provide a broad description of what is included.

The JUSTFAIR dataset is a result of linking together five datasets and as a result, it contains many variables. A high-level overview of the available information can be in Figure 1.

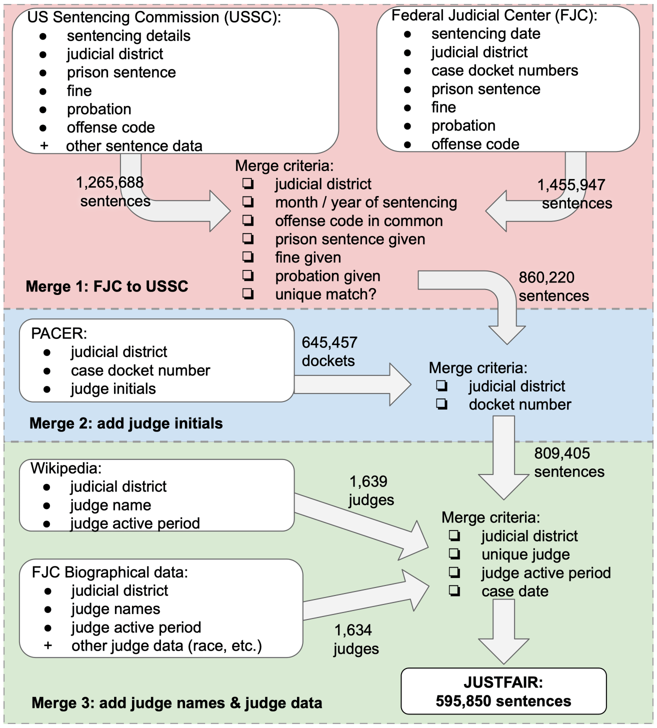

Fig. 1. An illustration of how the five datasets are linked together to form the JUSTFAIR dataset. Reprinted from [1].

Broadly, the information contains in the dataset consists of:

- Sentencing Details
- Prison sentence
- Fine
- Probation
- Offense
- Judicial District
- Judge Name
- Defendant demographics
- Judge demographics

II-F Is there a target label or associated with each instance?

Please include labels that are likely to be used as target labels, e.g. recidivism.

No. However, different aspects of sentencing, may be suitable to be used as target labels.

II-G Are there recommended data splits (e.g., training, development/validation, testing)?

If so, please provide a description of these splits, explaining the rationale behind them.

No.

II-H Does the dataset contain data on race and ethnicity?

If so, is it based on the individual's self-description, or based on officer's impression? Was it collected or derived in post-processing? For example, by name analysis.

Yes. The dataset contains defendant demographic information, as well as demographic information on the judge .

II-I Are there any known errors, sources of noise, bias or missing data, or variables collected for only part of the datasets?

If so, please provide a description.

JUSTFAIR uses a combination of Wikipedia and Federal Judicial Center biographical data to match the judge to the sentencing details. Based on 159 manual test cases, this matching has 2% error rate.

II-J Does the dataset contain data on criminal history or other data that might be considered confidential or sensitive in any way?

For example: sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; If so, please provide a description.

The dataset contains sentencing data on individuals which may be considered confidential.

II-K Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

If so, please describe how.

Yes, the dataset purposely identifies the judges involved in the cases. In addition, the defendant's name also appears in the data.

III. Uses

III-A What type of tasks, if any, has the dataset been used for?

If so, please provide examples and include citations.

The dataset has been used to study racial bias in the judicial system [1], [2], [3], [4].

III-B Is there a repository that links to any or all papers or systems that use the dataset?

If so, please provide a link or other access point.

No.

III-C What (other) tasks could the dataset be used for?

For example: testing predictive policing systems, predicting recidivism.

This dataset could be used to study research questions around sentencing and probation in the relevant courts.

III-D Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Results will have to be analyzed carefully due to the non-negligible error rate in judge-case matching.

IV. Collection Process

IV-A How was the data associated with each instance acquired?

e.g. the data collected survey, the raw data is routinely collected by the courts.

The dataset was created by merging five datasets:

1) the United States Sentencing Commission Database
2) the Federal Judicial Center Integrated Database
3) the Public Access to Court Electronic Records system
4) Wikipedia
5) the Federal Judicial Center Biographical Directory of Article III Federal Judges

IV-B Was the information self-reported?

If the data was self-reported, was the data validated/verified? If so, please describe how.

No. However, some of the information in the individual datasets, e.g., Wikipedia, may be self-reported.

IV-C Who was involved in the data collection process?

Was this done as part of their other duties? If not, were they compensated?

The data was collated by the QSIDE institute.

IV-D Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

If not, please describe the timeframe in which the data associated with the instances was created. If the collection was not continuous within the timeframe, please specify the intervals, for example, annually, every 4 years, irregularly.

The dataset was published October 26, 2020, and covers the years 2001 — 2018.

IV-E Were any ethical review processes conducted (e.g., by an institutional review board)?

If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

An ethical review is not mentioned by the authors [1].

IV-F Were the individuals in question notified about the data collection? Did they give their consent?

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

No.

IV-G Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

An impact analysis is not mentioned by the authors [1].

V. Pre-processing, cleaning, labeling

V-A Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, removal of instances, processing of missing values)?

If so, please provide a description and reference to the documentation. If not, you may skip the remaining questions in this section.

The pre-processing performed on each of the five datasets can be found in [1].

V-B Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?

If so, please provide a link or other access point to the "raw" data.

The raw data is available on the respective websites of each of five datasets.

V-C Is the software that was used to preprocess/clean/label the data available?

If so, please provide a link or other access point.

No.

VI. Distribution

VI-A Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?

Does the dataset have a digital object identifier (DOI)?

Yes, the dataset can be downloaded at: https://qsideinstitute.org/research/criminal-justice/justfair/

VI-B Is the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No license has been specified.

VII. Maintenance

VII-A Is the dataset maintained? Who is supporting/hosting/maintaining the dataset?

The dataset will be updated, conditional on the creators obtaining additional funding.

VII-B How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The QSIDE Institute can be contacted at: qside@qsideinstitute.org

VII-C Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes, conditional on the creators obtaining additional funding.

VII-D Are older versions of the dataset continue to be supported/hosted/maintained?

No.

VII-E If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

If so, please provide a description.

Try contacting qside@qsideinstitute.org.

References

[1] Maria-Veronica Ciocanel, Chad M Topaz, Rebecca Santorella, Shilad Sen, Christian Michael Smith, and Adam Hufstetler. Justfair: Judicial System Transparency Through Federal Archive Inferred Records. *Plos one*, 15(10):e0241381, 2020.

[2] Christian Michael Smith, Nicholas Goldrosen, Maria-Veronica Ciocanel, Rebecca Santorella, Chad M Topaz, and Shilad Sen. The Most Discriminatory Federal Judges Give Black and Hispanic Defendants at Least Double the Sentences of White Defendants. 2021.

[3] Jackson Sargent and Melanie Weber. Identifying Biases in Legal Data: An Algorithmic Fairness Perspective. *arXiv preprint arXiv:2109.09946*, 2021.

[4] Mikaela Meyer, Aaron Horowitz, Erica Marshall, and Kristian Lum. Flipping the Script on Criminal Justice Risk Assessment: An Actuarial Model for Assessing the Risk the Federal Sentencing System Poses to Defendants. *arXiv preprint arXiv:2205.13505*, 2022.