Virginia Pre-Trial Data Project
Datasheet

I. Motivation

I-A *For what purpose was the dataset created?*

The Virginia State Crime Commission has been studying various aspects of the pre-trial process since 2016. However, there was a significant lack of data readily available to answer many important questions related to the pre-trial process in the Commonwealth. As a result, the Virginia Pre-Trial Data Project was developed [1].

I-B *Who created the dataset?*

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

The project was lead by the Virginia State Crime Commission. Data was collected from the following agencies: Supreme Court of Virginia, Office of the Executive Secretary; Alexandria Circuit Court; Fairfax County Circuit Court; Virginia Department of Criminal Justice Services; Virginia State Police; Virginia Department of Corrections; Virginia Compensation Board.

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

Following individuals from pre-trial to final dispositions, including assigned risk levels. The Project consisted of two phases: (i) developing a cohort of adult defendants charged with a criminal offense in Virginia during October 2017 and (ii) tracking various outcomes within that cohort [1].

II. Composition

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

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

Each instance reports on a ‘contact event’, defined as: “all charges against a defendant in the same jurisdiction on the same day and having the same CBR number” (CBR stands for “Commit, Bond, Release” and refers to any one of these bail processes) [1].

Each instance includes information regarding the defendant, and the progression of the criminal case from the time a defendant is charged with an offense until the final disposition of the case, i.e., trial or sentencing [1].

If the same individual has more than one contact event during the month of October 2017, only the earlier contact event is reported in the data. If a defendant was charged with multiple offenses on the same day, but the offenses were heard in different courts, those records were grouped by court and reported as separate events [1].

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.

The dataset contains 22,986 adult defendants charged with a criminal offense during a October 2017.

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 cohort includes all defendants charged in Virginia in October 2017. However, some instances were excluded from the follow-up and as a result can not be used for most analysis. The reason for the excluding is reported in the variable ‘Exclude’ and include missing data, the defendant being under 18 when they were charged, and the offense not being punishable by incarceration, amongst other reasons.

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 dataset contains over 700 variables for each defendant. Broadly, these include:

1. Demographics: Sex, Race, Age, Indigency Status, Virginia Residency Status, Zip Code.
2. Pending charges.
3. State or local probation status.
4. October 2017 charge(s): number of offense, offense and offense type (up to 10).
5. Bond: Bond Type and amount at initial contact and at release.
6. Release status: Whether Defendant Was Released During Pre-trial Period, Pre-Trial Release Date, Pretrial Release Type.
7. Whether the defendant received pretrial services agency supervision: supervision Days, conditions. Whether defendant is on state or community supervision.
8. Case: attorney type, court type, court locality, sentence type, imposed and effective sentence, Final Disposition and disposition date.

9. Prior criminal history: age at first adult arrest; number prior arrests for felonies, misdemeanors, and specific crimes, e.g., domestic abuse; number of prior convictions overall, as an adult, in the past 2 and 5 years, for felonies, misdemeanors, and specific crimes, e.g., drug convictions. Prior sentencing for felonies and probation Revocation ; prior probation revocations, number of prior incarceration events for more than 14 days, less and more than a year.
10: Risk level: components to calculate VPRAI [2] and PSA [3] risk levels, and corresponding scores.
11. Court appearance and public safety: details on new failure to appear and offending in the follow up period.
12: Aggregate locality characteristics: locality Name, Region, population estimate, density and demographic (race, ethnicity, sex and age combined) composition, unemployment and education rates, number of law enforcement officers, income, health insurance, citizenship status; incident and arrest rate overall and for specific crimes.

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.

There is not pre-specified target label. However, variables under the court appearance and public safety, e.g., new offending, are particularly suitable to be used as target variables.

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.

Information on race is included. This information is taken from court records. It is unclear if it is based on self-description or not. Ethnicity is only partly recorded in the raw data. As a result, in the dataset Hispanic ethnicity is considered within the White racial category.

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.

Records with missing data were excluded from the follow on analysis. The reason for the exclusion is reported in the variable 'Exclude' (See pg. 270 is [1]).

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 information about criminal history, as well as demographic information.

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.

Unlikely. Only indirectly, and only if the case received significant media attention.

III. Uses

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

If so, please provide examples and include citations.

Yes. Finding from the project can be found in [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 can be used to study risk assessment scores, pre-trial custody status, sentencing and recidivism.

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?

Offender are from a single cohort, october 2017.

Although criminal records were extracted for all defendants in the cohort, the data does not include those records. We have standardized information regarding the criminal history of the defendant, which may not be suitable for all uses.

Hispanic ethnicity within the White racial category.

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 data is a compilation of information and variables provided by numerous state and local government agencies across Virginia:

1. Supreme Court of Virginia, Office of the Executive Secretary: eMagistrate Sytem; Circuit, General District, and Juvenile and Domestic Relations District Court Case Management Systems.
2. Alexandria Circuit Court: Alexandria Circuit Court Case Management System.
3. Fairfax County Circuit Court: Fairfax County Circuit Court Case Management System.
4. Virginia Department of Criminal Justice Services: Pretrial and Community Corrections Case Management System (PTCC).
5. Virginia State Police: Central Criminal Records Exchange (CCRE).
6. Virginia Department of Corrections: Corrections Information System (CORIS).
7. Virginia Compensation Board: Local Inmate Data System (LIDS).

IV-B Was the information self-reported?

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

No. Data was extracted from government record keeping systems. No direct validation of the data has been conducted.

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 collected by paid workers (assumed).

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.

Adult defendants charged with a criminal offense during October 2017 were included. They were tracked until final case disposition or December 31, 2018, whichever came first.

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.

Unknown.

IV-F Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?

The data was collected from government agencies.

IV-G 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?

Individuals likely know their data was entered into the agency's data collection system. It is unlikely they knew or consented the use of the data for research purposes.

IV-H 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.

Unknown.

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.

Details regarding pre-processing can be found in the codebook [1]. Broadly, the defendant's criminal history is not presented in its raw from. For example, we do not see that defendant a was charged with offense A in the year X and so on. Instead, we are told that defendant a has N prior charges from type A.

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.

No.

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 is publicly available on The Virginia State Crime Commission's website:
http://www.vcsc.virginia.gov/pretrialdataproject.htmlhttp://www.vcsc.virgi

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.

The data is available for download. We did not find information about a specific license.

VI-C Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

Unknown.

VII. Maintenance

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

Yes. The Virginia State Crime Commission.

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

At the point of this publication, it is stated that "if you are having trouble downloading the dataset, please email meredith.farrar-owens@vcsc.virginia.gov or call Sentencing Commission staff at 804.225.4398"

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

"Data continues to be reviewed, revised, and validated as necessary" [1].

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

N/A

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.

No.

References

[1] V. State Crime Commission, "Virginia Pre-trial Data Project: Data Codebook, October 2017 Project Dataset," September 2021. [Online]. Available: http://www.vcsc.virginia.gov/pretrialdataproject.html

[2] N. Institute of Corrections, "Virginia Pretrial Risk Assessment Instrument (VPRAI)." [Online]. Available: https://nicic.gov/virginia-pretrial-risk-assessment-instrument-vprai

[3] A. Pretrial Policy and Research (APPR), "Public Safety Assessment (PSA)." [Online]. Available: https://advancingpretrial.org/psa/about

[4] V. State Crime Commission, "Virginia Pre-Trial Data Project: Final Report," 2021. [Online]. Available: http://www.vcsc.virginia.gov/pretrialdataproject.html