NATIONAL AND STATE-LEVEL DATASETS OF UNITED STATES FORENSIC DNA DATABASES 2001–2025

A PREPRINT

Yemko Pryor[1], **João Pedro Donadio**[2], **Samantha C. Muller**[1], **Jenna Wilson**[3], **and Tina Lasisi**[1,4,*]
[1]Department of Anthropology, University of Michigan, Ann Arbor, USA
[2]Independent researcher, Vancouver, Canada
[3]Department of Human Genetics, University of Michigan, Ann Arbor, USA
[4]Center for Ethics, Society, and Computing, University of Michigan, Ann Arbor, USA
[*]Corresponding author: tlasisi@umich.edu

November 18, 2025

ABSTRACT

Forensic DNA databases in the United States have expanded substantially over the past two decades. However, comprehensive, harmonized data describing database structure and composition remain limited. This dataset series documents forensic DNA infrastructure across national and state levels from 2001 to 2025. It includes a reconstructed time series of monthly National DNA Index System (NDIS) statistics from FBI archives, capturing counts of offender, arrestee, and forensic profiles, participating laboratory totals, and investigations aided. A complementary dataset compiles publicly available state-level statistics and policy metadata on arrestee collection laws, familial search practices, and DNA collection statutes across all 50 states. A third dataset provides standardized demographic and annual collection data obtained through previously published public records requests, including racial and gender composition where reported. Together, these resources provide a foundation for studying the historical development of forensic DNA systems in the U.S., enabling longitudinal and cross-sectional analyses of database growth, policy variation, and reporting practices across jurisdictions.

Keywords Forensics · DNA Databases · Criminal justice · United States

1 Background & Summary

The Combined DNA Index System (CODIS) is a software platform originated by the FBI to store, compare, and search DNA profiles using a standardized set of short tandem repeat (STR) loci [1]. CODIS is implemented across three levels: local DNA index systems (LDIS) maintained by individual laboratories, state DNA index systems (SDIS) operated by state agencies, and the National DNA Index System (NDIS) administered by the FBI. DNA profile records flow upward from LDIS to SDIS to NDIS if they meet defined eligibility criteria and are organized into multiple indices, two of which are the Offender Index (which includes profiles from convicted individuals and, where authorized, arrestees) and the Forensic Index, which contains crime scene profiles of unknown offenders [2].

The FBI has published monthly NDIS statistics since the early 2000s, including counts of offender, arrestee, and forensic profiles, along with cumulative investigations aided [3]. These statistics are commonly referenced in the literature, with studies reporting point-in-time totals based on the FBI's website at the time of writing. For example, counts from June 2011 [4], January 2012 [5], May 2013 [6], October 2021 [7], November 2022 [8], and February 2024 [9]. However, these reports are released only as individual snapshots and are not preserved in a longitudinal, machine-readable format, limiting their use for tracking changes over time.

Data availability is inconsistent at the state level. Some states publish total SDIS counts or break down profiles by category, while others provide general department summaries or statutory references that are not specific to SDIS. Many do not publish SDIS-level statistics easily available to the public. Demographic composition is even more sparsely

documented. The only known compilation of racial and gender composition data comes from a 2020 publication by Murphy and Tong [10], who submitted Freedom of Information Act (FOIA) requests to all 50 states and received responses from seven. Their publication includes summary tables describing the demographic composition of state DNA databases and annual DNA profile collection counts. In this work, we digitize and standardize those materials to enable integration with other public data to support structured reuse.

This Data Descriptor introduces three datasets that document the structure, scale, and composition of U.S. forensic DNA databases. The first is a national-level NDIS time series (2001–2025) reconstructed from archived FBI webpages, including monthly counts by index and jurisdiction, participating laboratory counts, and investigations aided. The second is a current cross-sectional SDIS dataset that compiles reported totals and profile-type breakdowns if applicable, with accompanying metadata on arrestee collection, familial search authorization, and relevant state statutes for all 50 states. The third dataset provides structured versions of two resources derived from Murphy and Tong (2020): 1) racial and gender composition by state, year, and index type, and 2) annual state-level profile collection counts.

These datasets are intended to support research on the scale and governance of forensic DNA infrastructure in the United States, including cross-state comparison, time series modeling, and evaluation of demographic composition. The longitudinal NDIS dataset also enables investigation of how operational or policy shifts—such as the expansion of the CODIS core STR set to 20 loci in 2017 [11]—may have influenced profile accumulation. All datasets are versioned, documented, and released with processing code to facilitate transparent reuse.

2 Methods

2.1 Data Collection Strategy

We developed a three-pronged approach to capture different dimensions of the U.S. forensic DNA database landscape. Federal statistics were reconstructed from archived FBI websites, state policies were compiled through targeted web searches of current legislative sources, and demographic data were obtained through Murphy & Tong's supplementary data documenting FOIA responses on the demographic composition of seven states who responded to their requests.

2.2 Federal Statistics Extraction

We queried the Internet Archive's Wayback Machine API to identify all preserved snapshots of FBI NDIS statistics pages (Figure 1). Our search strategy accounted for the FBI's evolving web architecture: prior to 2007, statistics were distributed across jurisdiction-specific HTML pages (e.g., al.html for Alabama, ne.html for Nebraska), while post-2007 data were consolidated onto unified pages. We implemented a two-phase search algorithm that queried over 50 jurisdiction-specific URLs for the pre-2007 era and at least five distinct URL patterns for consolidated pages, testing both HTTP and HTTPS protocol variants to ensure comprehensive coverage across the archive's indexing variations.

Rather than sampling, we retrieved all available snapshots to maximize temporal resolution, yielding 11,359 unique captures spanning 255 distinct URL patterns. Each download attempt included robust error handling with exponential backoff (1s $\rightarrow$ 2s $\rightarrow$ 4s delays) and extended timeouts to manage network variability. This comprehensive approach captured the complete temporal evolution of NDIS reporting from July 2001 through August 2025 (Figure 2).

We processed downloaded snapshots using four era-specific parsers designed to accommodate the FBI's evolving reporting formats. The pre-2007 parser extracted data from individual state pages, identifying jurisdictions through both filename patterns (e.g., extracting "Pennsylvania" from "20040312_pa.html") and HTML content matching. Since these early pages lacked temporal metadata, we set report dates to null, relying instead on Wayback Machine capture timestamps. The 2008-2011 parser processed consolidated pages with jurisdiction sections separated by "Back to top" dividers and introduced extraction of "as of" date statements. The 2012-2016 parser captured the introduction of arrestee profile reporting alongside modified section headers. The post-2017 parser implemented regex patterns for the modern standardized format. Each parser extracted jurisdiction name, profile counts by category, laboratory counts, and investigations aided, along with temporal metadata from either explicit dates or snapshot timestamps.

The raw HTML data contained 169 unique text strings representing 54 actual jurisdictions, resulting from inconsistent naming conventions across time periods and data sources. We developed a comprehensive mapping table to standardize heterogeneous formatting of jurisdiction names while maintaining traceability to original sources. Examples included "Alabama", "Alabama Alabama", and "Alabama Stats Alabama" all representing the single jurisdiction of Alabama. Several non-state jurisdictions (i.e. Puerto Rico, DC/FBI Lab, DC/Metro PD, U.S. Army) were treated as distinct reporting entities and included when available.

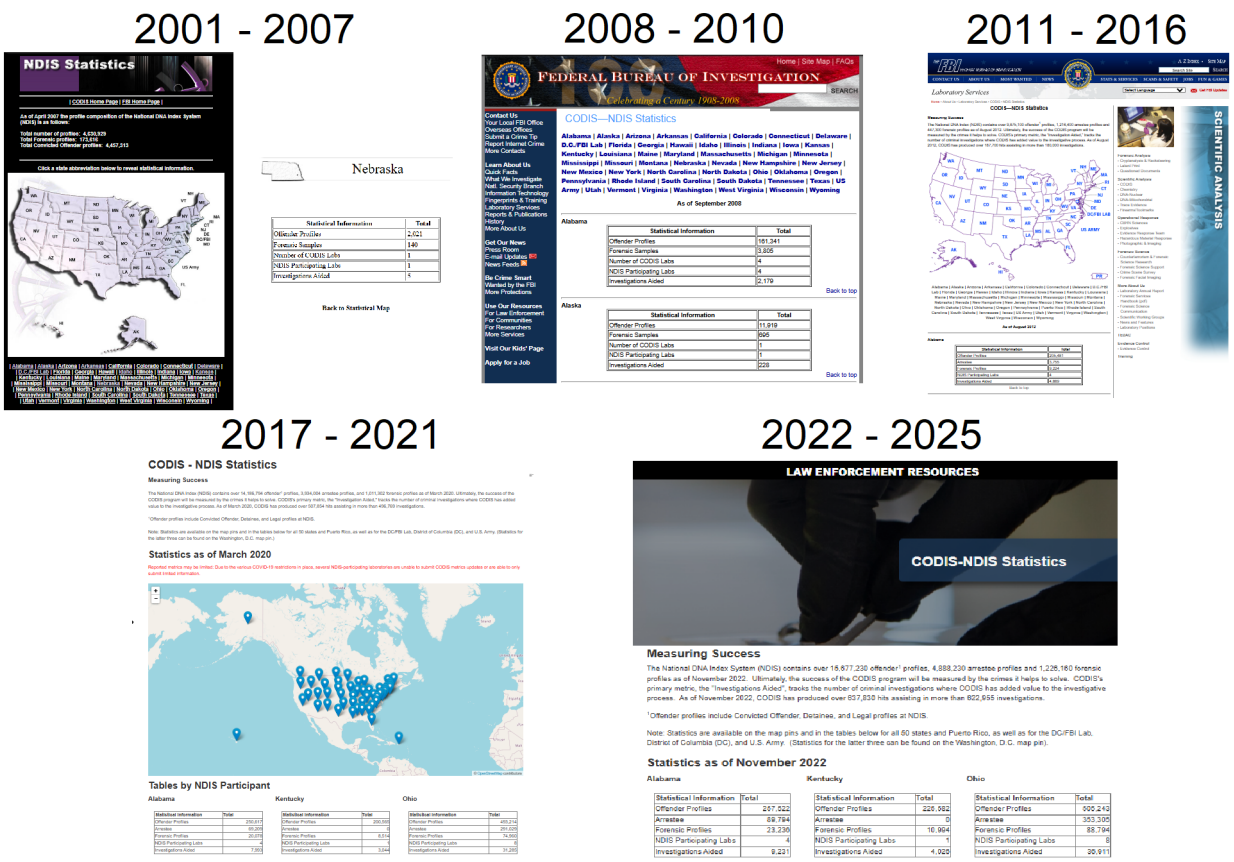

Figure 1: FBI formatting changes in NDIS statistics pages. Variations in the structure and organization of FBI NDIS webpages captured through FBI.gov in different eras. Webpages are grouped into five eras used to build the NDIS time series dataset. Eras of NDIS webpages are broken down into the pre-2007 era with jurisdiction-specific HTMLs, unified HTMLs with jurisdiction-based dividers from 2008-2010, modified section headers in 2011-2016, and more modernized unified pages in the post-2017 eras.

2.3 State DNA Database Counts and Policy Compilation

Targeted web searches were performed to find available information on state-level DNA databases. Data sources were identified and reviewed using official state government websites and annual reports. Searches were conducted by entering terms such as “SDIS,” “CODIS,” “Forensics”, or “DNA” into state website search bars, or by using general browser queries such as “[state name] AND DNA Database,” “[state name] AND annual review,” or “[state name] Police AND DNA Database.” Available information from state government webpages were codified to summarize the number of arrestees, offenders, forensic profiles, collection statutes and familial search practices per state. The most recent publicly available state-sourced database reports span from 2013 to 2025, depending on the state. Only 29 of the 50 states reported raw arrestee DNA totals. Policies regarding arrestee DNA collection were coded as binary variables (yes/no) when the statutory language was explicit (Figure 3A). Familial search policies were coded as "permitted" or "prohibited" when definitively stated or as “unspecified” when the statutes were silent or ambiguous on these particular provisions(Figure 3B).

2.4 Demographic FOIA Data Extraction and Standardization

We incorporated data from two appendices originally referenced in Murphy & Tong (2020) that were not accessible directly with the original article. At our request, the authors provided the original files, which include the scanned FOIA response letters from state agencies reporting demographic breakdowns of their forensic DNA databases. Additionally,

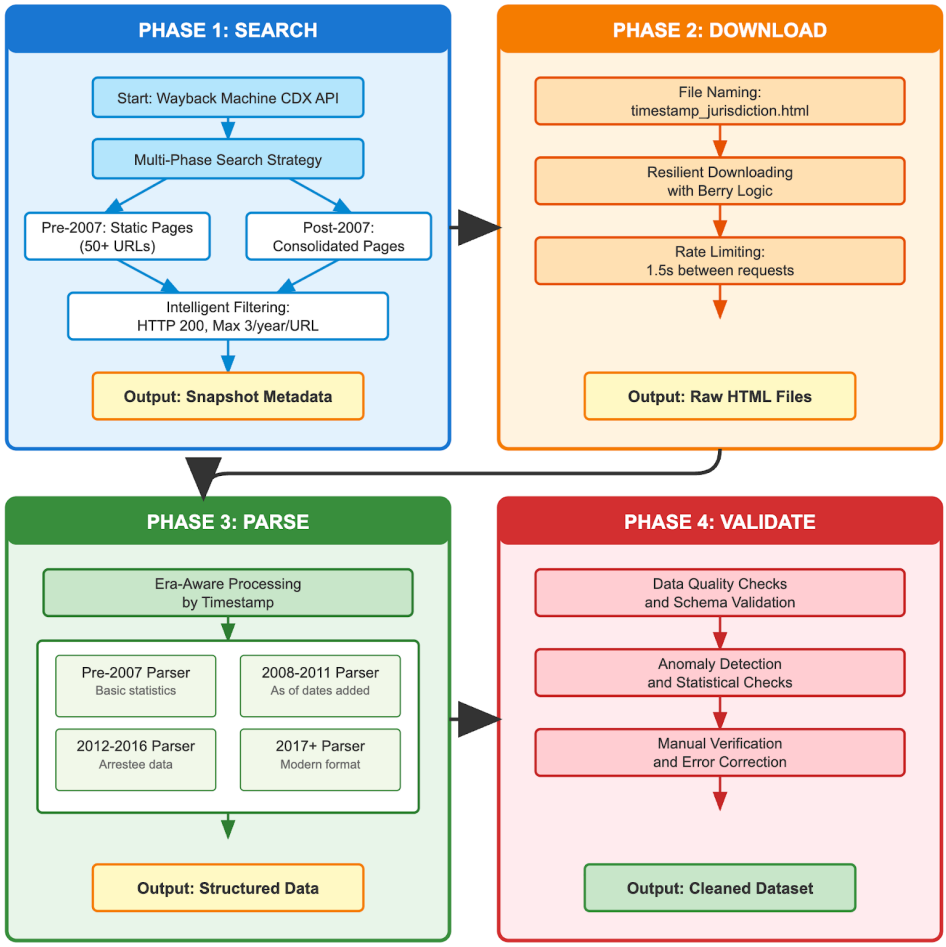

Figure 2: Parsing logic for NDIS statistics scraping. A diagram depicting the step-by-step approach used to scrape the Wayback Machine and consolidate FBI NDIS statistics using era-specific parsers, allowing for a validated cleaned dataset of complied NDIS time series data.

we provide an appendix with tables in which Murphy & Tong had already performed calculations combining state-level collection data, Census demographics, and estimated rates of annual DNA collection by race.

To make these appendices accessible and reusable, we transcribed all documents into machine-readable CSVs. For the first appendix, this primarily involved converting the demographic tables from scanned PDFs into structured data while retaining the categories used by the original agencies. For the second appendix, we treated Murphy & Tong's own calculated estimates as data values, preserving them alongside the raw counts they reported. This ensured that our release captures not only the state-provided figures but also the derived metrics that underpinned the analysis in their original article.

Together, these appendices provide two complementary resources: 1) agency-reported demographic counts from seven states and 2) Murphy & Tong's standardized estimates of annual DNA collection rates by race across all fifty states. We preserved the structure and terminology of the original materials while adding clear provenance fields so users can distinguish between reported values and author-calculated estimates.

3 Data Record

The datasets presented in this study are available on Zenodo [12]. Detailed information and reproducible code used to process each of the datasets presented in this article, together with the full analyzes and the scraping methodology, can be accessed through our accompanying research website: `https://lasisilab.github.io/PODFRIDGE-Databases/`.

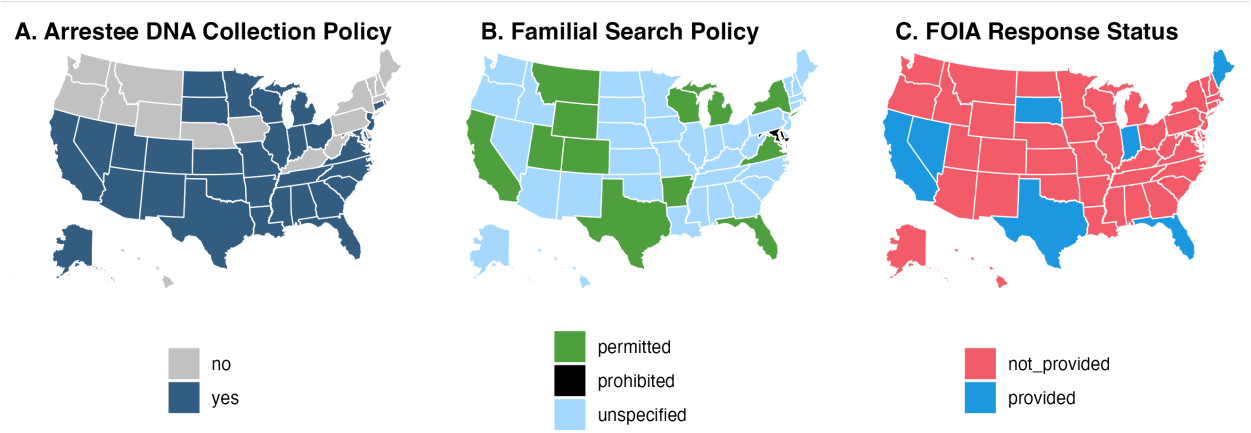

Figure 3: SDIS policies and FOIA availability mapped by state. (A) Arrestee DNA Collection Policy availability coded as yes (blue) or no(gray) variables, (B) familial Search Policy detailing states with permitted(green), prohibited(black) and unspecified(blue) data availability, and (C) FOIA Response Status coding availability status as not_provided(red) or provided(blue).

The raw NDIS time series, SDIS cross section, FOIA demographics and Annual DNA collection files are available for download in .csv format.

Figure 4 presents a visual representation of the hierarchical directories that reflect the data sources, processing stages, and the final outputs derived. Each major data stream (NDIS, SDIS, FOIA, and Annual DNA Collection) is subdivided into standardized processing levels (Raw, Intermediate, and Final) to ensure transparency and reproducibility between analyzes. The ndis_crossref subfolder houses a 2015 CODIS brochure, as well as the data from Table 1. The definitions for all organizational terminology used across all files are available as a data dictionary in the README.md file located in the data subfolder and also located on our Github repository.

4 Technical Validation

4.1 National DNA Index System (NDIS)

We validated the NDIS time series by testing temporal consistency within each jurisdiction and at the national aggregate level. Because the counts of offender, arrestee, and forensic profiles, as well as investigations aided, are cumulative measures, they are expected to increase monotonically across successive data snapshots. Although occasional expungements may occur, such legitimate decreases should be distinguishable from reporting artifacts through systematic anomaly detection. Here we describe our attempt to conservatively remove unambiguous (primarily typographical) errors and we encourage any users of this data to analyze both the raw and intermediate files in addition to the clean files in our 'data/ndis' folder (see 4) as they may wish to set their own definitions of what constitutes an anomaly in the data or analyze the anomalies themselves in their own right.

To identify and correct data-quality issues while preserving genuine profile removals, we developed a multi-category anomaly detection framework. For each jurisdiction j, metric x, and capture time t, we examined patterns in the recorded count $N_{j,t}^{(x)}$ relative to adjacent observations.

Spike-Dip detection. We flagged isolated deviations from the trajectory established by adjacent observations:

- $N_{j,t}^{(x)} > 2 \times N_{j,t-1}^{(x)}$ (more than double the previous value)
- $N_{j,t}^{(x)} < 0.5 \times N_{j,t-1}^{(x)}$ (less than half the previous value)
- $N_{j,t}^{(x)} > 2 \times N_{j,t+1}^{(x)}$ (more than double the next value)
- $N_{j,t}^{(x)} < 0.5 \times N_{j,t+1}^{(x)}$ (less than half the next value)

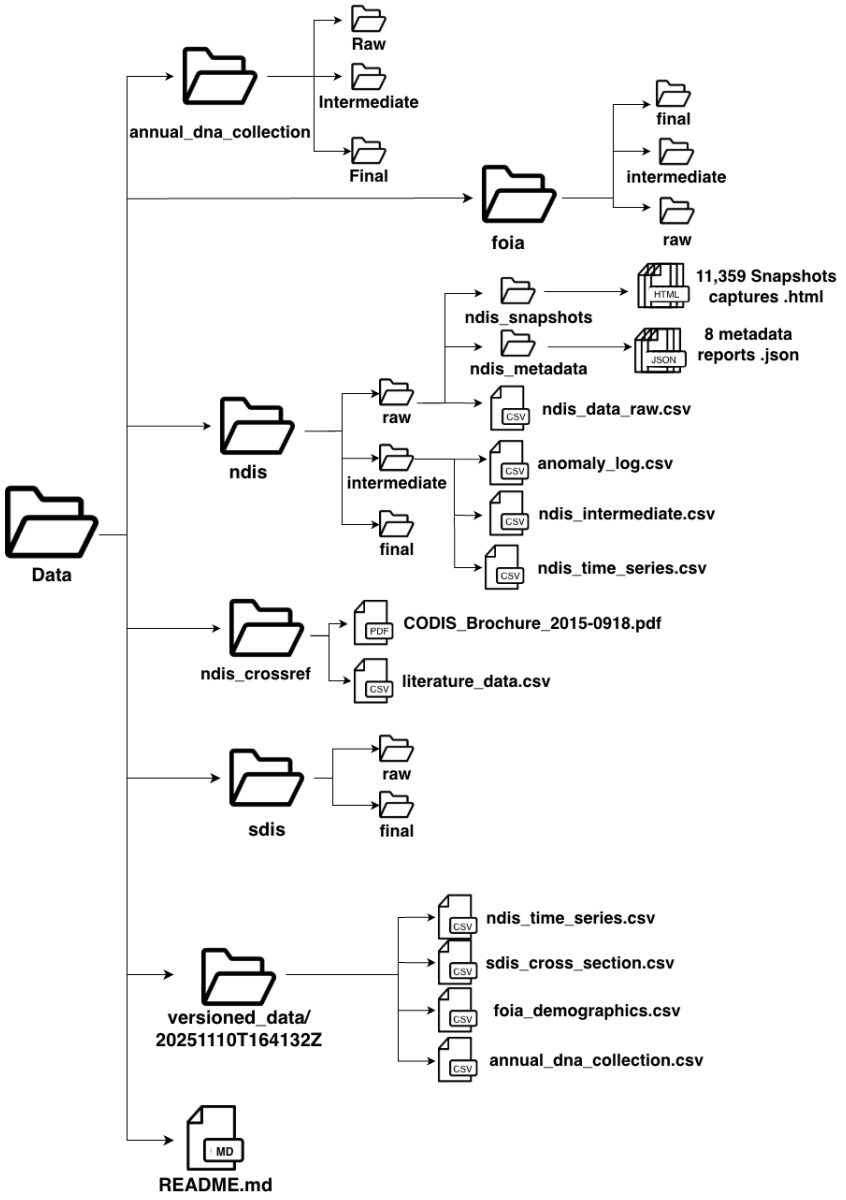

Figure 4: File and folder structure. A file map showcasing the hierarchical organization of directories and data sources present in the public GitHub repository: `https://github.com/lasisilab/PODFRIDGE-Databases`

Continuation of spike-dip patterns was detected when a previously flagged point showed recovery toward normal levels. These temporary surges or isolated dips indicate reporting synchronization anomalies rather than genuine profile count changes.

Zero error detection. We identified instances where zero values appeared following positive counts, which cannot occur in cumulative profile data: (1) A zero appearing after positive values was flagged as "zero_error" or (2) Consecutive zeros following the initial error were flagged as "cont_zero_error". All consecutive zero values from the first anomalous zero through recovery to positive numbers were removed as a cohesive error sequence.

Update lag detection. We detected system synchronization delays manifesting as oscillations within compressed timeframes. Values showing decreases or stability relative to neighbors within a 48-hour window (2 days between

successive observations) indicated transient states during database synchronization across multiple systems. Within each oscillation cluster, only the highest value, representing the true profile count, was retained; lower values reflecting intermediate synchronization states were removed. For Investigations Aided and Forensic Profiles, the update lag detection window was calibrated to the specific patterns observed in those metrics.

Value propagation. Following identification of flagged anomalies, any subsequent data point within the jurisdiction matching the exact value of a previously flagged point was also removed, as this pattern indicated persistent propagation of the initial error across reporting cycles. This rule was not applied to Participating Laboratories given the discrete nature of laboratory counts.

Metric-specific adjustments. The same logic was applied across metrics with modest threshold adjustments. Offender and forensic series used the two-fold spike–dip thresholds described above, with forensic series using a slightly higher upper threshold ($2.5\times$) and shorter update-lag windows. Arrestee profiles (from January 1, 2012 onward) exhibited only zero-after-positive anomalies; thus only the zero-error rule was required in practice. Investigations-aided series primarily required the 10-fold spike threshold together with the zero-error and propagation rules. NDIS laboratory counts are small integers; therefore only empirically justified jurisdiction-specific thresholds (threefold increases in Oklahoma and drops to 25% or less in Michigan) were used. Zero-after-positive and propagation rules were not applied to laboratory counts.

All anomaly flags (zero errors, spike–dips, continuation events, oscillations, and propagated anomalies) are preserved in the anomaly log. The cleaned series represents a deliberately conservative approach: we remove only values that clearly contradict the behavior of cumulative reporting systems while retaining plausible decreases.

Anomaly summary. Using these rules, we detected 477 anomalies across all metrics and jurisdictions (Figure 5). These included spike–dip events (278), update-lag oscillations (186), and zero-errors (13). Offender profiles showed the highest number of anomalies (213), followed by forensic profiles (177), investigations aided (71), NDIS laboratories (10), and arrestee profiles (6). Figure 6 compares raw and cleaned snapshot coverage.

External calibration. After cleaning, we compared national totals to independent reference points from FBI CODIS brochures (2000–2015) and published estimates. For each year, national totals were computed by summing the maximum jurisdiction-level values for that year. The reconstructed offender, arrestee, forensic, total-profile, and investigations-aided series closely match values from a 2015 FBI Brochure (available in our Zenodo 'ndis_crossref' folder) and the point-in-time estimates reported by Ge et al. (2012, 2014), Wickenheiser (2022), Link et al. (2023), and Greenwald & Phiri (2024) (Figure 7; Table 1). This correspondence supports the fidelity of the reconstructed national trajectory.

Table 1: Reference values from FBI publications and academic literature for NDIS database growth.

Citation	Date	Offender Profiles	Arrestee Profiles	Forensic Profiles	Total Profiles	Investigations Aided
FBI (Dec 2000)	12/1/00	441,181	NA	21,625	NA	1,573
FBI (Dec 2002)	12/1/02	1,247,163	NA	46,177	NA	6,670
FBI (Dec 2004)	12/1/04	2,038,514	NA	93,956	NA	21,266
FBI (Dec 2006)	12/1/06	3,977,435	54,313	160,582	NA	45,364
FBI (Dec 2008)	12/1/08	6,399,200	140,719	248,943	NA	81,955
FBI (Dec 2010)	12/1/10	8,564,705	668,849	351,951	NA	130,317
FBI (Dec 2012)	12/1/12	10,086,404	1,332,721	446,689	NA	190,560
FBI (Jun 2015)	6/1/15	11,822,927	2,028,734	638,162	NA	274,648
Ge et al. 2012	6/1/11	NA	NA	NA	10,000,000	141,300
Ge et al. 2014	5/1/13	NA	NA	NA	12,000,000	185,000
Wickenheiser 2022	10/1/21	14,836,490	4,513,955	1,144,255	NA	587,773
Link et al. 2023	11/1/22	NA	NA	NA	21,791,620	622,955
Greenwald & Phiri 2024	2/1/24	17,000,000	5,000,000	1,300,000	NA	680,000

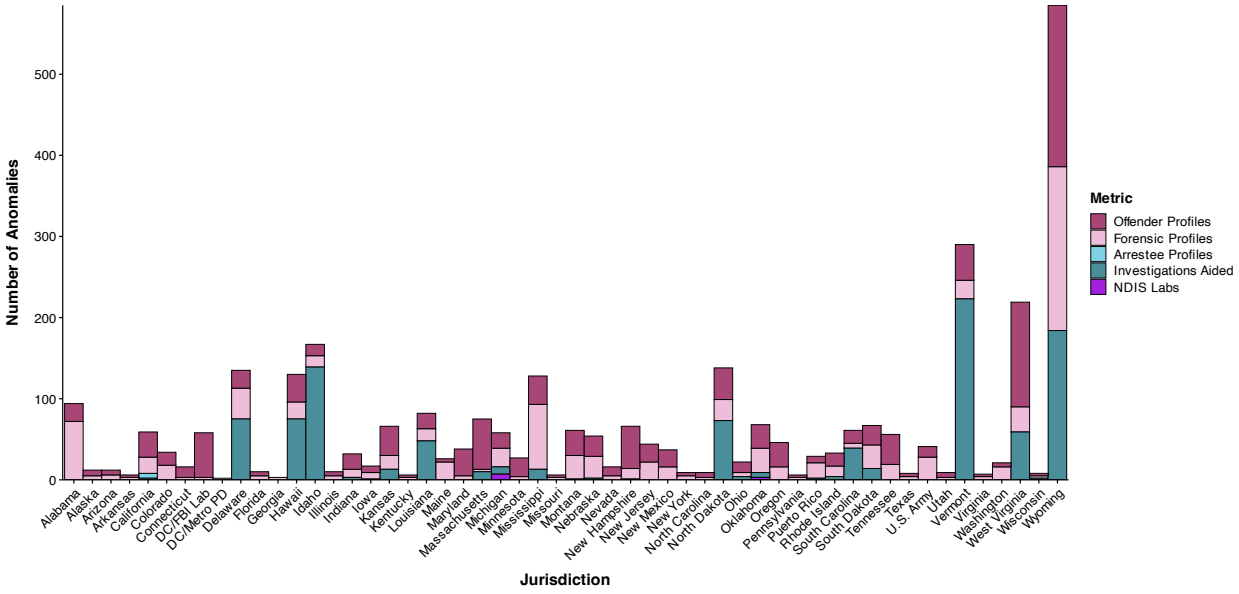

Figure 5: Anomalies distribution per state. A stacked bar plot detailing anomalies in Offender Profiles (magenta), Forensic Profiles (pink), Arrestee Profiles (cyan), Investigations Aided (turquiose) and NDIS Labs (purple) cataloged by state and corrected across all metrics and jurisdictions. Wyoming's high values can be explained by flagged value propagation in the early years, which may be due to periods where no updates were made.

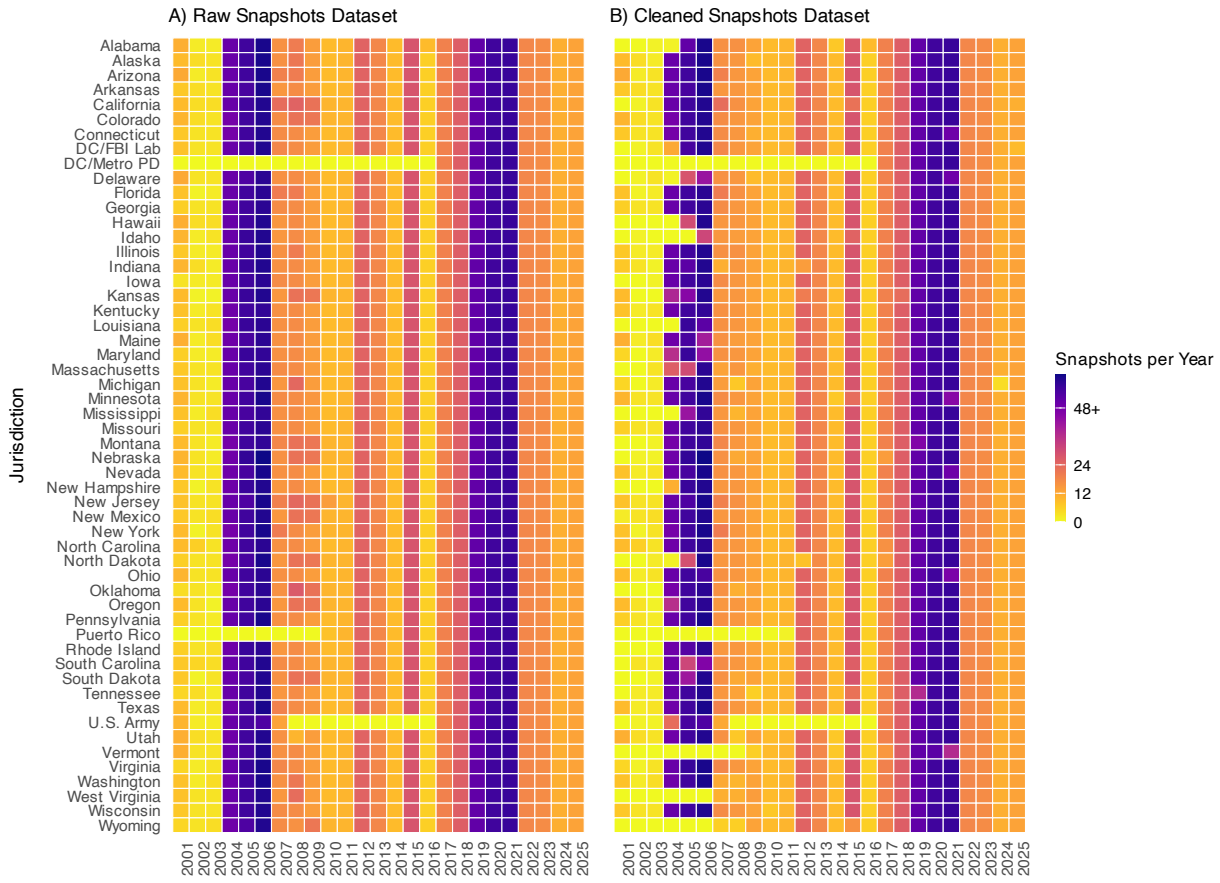

Figure 6: NDIS snapshot heatmaps. Heatmaps showcasing NDIS snapshots captured raw (A) and those retained after cleaning (B) the NDIS time series dataset.

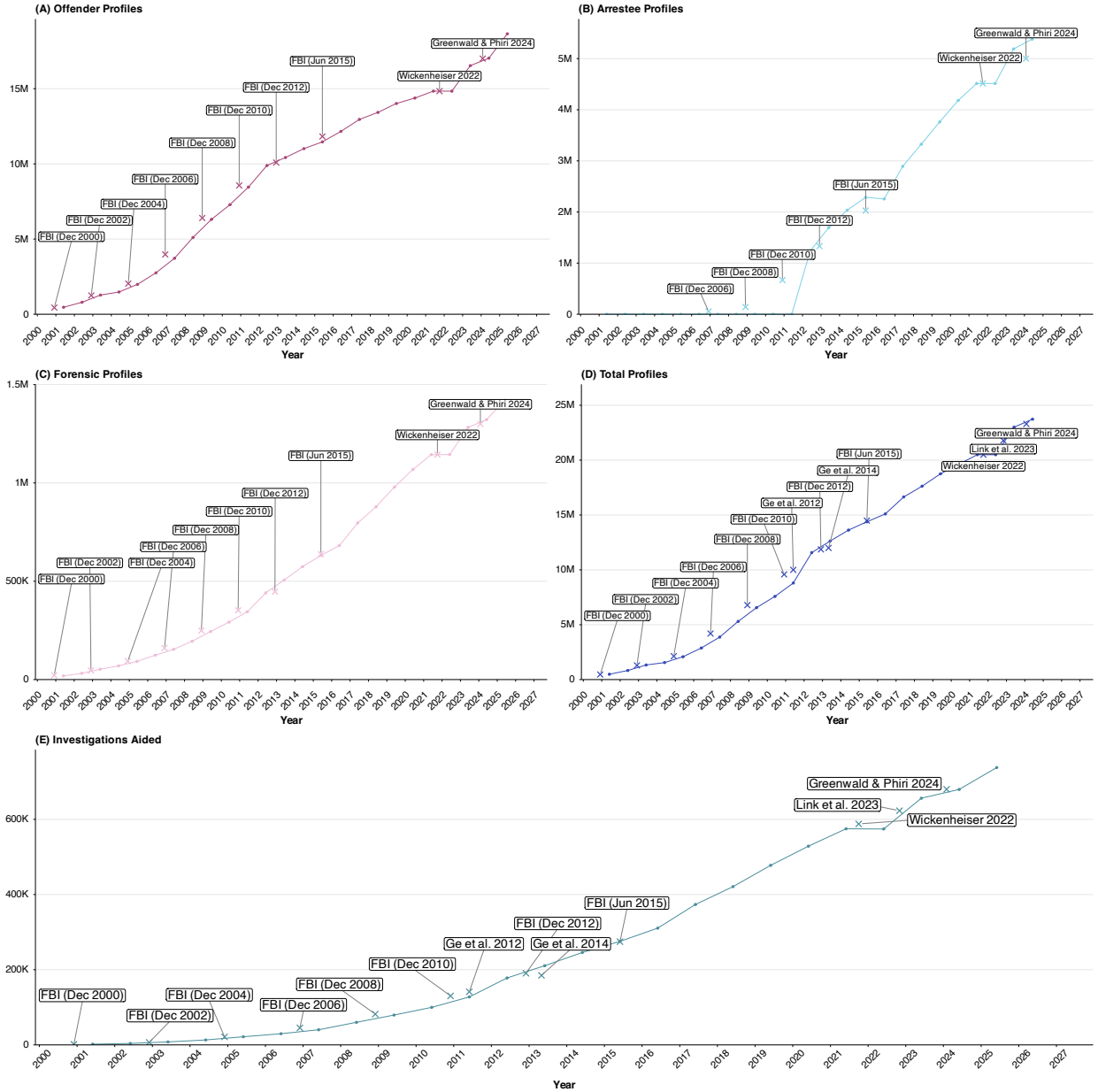

Figure 7: External calibration of NDIS data. NDIS data cross-referenced with both citations from the literature and numbers reported by an FBI 2015 CODIS brochure are plotted by year and split into five major metrics, (A)Offender Profiles, (B) Arrestee Profiles, (C) Forensic Profiles, (D) Total Profiles, and (E) Investigations Aided, showing NDIS time scaled growth from 2001-2025.

5 Data Availability

The datasets are archived on Zenodo (DOI: 10.5281/zenodo.17215677). Contents include NDIS (time series), SDIS (state cross-section and arrestee/familial search policies), and FOIA components. The FOIA folder contains the original scanned state responses received by Murphy & Tong (2020) and machine-readable transcriptions. Reuse of the FOIA scans should cite Murphy & Tong (2020) [10] in addition to this dataset.

6 Code Availability

All code used to generate the datasets is available in a public GitHub repository: `https://github.com/lasisilab/PODFRIDGE-Databases`. The repository contains Quarto notebooks, parsing scripts, and workflow documentation used to process the raw inputs into the released CSVs.

Acknowledgements

We thank Erin Murphy for providing access to the original FOIA-acquired state-level demographic disclosures and appendices from her original publication.

Author Contributions

Conceptualization: Lasisi, T.
Data curation: Donadio, J. P.; Lasisi, T.; Muller, S. C., Wilson, J.; Pryor, Y.
Methodology: Lasisi, T.; Donadio, J. P.
Validation: Lasisi, T.; Donadio, J. P.
Visualization: Donadio, J. P.; Lasisi, T., Pryor, Y.
Writing – original draft: Lasisi, T.
Writing – review and editing: Lasisi, T.; Donadio, J. P.; Muller, S.C.; Pryor, Y.; Wilson, J.
Supervision: Lasisi, T.

Competing Interests

The authors declare no competing interests.

References

[1] CODIS archive. `https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints/codis-2`, June 2022. Accessed: 2025-10-17.

[2] Forensic DNA education for law enforcement decisionmakers. `https://nij.ojp.gov/nij-hosted-online-training-courses/forensic-dna-education-law-enforcement-decisionmakers/partial-matches/codis-hierarchy-cont`. Accessed: 2025-9-26.

[3] CODIS-NDIS statistics. `https://le.fbi.gov/science-and-lab/biometrics-and-fingerprints/codis/codis-ndis-statistics`, July 2022. Accessed: 2025-9-26.

[4] Jianye Ge, Arthur Eisenberg, and Bruce Budowle. Developing criteria and data to determine best options for expanding the core CODIS loci. *Investig. Genet.*, 3(1):1, January 2012.

[5] Jianye Ge. DNA based familial searching and related statistical issues. *J. Forensics Res.*, 03(08), 2012.

[6] Jianye Ge, Hongyu Sun, Haiyan Li, Chao Liu, Jiangwei Yan, and Bruce Budowle. Future directions of forensic DNA databases. *Croat. Med. J.*, 55(2):163–166, April 2014.

[7] Ray A Wickenheiser. Expanding DNA database effectiveness. *Forensic Sci. Int. Synerg.*, 4(100226):100226, April 2022.

[8] Vivian Link, Yuómi Jhony A Zavaleta, Rochelle-Jan Reyes, Linda Ding, Judy Wang, Rori V Rohlfs, and Michael D Edge. Microsatellites used in forensics are in regions enriched for trait-associated variants. *iScience*, 26(10):107992, October 2023.

[9] Emily Greenwald and Linda Phiri. On accountability: Genetic tools for justice and injustice in criminal proceedings. *J. Sci. Policy Gov.*, 25(1), October 2024.

[10] Erin Murphy and Jun Tong. The racial composition of forensic DNA databases. 108, 2020.

[11] Douglas R Hares. Selection and implementation of expanded CODIS core loci in the united states. *Forensic Sci. Int. Genet.*, 17:33–34, July 2015.

[12] Pryor, Y., Donadio, J. P., Muller, M., Wilson, J., Lasisi, T. United states forensic DNA databases: NDIS, SDIS, and FOIA datasets, November 2025.