Methodology for count of unique individuals on child abuse and neglect registries

Introduction to NCANDS

NCANDS, or the National Child Abuse and Neglect Data System, is a federal data set collected by the Administration of Children and Families and managed by the National Data Archive on Child Abuse and Neglect at Cornell University, which handles its distribution to the public. The files contain de-identified records about child abuse reports, investigations and their outcomes.

The level of detail is such that access is heavily restricted and typically only provided to academic researchers using institutional review boards registered with the U.S. Office for Human Research Protections.

To gain access to the data, BuzzFeed News hired an independent institutional review board, submitted a proposal for study, and completed a course in Ethics and Human Subject Protection from the Association of Clinical Research Professionals.

As part of our agreement with NDACAN, we are not publishing any raw data used to complete this analysis.

Goal

To produce a reasonable “lower bound” estimate of all unique individuals added to state child abuse/neglect registry between 2008 and 2020 using data about confirmed perpetrators of child abuse/neglect.

Challenges

There are multiple challenges to obtaining a full count, and BuzzFeed News removed as many of them as possible, obtaining a particularly conservative lower bound estimate of the registry’s size across the country. The actual number is likely millions larger.

First, there is no variable in NCANDS indicating an individual’s status on the registry. In most states, all individuals with substantiations are added to the registry, so a unique count of substantiated individuals should be roughly the same as a unique count of people added to the registry. In other states, such as Michigan, the state only adds substantiated cases to the registry when the case indicates a high level of risk. It’s difficult to determine from NCANDS data alone, for instance, whether a particular substantiated individual would be added to Michigan’s registry.

Second, NCANDS is generally used in academic literature primarily to analyze children and reports, not perpetrators. There are fewer variables associated with perpetrators and that information is blank when the allegation hasn't been substantiated.

Lastly, it is not uncommon for individuals to have multiple substantiations either within the same report or even across years. Care must be taken to avoid counting the same perpetrator more than once.

Missing data

Some states did not submit NCANDS for certain years and others did not include perpetrator data. These are the states and years where the data was missing or unusable.

state	**year**
Hawaii	2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015
Georgia	2008, 2009, 2010, 2011, 2012, 2015, 2016
Iowa	2008, 2009, 2010, 2011, 2012, 2013
Oregon	2008, 2009, 2010, 2011
Maryland	2008, 2009
North Dakota	2008, 2009

Methodology and steps

First, the data was reshaped so that instead of each row representing a child-report combination, it represented a perpetrator-child-report combination. The table, as you might expect, is significantly longer than the very "wide" format of the original. An example of the format, with dummy data is below:

perpetrator_id	**report_id**	**child_id**	**Additional columns….**
2323423556	564767568	345346546	1.0
3423546536	456547657	435445656	2.0

The remaining columns are the ones indicating the perpetrator race, ethnicity, sex, maltreatment disposition and maltreatment type.

Next, for each state, perpetrator IDs were deduplicated and compared across years to look for overlaps in perpetrators. The expectation was that each consecutive year-pair should generate a small portion of natural repeats for individuals who accumulate multiple substantiations across a period of time. States with 0 or only a handful of overlaps between years are likely using a new perpetrator ID for each interaction, resulting in a new ID even if the individual is already in the state's system. We flagged all year-pairs with fewer than 1% matching and removed the latter year in the pair from the analysis.

The next thing we looked for was errors in the matching of perpetrators. For child IDs this can be accomplished by comparing the dates of birth between two matching child IDs. NCANDS does not include date of birth for perpetrators but we can approximate an identity match by comparing the sex and race values of a perpetrator ID across years.

For the purposes of a lower-bound count, we can tolerate a relatively high rate of errors. If an ID is re-used, it will result in an undercount of unique perpetrator IDs, not an overcount. For each state, I flagged all year-pairs with more than 5% errors and removed them from the analysis.

After both of these tests, the following states and years were excluded from the analysis:

state	**year**
Oklahoma	2009, 2011, 2012, 2013, 2015, 2016, 2017, 2018, 2019, 2020
North Carolina	2012, 2013, 2014, 2015, 2017, 2018, 2019, 2020
Pennsylvania	2011, 2013, 2015, 2018, 2019
Texas	2017, 2018, 2019, 2020
Virginia	2010, 2013, 2014, 2015
Illinois	2009, 2012, 2013
Nebraska	2013, 2016, 2017
California	2017, 2018, 2019
Florida	2012, 2013, 2014
Hawaii	2017, 2018
Indiana	2012, 2013
Maryland	2011, 2012
Tennessee	2010, 2012
New Hampshire	2011, 2020
New Jersey	2011, 2014
Vermont	2020
Utah	2009
Alabama	2009

Nevada	2018
Ohio	2011
New York	2019
Michigan	2015
Maine	2009
Kentucky	2012
Kansas	2009
Georgia	2015
Washington	2009

In addition to these two filters, we also removed several states entirely. The state of Georgia repealed its registry law in 2020, so I used no data from that state. Additionally, many states have selective registries that only add some substantiated individuals to the registry depending on type, severity or some other variable. Colleen Henry performed a very useful, recent survey of states, and indicated in a 2021 research paper the states that had more limited registries. We removed all states that did not add “all” substantiations to the registry.

The following states were removed:

Arkansas, California, Connecticut, Delaware, Iowa, Louisiana, Maine, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, North Carolina, North Dakota, Utah, Wisconsin.

Results

Using only the year-pair diagnostics, the final count of perpetrators was: 4,706,050.

Excluding selective states from the count, the absolute lower bound was: 3,087,687