DATA USER GUIDE AND METHODOLOGY DOCUMENTATION (VERSION 1.0)

Jaclyn Schildkraut, Ph.D., M. Hunter Martaindale, Ph.D., & Emily Greene-Colozzi, Ph.D.

This codebook is part of a project funded in part by the U.S. Department of Homeland Security's Center for Prevention Programs and Partnerships, Opportunity No. DHS-24-TTP-132-00-99.

Contents

Overview of the Dataset

The Sharing Information to Stop Mass Shootings (SISMS) dataset is a structured, multi-source database of mass public shooting incidents in the United States. The data capture incidents occurring between 1999 and 2024 and include detailed information on both event-level characteristics and perpetrator-level behaviors. The current release includes information on 171 incidents and 175 perpetrators. The dataset is comprised of two linked components: an event-level database of mass public shooting incidents and a perpetrator-level database capturing pre-attack communications and behaviors. Information was compiled from a combination of official records, secondary reports, and triangulated open-source materials to document observable characteristics of each case.

The database is designed to support the systematic study of pre-attack behaviors, situational factors, and incident characteristics associated with mass public shootings. A particular emphasis is placed on documenting behaviors that were observable to others prior to the attack, allowing for analyses focused on prevention, intervention, and response. These data are structured to distinguish between incident-level and perpetrator-level information while allowing for integrated analysis across both units.

The analytic sample represented in the dataset reflects incidents for which sufficient information was available to document relevant characteristics, including pre-attack behaviors. Accordingly, the dataset should be understood as a subset of cases with documented information rather than a complete census of all mass public shootings during the study period. This distinction is important for interpretation, as the absence of a coded behavior does not necessarily indicate that the behavior did not occur, but rather that it was not identified in available source materials.

The dataset is intended for use by researchers, practitioners, and policymakers seeking to better understand patterns associated with mass public shootings, particularly in relation to observable warning behaviors and opportunities for prevention and intervention. By combining detailed behavioral indicators with incident-level characteristics, the data provide a foundation for examining how individual, situational, and contextual factors intersect in these events.

Suggested Citation (APA 7th Edition)

Users of the SISMS dataset should cite the data as follows:

Schildkraut, J., Martaindale, M. H., & Greene-Colozzi, E. A. (2026). *Sharing Information to Stop Mass Shootings (SISMS) dataset: Mass public shooting incidents in the United States, 1999–2024*. Regional Gun Violence Research Consortium, Rockefeller Institute of Government.

Definition and Case Selection Criteria

For the purposes of this dataset, mass public shootings are defined as targeted acts of violence occurring in public or populated settings, involving one or more perpetrators and multiple victims (fatalities and/or injuries) within a single 24-hour period. Incidents were included when victims and/or locations were selected at random or for symbolic purposes.[1] Both single-perpetrator and multiple-perpetrator incidents are included in the dataset.

Incidents were excluded if they were associated with gang-related activity or organized terrorist action. Events that occurred in private residences without a broader public or community impact were not included unless they met the criteria for a public or populated setting. These inclusion and exclusion criteria were applied consistently across cases to ensure alignment with established definitional parameters used in prior research.

The dataset includes incidents occurring between 1999 and 2024. Cases were identified through systematic searches of official records, secondary reports, and open-source materials. To be included in the analytic sample, incidents were required to have sufficient available information to document core characteristics of the event and, where applicable, perpetrator behaviors. In particular, inclusion in the behavioral dataset required documented evidence of pre-attack communications or behaviors, consistent with prior work examining leakage and related warning signs.[2] As a result, the dataset reflects cases with documented information rather than a complete census of all incidents meeting the definitional criteria.

Data Structure and File Organization

The SISMS dataset is comprised of two linked databases: the Mass Public Shootings Event (MPSE) database and the Pre-Attack Communications and Behaviors of Mass Public Shooters (PACB-MPS) database. Each database is structured to capture a distinct unit of analysis while allowing for integrated analysis across datasets.

The MPSE database contains event-level data, with each row representing a single mass public shooting incident. Variables in this file describe characteristics of the incident, including location type, date, number of victims, and other situational factors. The MPSE database includes 171 incidents.

---

[1] Schildkraut, Jaclyn, and H. Jaymi Elsass. *Mass Shootings: Media, Myths, and Realities*. Santa Barbara, CA: Praeger, 2016.

[2] Greene-Colozzi, Emily A. *Mitigating the Harm of Public Mass Shooting Incidents through Situational Crime Prevention*. PhD diss., The Graduate Center, City University of New York, 2022. https://academicworks.cuny.edu/gc_etds/4949/.

The PACB-MPS database contains perpetrator-level data, with each row representing an individual perpetrator associated with a mass public shooting incident. Variables in this file capture demographic characteristics as well as documented pre-attack communications and behaviors. The PACB-MPS database includes 175 perpetrators.

The difference in the number of incidents and perpetrators reflects the inclusion of cases involving multiple perpetrators. In such cases, a single incident in the MPSE database may be associated with more than one perpetrator in the PACB-MPS database. As a result, the data reflect a one-to-many relationship between incidents and perpetrators.

The two databases are linked through shared unique identifiers that allow users to merge files for analysis. Each incident in the MPSE database is assigned a unique event identifier (EID), which also appears in the PACB-MPS database for each associated perpetrator. In addition, each perpetrator in the PACB-MPS database is assigned a unique perpetrator identifier (PID), which is also included in the MPSE database. These identifiers allow the datasets to be merged using either EID or PID, depending on the unit of analysis and analytic approach. This structure enables users to conduct analyses at either the incident level or the perpetrator level, as well as to combine data across levels when appropriate.

Users should be aware that the appropriate unit of analysis depends on the research question being examined. Analyses focused on incident characteristics should be conducted using the MPSE database, while analyses focused on individual behaviors or characteristics should be conducted using the PACB-MPS database. When merging the datasets using EID, users should account for the one-to-many relationship between incidents and perpetrators, which may result in multiple rows per incident and duplication of incident-level variables across perpetrators. In contrast, merging or analyzing data using PID will retain a one-row-per-perpetrator structure. Careful consideration of the unit of analysis is essential to ensure accurate interpretation of results.

Data Collection and Sources

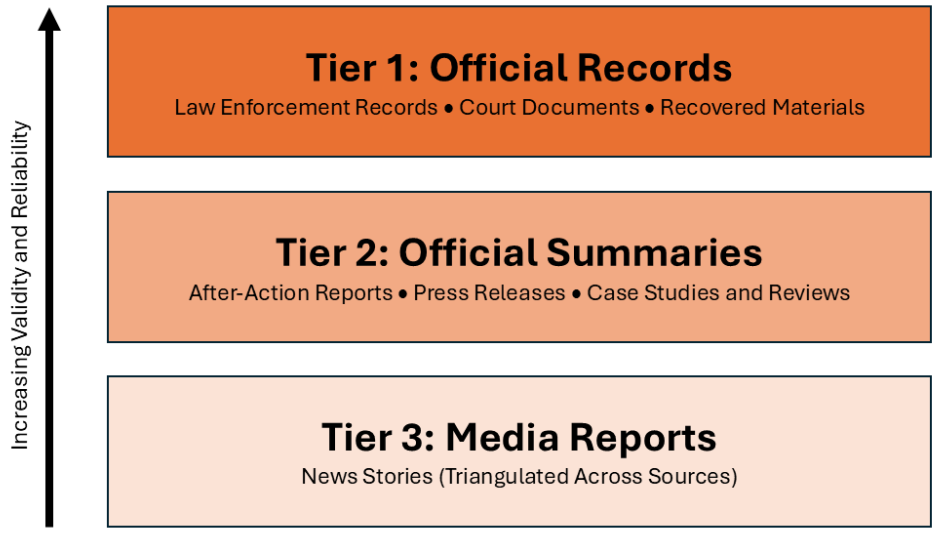
Data for the SISMS dataset were compiled using a structured, multi-source approach designed to prioritize accuracy, completeness, and verification across cases. Source materials were drawn from three primary tiers: official records, official summaries and interpretive materials, and open-source media reports. This tiered approach reflects differences in the level of detail and reliability across source types, with official records representing the most comprehensive and authoritative sources of information (see Figure 1).

Tier 1 sources consisted of official records, including law enforcement investigative files, court documents, and recovered materials. These sources were obtained through public records requests and other official channels and provide the most detailed and direct accounts of incidents and perpetrator behaviors. Across the project, more than 280 public

records requests were submitted, resulting in the collection of over 144,000 pages of official documentation.

Tier 2 sources included official summaries and interpretive materials, such as after-action reports, agency press releases, and case studies. These sources often synthesize information from primary records and provide additional context regarding incident dynamics and investigative findings.

Figure 1. Three-Tier Data Collection Framework



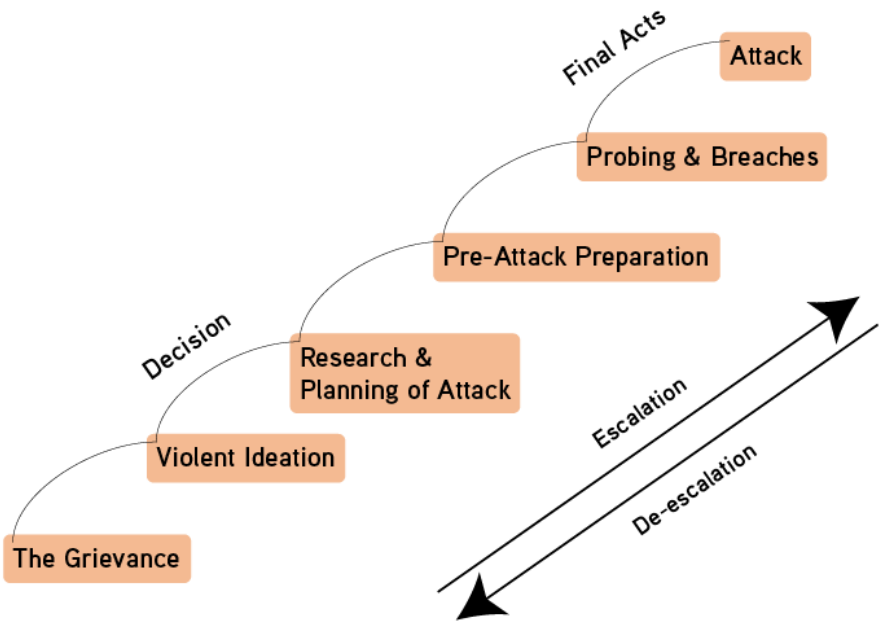
Tier 3 sources consisted of open-source media reports. Information from media sources was included only when it could be corroborated across at least three independent and reputable outlets. Open-source searches were conducted using a wide range of search engines and media platforms to ensure comprehensive coverage of each case.

Data collection was guided by a structured coding instrument informed by the Path to Intended Violence framework, which conceptualizes targeted violence as a progression of observable behaviors and experiences. This framework informed the identification and organization of variables related to grievance development, research and planning, preparation, leakage, and other pre-attack indicators. The use of this framework supports the systematic documentation of behaviors that may be observable to others prior to an attack, while allowing for variation in how such behaviors manifest across cases.

All cases were coded by trained research assistants using a standardized coding protocol. Following initial coding, each case underwent a multi-stage verification process. A lead researcher independently reviewed coded variables and source materials, after which a verification memo was prepared and reviewed by a second lead researcher. Discrepancies were resolved through discussion and, when necessary, adjudication by an additional

reviewer. This process was implemented to promote consistency and reliability in coding across cases (see Figure 2).[3]

Figure 2. The Path to Intended Violence Model[4]



Coding Procedures and Verification

All variables in the SISMS dataset were coded using a standardized coding instrument designed to capture observable characteristics of incidents and perpetrators based on available source materials. Coding was conducted by trained research assistants following a structured protocol to promote consistency across cases.

The dataset employs a presence-based coding framework. Variables were coded as present (1) only when sufficient confirmatory evidence was identified in source materials. Variables were coded as absent (0) only when there was explicit evidence indicating that the behavior or characteristic did not occur. In cases where available information was insufficient to determine whether a behavior or characteristic was present or absent, the variable was coded as missing and is denoted as −99 in the dataset. This approach prioritizes the documentation of verified information and avoids assumptions about the absence of behaviors that may not have been reported.

---

[3] Calhoun, Frederick S., and Stephen W. Weston. *Contemporary Threat Management*. Leesburg, VA: Specialized Training Services, 2003.

[4] Adapted from White, Stephen, and J. Reid Meloy. "Threat Assessments 101: Understanding the Red Flags of Workplace Violence." Webinar, September 20, 2017. https://vimeo.com/234743780.

As a result of this coding framework, missing data are an expected feature of the dataset and reflect the availability of information rather than measurement error. Users should not interpret missing values as evidence that a behavior did not occur. Instead, missing values indicate that the presence or absence of the behavior could not be determined based on available sources.

To enhance reliability, all cases underwent a multi-stage verification process following initial coding. First, a lead researcher independently reviewed coded variables alongside the underlying source materials. Second, a verification memo was prepared summarizing key aspects of the case and any uncertainties in coding. This memo was then reviewed by a second lead researcher. Discrepancies in coding were resolved through discussion and, when necessary, adjudication by an additional reviewer. This process was implemented to promote consistency and accuracy in the final dataset.

In some instances, variables may include additional codes to distinguish between different types of missingness or to capture cases where information was explicitly unavailable. These coding conventions are described in detail in the accompanying codebooks and should be consulted when preparing data for analysis.

Key Measures and Constructs

The SISMS dataset includes a broad set of variables capturing incident characteristics and perpetrator-level information, organized across the event-level (MPSE) and perpetrator-level (PACB-MPS) databases. Detailed variable definitions, coding schemes, and response categories are provided in the accompanying codebooks and should be consulted for all analytic use.

In the MPSE database, variables primarily describe incident characteristics. These include temporal and geographic information (e.g., date, time, city, state, and location identifiers), location type, and victim outcomes. Victim outcome measures are captured as count variables and include the number of individuals killed at the scene, those who died at a later time, and individuals sustaining nonfatal injuries, including both gunshot injuries and total injuries. The dataset also includes indicators related to incident structure, such as whether multiple perpetrators were involved.

The PACB-MPS database captures perpetrator-level information across a range of domains. These include demographic characteristics (e.g., age, sex, race, occupation), prior system contact (e.g., criminal history, law enforcement contact), and event outcomes, such as how the incident was resolved (e.g., suicide, police interdiction, surrender, or flight).

The dataset also includes detailed information on firearms and weapons associated with each incident. These variables capture firearm type as well as specific weapon characteristics, including make, model, and caliber when available. Additional measures

document firearm acquisition (e.g., source, timing, and location), as well as contextual details such as whether background checks or waiting periods were completed. The dataset further includes information on ammunition and magazine characteristics (e.g., quantity of ammunition brought to the scene, rounds fired, total ammunition in possession, number of magazines brought to and used during the attack, and the presence of extended or drum magazines), along with the presence and use of multiple firearms or other weapons and costuming elements associated with the perpetrator. Where documented, the dataset also captures firearms and weapons to which the perpetrator had access but that were not brought to or used during the attack.

A central component of the perpetrator-level data is the documentation of pre-attack communications and behaviors. These variables are organized into domains informed by the Path to Intended Violence framework and are intended to capture a range of documented behaviors and experiences that may occur prior to an attack.

Grievances and Motivations
Variables in this domain capture documented stressors, perceived injustices, or motivating factors associated with the perpetrator prior to the attack. These may include interpersonal, occupational, or other forms of strain reflected in available source materials.

Leakage
Leakage is defined as the communication of intent to do harm to a third party prior to the attack.[5] This includes verbal statements, written communications, text messages, online posts, or other forms of expression indicating potential or planned violence. Variables in this domain capture the presence of leakage, the modality through which it occurred (e.g., in-person or online), the individuals or groups who were aware of the communication, the content of the communication, and whether the information was reported.

Planning and Preparation
This domain includes variables capturing behaviors associated with preparing for the attack. These measures include research behaviors, planning activities, weapons acquisition and practice, and the creation of legacy materials intended to document or communicate the perpetrator's actions or motivations.

Concerning Behaviors
In addition to pathway-aligned behaviors, the dataset includes variables capturing other documented concerning behaviors that may co-occur with pre-attack activities. These behaviors reflect a range of observable actions or patterns identified in source materials and provide additional context for understanding the lead-up to the incident.

[5] Meloy, J. Reid, and Mary Ellen O'Toole. "The Concept of Leakage in Threat Assessment." *Behavioral Sciences & the Law* 29, no. 4 (2011): 513–527. https://doi.org/10.1002/bsl.986.

Users should note that these domains are used to organize variables conceptually and do not imply a fixed or linear sequence of behaviors across cases. Behaviors may occur in different orders, overlap, or be absent depending on the availability of documented information. These domains are not mutually exclusive, and multiple categories of behaviors may be documented within a single case.

Many variables within these domains are coded as dichotomous indicators reflecting the presence of documented evidence. As described in the coding procedures, these measures should be interpreted within the context of the dataset's presence-based coding framework. Additional detail on all variables, including coding decisions and category definitions, is provided in the codebooks.

Missing Data and Analytical Considerations

Missing data are an expected feature of the SISMS dataset and reflect the availability of information across cases rather than measurement error. As described in the coding procedures, variables are coded as missing (−99) when sufficient information is not available to determine whether a behavior or characteristic was present or absent. Accordingly, missing values should not be interpreted as evidence that a behavior did not occur.

The presence-based coding framework used in the dataset has important implications for analysis. Since variables are coded as present only when supported by documented evidence, observed frequencies reflect confirmed instances rather than the full universe of behaviors that may have occurred. As a result, estimates derived from the data should be interpreted as conservative.

Analytical decisions regarding the treatment of missing data may affect sample size and statistical power. For example, the use of listwise deletion in multivariate analyses may result in smaller analytic samples when variables have differing levels of missingness. In contrast, pairwise approaches may retain more cases for bivariate analyses but can result in varying sample sizes across estimates. Users should consider these tradeoffs when designing analyses.

Given these considerations, users are encouraged to assess patterns of missingness and to report analytic decisions transparently. Reporting the number of cases included in each analysis can help contextualize findings and improve comparability across studies. Where appropriate, users also are encouraged to report the extent of missingness for key variables to provide additional context for interpretation.

Strengths and Limitations

The SISMS dataset represents one of the most comprehensive, systematically collected sources of information on mass public shootings and associated pre-attack communications and behaviors. A key strength of the dataset is the use of a structured, multi-source data collection strategy, including extensive use of official records, which allows for detailed documentation of both incident characteristics and perpetrator-level information. The application of a standardized coding instrument, combined with a multi-stage verification process, further supports the consistency and reliability of the data.

Another important strength is the breadth and depth of the variables captured across domains. The dataset includes detailed information on incident characteristics, perpetrator demographics, firearms and weapons, and a wide range of documented pre-attack behaviors. In particular, the inclusion of granular information on firearms (e.g., weapon characteristics, acquisition, and ammunition and magazine details) and pre-attack communications provides a level of detail that is not consistently available in existing data sources. By documenting behaviors that were observable to others prior to the attack, the dataset also supports analyses focused on prevention, intervention, and response.

At the same time, the dataset is subject to several limitations. As with all research relying on retrospective and open-source data, the availability and completeness of information vary across cases. Some behaviors or characteristics may not be documented in available sources, resulting in missing data. In addition, the dataset reflects cases for which sufficient information was available to support coding and therefore should be understood as a subset of incidents with documented information rather than a complete census of all mass public shootings.

These limitations should be considered when interpreting findings, particularly in relation to the absence of documented behaviors and the generalizability of results. At the same time, the structured and transparent approach to data collection and coding provides a strong foundation for examining patterns in mass public shootings using documented evidence.

Guidance for Data Use

The SISMS dataset is intended to support research, policy analysis, and applied work focused on mass public shootings and prevention efforts. Users should carefully consider the structure of the data, including the distinction between event-level and perpetrator-level information, when designing analyses.

Given the presence-based coding framework, users should take care not to interpret missing values as evidence of absence and should clearly document analytic decisions related to missing data. Transparency in reporting methods, including the treatment of missingness

and the number of cases included in analyses, will support accurate interpretation and comparability across studies.

Users are encouraged to consult the accompanying codebooks when selecting and interpreting variables, as these provide detailed information on coding decisions, variable construction, and response categories. Careful alignment between research questions and the appropriate unit of analysis is essential to ensure valid and reliable findings.

When using the SISMS dataset in publications or presentations, users should cite the dataset as described in the Suggested Citation section. Proper attribution supports transparency and acknowledges the work involved in the development and maintenance of the dataset.

Dataset Version and Release Information

- Version: 1.0
- Release Date: June 2026
- Coverage Period: 1999–2024

Future updates to the dataset may expand coverage or incorporate additional information as it becomes available.

Data Files Included

The SISMS dataset release includes the following data and documentation files:

- MPSE dataset (event-level data; N = 171 incidents)
- PACB-MPS dataset (perpetrator-level data; N = 175 perpetrators)
- MPSE Codebook (variable definitions and coding structure for event-level data)
- PACB-MPS Codebook (variable definitions and coding structure for perpetrator-level data)

Contact Information

For questions about the SISMS dataset, including data structure, variable definitions, or appropriate use, please contact:

Jaclyn Schildkraut
Rockefeller Institute of Government
Jaclyn.Schildkraut@rock.suny.edu