# F Datasheet

## F.1 Motivation

**For what purpose was the dataset created?** The raw case records were created presumably for public inspection and for record keeping in the courts.

The external census dataset used in curation was created for census.

The final curated dataset, contributed in this paper, was created for academic research on algorithmic fairness.

**Who created the dataset, and on behalf of which entity?** According to Wisconsin Circuit Courts Access website, the raw data records that we accessed through their website are an exact copy of the case information entered into the circuit court case management system by court staff in the counties where the case files are located.

Nianyun Li, Claudia Marangon and Peiyao Sun curated the dataset in its current form with the help of Elliott Ash and Naman Goel.

**Who funded the creation of the dataset?** The funding of creation of original data records is unclear, but presumably the state funded the court staff in the counties where the case files are located.

The creation of the curated dataset was funded by ETH Zurich.

## F.2 Composition

**What do the instances that comprise the dataset represent?** The instances represent case and defendant information.

**How many instances are there in total (of each type)?** There are around 1.5 million instances.

**Does the dataset contain all possible instances or is it a sample of instances from a larger set?** The dataset is a sample of instances from a larger set.

**What data does each instance consist of?** Each instance contains defendant's new_id, age, sex, type of offense, wcisclass, year of filing, race, age at judgment, age at first offense, 9 neighborhood characteristics (population density, proportion who attended college, proportion eligible for food stamp, African American population share, Hispanic population share, proportion who live in rural and urban area, median household income, highest charge severity, not_detained, probation, recid (180 days sentence length cutoff), recid (2 years sentence length cutoff), violent_recid (180 days sentence length cutoff), jail, county, violent_crime. The details of each of these variables are provided in the accompanying paper and the metadata file in the data directory. Each instance also contains prior criminal count for each type of offense, prior sentence length statistics, prior charge severity counts. The counts were created from the **available** raw case records by performing database search, and are therefore possibly underestimated.

**Is there a label or target associated with each instance?** For research, recid (180 days sentence length cutoff) and recid (2 years sentence length cutoff) are two variables associated with each instance. There is also a violent_recid (180 days sentence length cutoff) variable available for research. These are not ground truth labels in a traditional sense but only variables defined by the authors. Details in Section 3.5and ethics discussion in Section 5.3 of the accompanying paper.The first variable is recidivism as observed in the case records by performing database search, within a 2 year follow-up period since judgment disposition date, using a 180 days cut-off for sentence. The second is obtained by using a 2 year cut-off sentence and extending the follow-up period of 2 years by adding the sentence length.

**Is any information missing from individual instances?** Yes, the recidivism variables can not be observed for defendants depending on their sentence. Thus, it is missing for some defendants. Details in Section 3.5.

**Are relationships between individual instances made explicit?** The dataset is anonymized and therefore some relationship between individual instances may be lost. We have included defendant pseudo-identifier in the dataset constructed based on first name, last name and date of birth.

**Are there recommended data splits?** No. But we have included two possible random splits in the dataset (one is completely random thus only ensuring different cases in train and test splits, the other also ensures different defendants in train and test splits).

**Are there any errors, sources of noise, or redundancies in the dataset?** The errors are possible in the raw case information entered by the court staff. Known errors in the curated dataset construction are discussed in Section 5.2.

**Is the dataset self-contained, or does it rely on external resources?** The dataset has been curated using case records from WCCA and census data from 2010. The curated dataset is self-contained.

**Does the dataset contain data that might be considered confidential?** No

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, or threatening?** No

**Does the dataset relate to people?** Yes, the raw case records relate to the defendants. However, directly identifiable information such as names, addresses, date of birth, case numbers etc has been removed in the curated dataset.

**Does the dataset identify any subpopulations?** Yes, the dataset contains data from five racial groups as marked by WCCA, sex and age groups.

**Is it possible to identify individuals?** The raw case records available on WCCA are public information and it is possible to identify individuals there. For the curated dataset that we release, we have removed directly identifiable information such as names, addresses, date of birth, case numbers etc.

**Does the dataset contain data that might be considered sensitive in any way?** The raw case records available on WCCA are public information and some of the information such as defendant's personal information may be considered sensitive. For the curated dataset that we release, we have removed directly identifiable information such as names, addresses, date of birth, case numbers etc.

### F.3 Collection Process

**How was the data associated with each instance acquired?** Through the REST interface of WCCA.

**What mechanisms or procedures were used to collect the data?** Through the REST interface of WCCA. We queried all case numbers in each county during the period and subsequently, using these case numbers, we queried individual cases for all available information. Different attributes were then derived using the process described in the accompanying paper.

**If the data are a sample from a larger set, what was the sampling strategy?** The dataset is composed primarily of new cases filed between 2000-2018. The dataset excludes dismissed cases that do not result in conviction, records of defendants that do not have sex and/or race data and cases that only have forfeiture (non-crime) charge. `https://wcca.wicourts.gov/faq.html` provides more information on records that might have been deleted. WCCA also informed us of a few limitations of the data as part of the subscription. These are listed as follows:

1. WCCA Information includes only court records open to public view under Wisconsin's Open Records Law, Wis. Stat. 19.31-19.39. Court records not open to public inspection by law are not available.
2. WCCA Information does not include information that may be confidential, sealed, or redacted in accordance with all applicable statutes, court orders, and rules related to confidentiality, sealing, and redaction.
3. WCCA Information consists of information entered into the CCAP [3] case management system by the Clerk of Circuit Court or Register in Probate in each county. CCAP is not responsible for the accuracy or timeliness of WCCA information.
4. WCCA Information does not comprise the complete court record. Copies of documents must be obtained from the Clerk of Circuit Court or Register in Probate.
5. WCCA Information is only a snapshot of the information accessible in the CCAP case management system on the date the information is downloaded by the Subscriber.

[3] WCCA was formerly CCAP.

6. WCCA Information is not the Judgment and Lien Docket under Wis. Stat. 806.10. The Judgment and Lien Docket is available from the Clerk of Circuit Court.
7. Court records which predate the implementation of the CCAP case management system in the county in which the records were created are not accessible under this Agreement, except to the extent such records have been back loaded.
8. In criminal cases, any designation in any race field contains subjective information generally provided by the agency that filed the case.
9. Searching WCCA Information by a particular field or code may not return all cases in which a particular event occurred unless at the time the record was created the case management system required the field or code to be completed in order to proceed to make the rest of the record

**Who was involved in the data collection process and how were they compensated?** The case records were created by court staff in respective county courts and were presumably, compensated by state.

The authors of this paper collected the case records from WCCA and were employees of ETH Zurich during the data curation process. They were compensated by ETH Zurich in the form of fixed monthly salaries. Nianyun Li, Naman Goel, Peiyao Sun, Claudia Marangon were/are on fixed-term contracts with ETH Zurich.

**Over what time frame was the data collected?** Authors had access to the raw case records through WCCA REST interface during the period July 2020 - July 2021. However, data collection was finished by Feb 2021.

**Were any ethical review processes conducted?** The authors are not aware of the ethical review process followed in WCCA or county courts for creation of the case records. As part of the curation of the dataset that we contribute, no formal/institutional ethical review process was conducted.

**Does the dataset relate to people?** Yes, the raw case records relate to defendants. However, directly identifiable information such as names, addresses, date of birth, case numbers etc has been removed in the curated dataset.

**Did you collect the data from the individuals directly, or obtain it via third parties?** We obtained the raw case records from a third party, WCCA (`https://wcca.wicourts.gov`).

**Were the individuals notified about the data collection?** The authors are not aware of it. However, presumably, the defendants were aware that the information about their cases (and hence the related information about them) is kept in court records and is public information under Wisconsin state laws, when exceptions do not apply.

**Did the individuals in question consent to the collection and use of their data?** The authors are not aware of it. The information is available publicly under Wisconsin state laws.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** WCCA provides option in certain limited cases to petition to have case-records removed. Please see `https://wcca.wicourts.gov/faq.html`.

For the curated dataset, directly identifiable information such as names, addresses, date of birth, case numbers etc has already been removed.

**Has analysis of the potential impact of the dataset and its use on data subjects been conducted?** Authors are not aware if Wisconsin state or WCCA had done any such analysis.

For the curated dataset, we have thought carefully about it and redacted all the information that, according to us, could potentially affect subjects.

### F.4 Pre-processing and Cleaning

**Was any preprocessing of the data done?** The accompanying paper describes the details of curating the dataset from the raw case records.

**Was the "raw" data saved in addition to the cleaned data?** Yes, we saved the data on our institute's secure servers (until our research requires). We do not plan to make this data available to others.

**Is the software used to clean the data available?** We didn't use any 'data cleaning software'. We have described the steps taken in curating the dataset in the accompanying paper. If useful, we can also provide SQL commands for these steps.

### F.5 Uses

**Has the dataset been used for any tasks already?** The curated dataset that we release has only been used for academic research. We are not aware who else has used the raw case records from WCCA and for which tasks.

**Is there a repository that links to any or all papers that use the dataset?** Not to our knowledge.

**What (other) tasks could the dataset be used for?** The dataset is for academic research.

**Is there anything about the composition of the dataset or the way it was collected and cleaned that might impact future uses?** We have listed limitations of the data in Section 5.2 and in earlier parts of this datasheet (for example, see Section F.3).

**Are there tasks for which the dataset should not be used?** The dataset should not be used for purposes other than academic research.

### F.6 Distribution

**Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** Yes, public data.

**How will the dataset be distributed?** The dataset is hosted at `http://clezdata.github.io/wcld/`. Downloads are subject to research only use acknowledgement. In case of any difficulties in accessing the data in the future, interested readers can contact the authors.

**When will the dataset be distributed?** The dataset is hosted at `http://clezdata.github.io/wcld/`.

**Will the dataset be distributed under a copyright, other IP license, or terms of use?** The dataset is distributed under the Creative Commons 4.0 BY-NC-SA license.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.

**Do any export controls or other regulatory restrictions apply to the data?** No.

### F.7 Maintenance

**Who is supporting/hosting/maintaining the dataset?** Elliott Ash.

**How can the data owner/curator be contacted?** Through email: elliott.ash@gess.ethz.ch

**Is there an erratum?** Not at the time of publishing this paper.

**Will the dataset be updated?** The existing entries in the dataset are unlikely to be modified. New information may be added.

**If the dataset relates to people, are there applicable limits on the retention of data associated with the instances?** Public information under applicable law.

**Will older versions of the dataset continue to be supported/hosted/maintained?** In the unlikely event that the entries in the dataset are to be modified, older version will also be made available, for example, using a version control system.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Individuals interested in contributing are encouraged to contact Elliott Ash at elliott.ash@gess.ethz.ch.

## G Additional Information

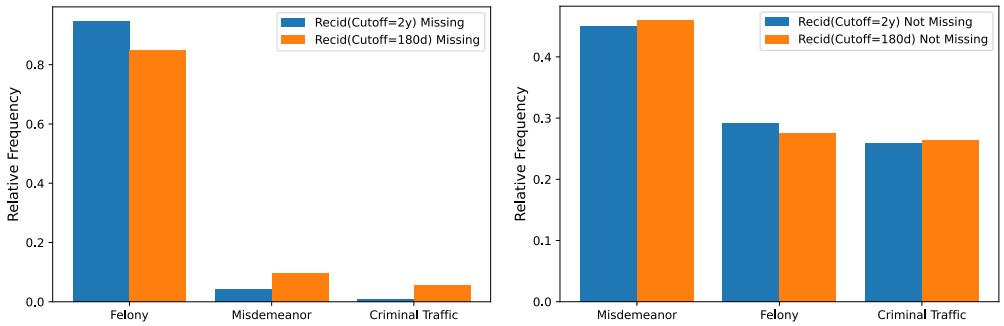
Table 5 shows the mappings from charge severity categories on WCCA and the numerical ranking that we assigned to these categories. Values 1-6 were assigned for forfeiture charges and hence, not shown in the table. Table 6 provides summary statistics for the column charge severity.

Table 5: Mapping from Charge Severity to Numerical Ranking

| Charge | `highest_charge_severity` |
|---|---|
| Felony A | 21 |
| Felony B | 20 |
| Felony BC | 19 |
| Felony C | 18 |
| Felony D | 17 |
| Felony E | 16 |
| Felony F | 15 |
| Felony G | 14 |
| Felony H | 13 |
| Felony I | 12 |
| Felony U | 11 |
| Misdemeanor A | 10 |
| Misdemeanor B | 9 |
| Misdemeanor C | 8 |
| Misdemeanor U | 7 |

Table 6: Highest Charge Severity in the Dataset

| `highest_charge_severity` | Count | Percentage |
|---|---|---|
| 7 | 516004 | 34.94 |
| 10 | 460898 | 31.21 |
| 9 | 145750 | 9.87 |
| 13 | 116008 | 7.85 |
| 12 | 74062 | 5.01 |
| 15 | 37580 | 2.54 |
| 14 | 30561 | 2.07 |
| 18 | 26678 | 1.81 |
| 11 | 21893 | 1.48 |
| 16 | 21217 | 1.44 |
| 17 | 17088 | 1.16 |
| 20 | 6187 | 0.42 |
| 19 | 1581 | 0.11 |
| 21 | 845 | 0.06 |
| 8 | 615 | 0.04 |




(a) Missing Recidivism Outcome

(b) Not-Missing Recidivism Outcome

Figure 2: Differences in the distribution of type of offense depending on the sentence cutoff length in recidivism variables.

Table 7 shows results of the machine learning classifiers for the recidivism variable with 2 years sentence cut-off.

Table 7: Recidivism (2 year Sentence Cut-Off) Prediction with 95% Confidence Intervals

| | Overall | Caucasian | African American |
|---|---|---|---|
| *XGBoost* | | | |
| Accuracy | $0.6589 \pm 0.0016$ | $0.6652 \pm 0.0019$ | $0.6460 \pm 0.0038$ |
| AUC | $0.7018 \pm 0.0018$ | $0.7026 \pm 0.0021$ | $0.7005 \pm 0.0034$ |
| FPR | $0.2177 \pm 0.0029$ | $0.2094 \pm 0.0029$ | $0.2353 \pm 0.0046$ |
| FNR | $0.5136 \pm 0.0043$ | $0.5232 \pm 0.0049$ | $0.4970 \pm 0.0064$ |
| PR | $0.3298 \pm 0.0030$ | $0.3163 \pm 0.0032$ | $0.3567 \pm 0.0039$ |
| *Logistic Regression* | | | |
| Accuracy | $0.6457 \pm 0.0014$ | $0.6555 \pm 0.0016$ | $0.6215 \pm 0.0029$ |
| AUC | $0.6784 \pm 0.0015$ | $0.6794 \pm 0.0019$ | $0.6760 \pm 0.0034$ |
| FPR | $0.1478 \pm 0.0022$ | $0.1430 \pm 0.0023$ | $0.1571 \pm 0.0026$ |
| FNR | $0.6429 \pm 0.0024$ | $0.6469 \pm 0.0030$ | $0.6453 \pm 0.0047$ |
| PR | $0.2351 \pm 0.0018$ | $0.2270 \pm 0.0020$ | $0.2467 \pm 0.0028$ |
| | Hispanic | Native American | Asian |
| *XGBoost* | | | |
| Accuracy | $0.6576 \pm 0.0054$ | $0.6273 \pm 0.0064$ | $0.6733 \pm 0.0149$ |
| AUC | $0.6724 \pm 0.0056$ | $0.6852 \pm 0.0066$ | $0.7016 \pm 0.0182$ |
| FPR | $0.1997 \pm 0.0079$ | $0.3223 \pm 0.0123$ | $0.2153 \pm 0.0130$ |
| FNR | $0.5720 \pm 0.0097$ | $0.4125 \pm 0.0105$ | $0.5140 \pm 0.0245$ |
| PR | $0.2872 \pm 0.0076$ | $0.4706 \pm 0.0091$ | $0.3163 \pm 0.0116$ |
| *Logistic Regression* | | | |
| Accuracy | $0.6540 \pm 0.0054$ | $0.5988 \pm 0.0050$ | $0.6728 \pm 0.0152$ |
| AUC | $0.6533 \pm 0.0061$ | $0.6699 \pm 0.0070$ | $0.6878 \pm 0.0176$ |
| FPR | $0.1293 \pm 0.0058$ | $0.2358 \pm 0.0090$ | $0.1339 \pm 0.0113$ |
| FNR | $0.6948 \pm 0.0087$ | $0.5316 \pm 0.0082$ | $0.6521 \pm 0.0222$ |
| PR | $0.1967 \pm 0.0055$ | $0.3658 \pm 0.0069$ | $0.2138 \pm 0.0096$ |

## H Violent/Non-Violent Labels for Charge Descriptions using GPT-4

The prompt used for GPT was as follows:

"In the FBI's Uniform Crime Reporting (UCR) Program, violent crime is composed of four offenses: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. Violent crimes are defined in the UCR Program as those offenses which involve force or threat of force.
I will next provide you a list of charge descriptions. For each charge description in the list, I want you to provide me a single word answer (Violent/Non-Violent), depending on whether that charge description refers to a violent crime or not. Before providing the answer, I want you to provide an explanation or thought process of how you go from charge description to the answer. The format of your response should be Charge Description;;Thought Process;;Violent/Non-Violent. Do not include any other text in your response. The charge description in your response should be exactly same as the charge description in my list (do not correct the formatting or spellings etc in charge descriptions) and in the same order as my list.
Here is the list: "

We appended charge descriptions (50 at a time) to the prompt and repeatedly prompted the model until we obtained labels for all the charge descriptions. The model used was 'gpt-4' as on 20-Sep-2023. We set the system message as 'You are a helpful assistant.' The cost of inference was approximately $150, including the costs of trial and error with different prompting and models etc.