National Survey on Drug Use and Health (NSDUH)
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The National Survey on Drug Use and Health (NSDUH) was created to provide up-to-date information on tobacco, alcohol, and drug use, mental health and other health-related issues in the United States [1].

The data provide estimates of substance use and mental illness at the national, state, and substate levels. NSDUH data also help to identify the extent of substance use and mental illness among different subgroups, estimate trends over time, and determine the need for treatment services [1].

I-B Who created the dataset?

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

The dataset is created by the Substance Abuse and Mental Health Services Administration (SAMSA), with data collection and analysis conducted under contract with RTI International.

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

NSDUH was created to support prevention and treatment programs, monitor substance use trends, estimate the need for treatment and inform public health policy. Prior to NSDUH, there was no annual health survey that reported on substance abuse and other health issues.

II. Composition

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

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

Instances in the dataset comprise of individual survey responses.

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 is released annually, containing information from $\sim 55,000$ respondents. 2020 was an exception due to the Covid-19 pandemic, when there were only $\sim 30,000$ respondents.

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.

Survey responses are collected from a sample of the population. The target population of the survey is defined as the civilian, noninstitutionalized population of the United States. NSDUH collects information from residents of households and non-institutional group quarters (e.g., shelters, rooming houses, dormitories) and from civilians living on military bases. The survey excludes homeless people who do not use shelters, military personnel on active duty, and residents of institutional group quarters, such as jails and hospitals [2].

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.

Each data instance includes age at first use, as well as lifetime, annual, and past-month use of the following drugs:

- alcohol
- marijuana
- cocaine (including crack)
- hallucinogens
- heroin
- inhalants
- tobacco
- pain relievers
- tranquilizers
- stimulants
- sedatives

Respondents are asked about personal and family income, health care access and coverage, illegal activities and arrest records, problems resulting from the use of drugs, and perceptions of risks. For marijuana, respondents are asked about how and how often they obtain the drug. Demographic data collected include gender, race, age, ethnicity, educational level, employment status, income level, veteran status, household composition, and population density.

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.

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, both are self-reported by the respondent.

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.

Missing and erroneous data are dealt with in pre-processing. For detailed information, see section 2.3.2 and 2.3.3 from the documentation [2].

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.

Yes, the data contains information on substance use and abuse as well as confidential medical information on each subject. In addition, respondents are asked about criminal activity and arrests. Demographic information and information of income and employment is also disclosed.

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.

No.

III. Uses

III-A Has the dataset been used for any tasks already?

If so, please provide a description.

The dataset has been used in over 1,700 publications since it's establishment in 1979. A repository of which can be found:

https://www.icpsr.umich.edu/web/ICPSR/series/64

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.

Yes. Please see above.

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

For example: testing predictive policing systems, predicting recidivism.

The NSDUH dataset can be used for to investigate questions around drug use and abuse, as well as its relationship to mental health and involvement with the criminal justice system. Age at first use, past-month, annual, and lifetime use is reported for many drugs (listed in a subsequent section), as well as treatment history. The following demographics are also self-reported: age, race, sex, level of education, employment status, income, and veteran status.

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?

The data is self-reported and is not corroborated in anyway. Although the survey is anonymous, respondents may self-report the amount of drug use or not disclose other details. This may be more prevalent in some demographics compared to others.

Due to the Covid-19 pandemic, collection methodology changed for 2020, making direct comparison to other years difficult.

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 collection methods used for NSDUH to conduct in-person interviews with sampled individuals. Confidentiality is stressed in all written and oral communications with potential respondents. Respondents' names are not collected with the data, and computer-assisted interviewing (CAI) methods are used to provide a private and confidential setting to complete the interview.

IV-B Was the information self-reported?

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

Yes. The data consists of the survey responses, although some variables are retracted in the publicly available version of the data.

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

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

A scientific random sample of household addresses are selected across the United States. Once selected, no other address be substituted for any reason. At the end of the completed interview, participants receive $30 as a token of appreciation for their help. Interviews are facilitated by paid field interviewers [2].

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.

Data is collected on an annual basis, with surveys taking place throughout the year. Data is available from the year 1979 onward.

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.

All projects involving human subjects must be approved by SAMHSA's Office of Research Protection, which serves as RTI's Institutional Review Board (IRB) under federal regulations. This committee looks very closely at the written introduction to the study to be sure the respondents are being properly informed [3].

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?

Yes, responding to the survey is optional.

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.

Unknown.

V. Pre-processing, cleaning, labeling

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

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

Pre-processing steps are described in sections 2.3.2 and 2.3.3 in the documentation [2].

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.

SAMHSA will have access to the raw survey responses, but these are not publically available.

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)?

The dataset is avilable to download on the SAMSA website, here.

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 is mentioned on the website or in the codebook.

VII. Maintenance

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

Yes. The dataset is supported and hosted by SAMSA in collaboration with RTI international.

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

RTI international, the company which manage the collection and processing can be contacted at: NSDUH-Helpdesk@rti.org

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

New data is released annually.

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

Yes. Data from previous years continue to be hosted.

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] SAMHSA, "National survey on drug use and health (NSDUH) population," 2021. [Online]. Available: https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001

[2] S. Abuse and Mental Health Services Administration, "2020 National Survey on Drug Use and Health (NSDUH): Methodological Summary and Definitions," 2021. [Online]. Available: https://www.samhsa.gov/data/sites/default/files/reports/rpt35330/2020NSDUHMethodSummDefs091721.pdf

[3] ——, "2019 National Survey on Drug Use and Health (NSDUH): Field Interviewer Manual," 2020. [Online]. Available: https://www.samhsa.gov/data/sites/default/files/reports/rpt23074/NSDUHmrbFIManual2019.pdf