Blending Probability Sampling with API and Administrative Data

Estimating Recall Rates for Rideshare Vehicles in the 50 States

Abinash “Nash” Mohanty
Sirin Yaemsiri
Joanie Lofgren

Applied Research and Methods
U.S. Government Accountability Office (GAO)

FCSM 2023 Research and Policy Conference

(Based on GAO Report: GAO-23-105996)

Why GAO Did This Study

- The Infrastructure Investment and Jobs Act includes a provision for GAO to study the extent of open recalls in passenger vehicles used for ridesourcing
- Congress requested national and state estimates of open recalls among rideshare vehicles

Overview

- GAO obtained confidential data from 2 ridesourcing companies
- Recall information is available from CARFAX and NHTSA.
- GAO used the CARFAX tool for about 98 percent of the vehicles. For the remainder, GAO used a sampling method and the NHTSA search tool

What is a rideshare company

- “Ridesourcing,” also called ridesharing, involves transportation network companies using a digital network to connect passengers with drivers of, most commonly, personally owned vehicles
- For confidentiality reasons we are not going to name individual manufacturers or the rideshare companies

Recalls flagged by Vehicle Manufacturers and the National Highway Transportation and Safety Administration (NHSTA)

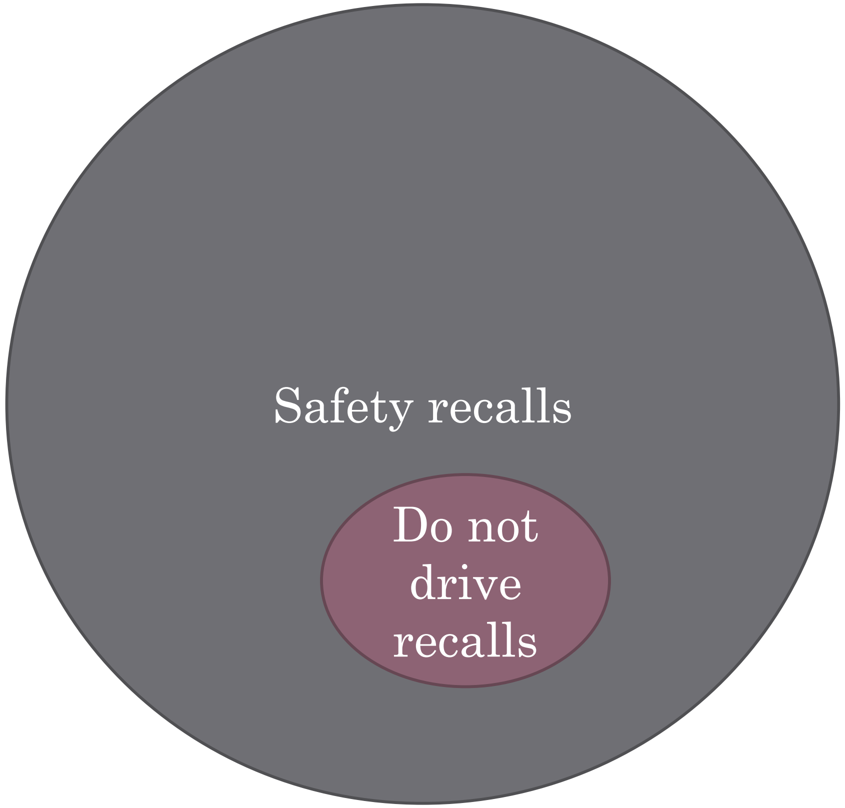
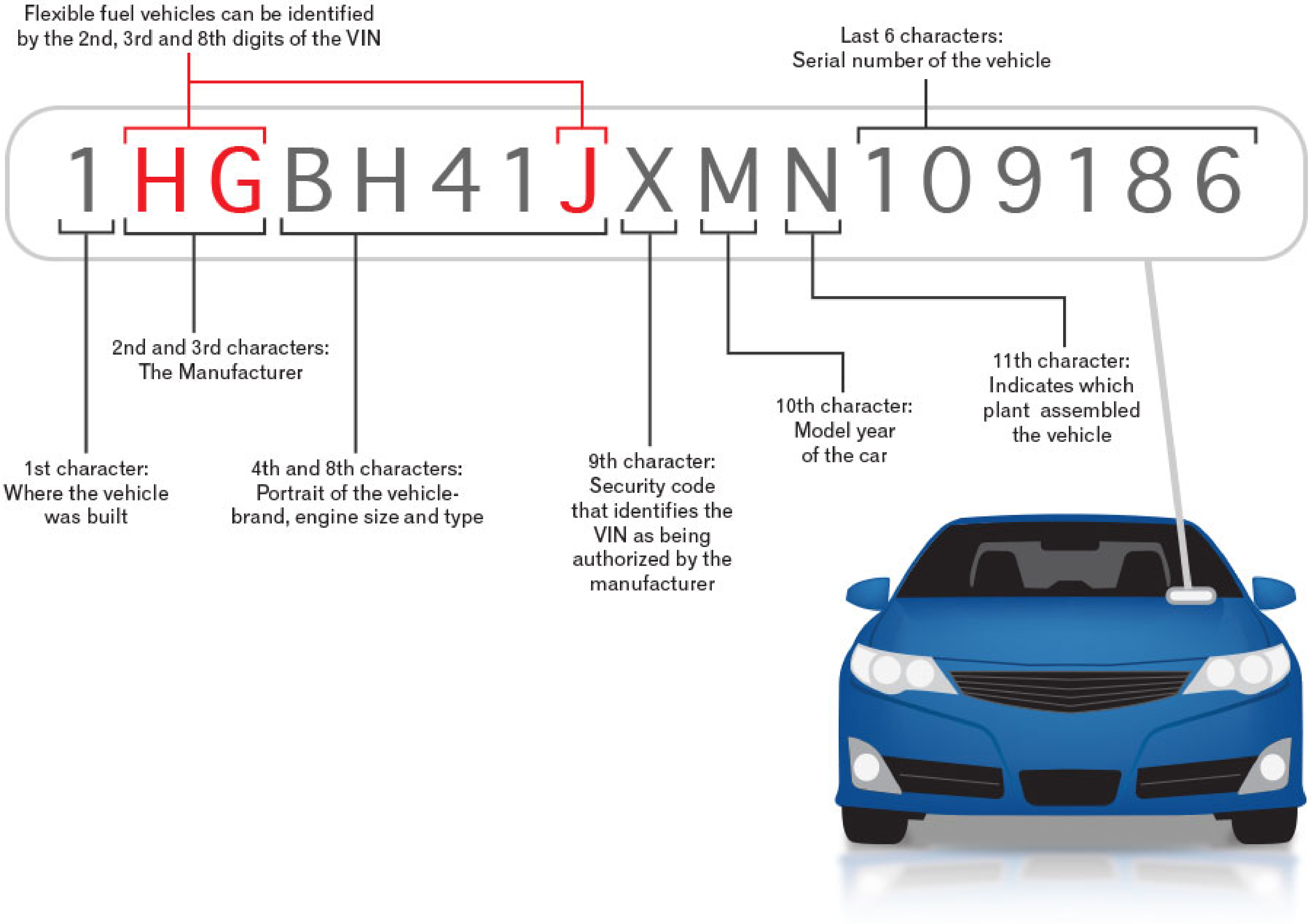


Diagram not to scale



Flexible fuel vehicles can be identified by the 2nd, 3rd and 8th digits of the VIN

Last 6 characters: Serial number of the vehicle

1HGBH41JXMN109186

1st character: Where the vehicle was built

2nd and 3rd characters: The Manufacturer

4th and 8th characters: Portrait of the vehicle- brand, engine size and type

9th character: Security code that identifies the VIN as being authorized by the manufacturer

10th character: Model year of the car

11th character: Indicates which plant assembled the vehicle

Comparing CARFAX and NHSTA recall search tools

CARFAX Vehicle Recall Search Service (VRSS)

- CARFAX, in a joint effort with the Alliance for Automotive Innovation, created a recall search tool available to approved private businesses and government entities at no cost.
- Some manufacturers do not participate.
- CARFAX VRSS API allows an approved user to search recall information provided by participating manufacturers for up to **10,000 VINs** at a time.

NHTSA

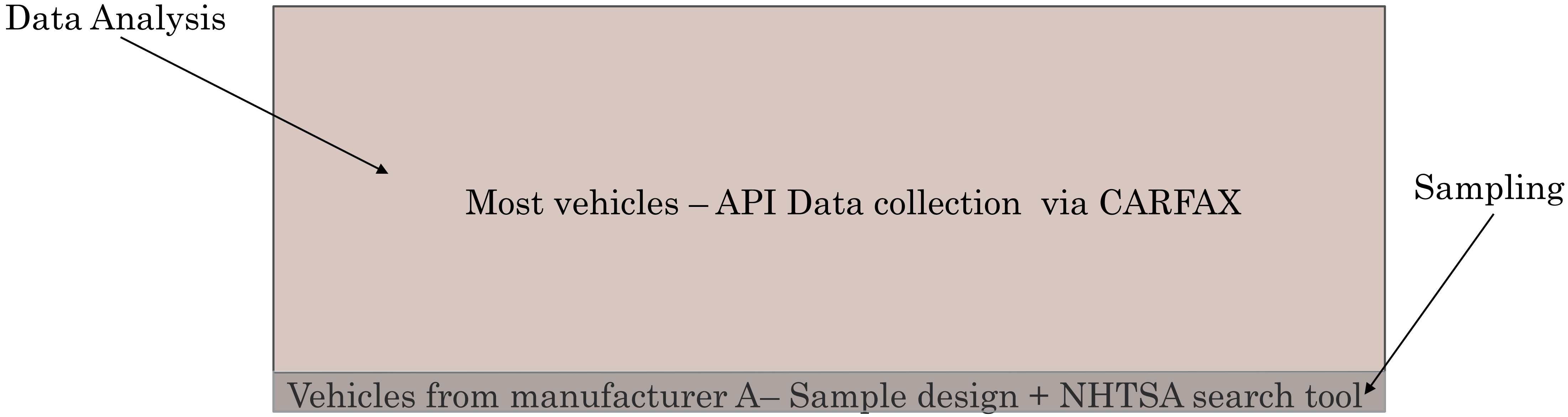
- The recall search tool queries the manufacturer-maintained data and returns information on any open safety recalls on the vehicle that began within the previous 15 years.
- NHTSA requires that **nearly all** manufacturers maintain data on vehicles subject to recalls.
- Only **one** VIN can be searched in the tool at a time

Challenges

1. Deduplication
   - Some vehicles were shared between the 2 rideshare companies
2. CARFAX did not have the data from one manufacturer
   - Some manufactures (less than 2% of VINs) were not included in the CARFAX VRSS
     - Manufacturer A's VINs were an overwhelming majority of these cases, so the others were dropped from our analysis
   - NHSTA tool has the recall information for the manufacturers (including Manufacturer A) missing from the CARFAX tool
3. Resources
   - It takes 2 minutes to retrieve recall information for a single VIN using the NHSTA search tool. Inputting all of Manufacturer A's VINs could take hundreds of hours
   - Even as stratified random sample with each state as a stratum, inputting thousands of VINs would have taken too long
   - Sampling with fewer strata was chosen as a way to overcome time restraints

API Data Collection

- Used CARFAX API to gather recall data on all non-Manufacturer A VINs
- Aggregated the recall status of VINs at the state and national levels to get national and state level recall statistics for non-Manufacturer A VINs



Data Analysis

Most vehicles – API Data collection via CARFAX

Sampling

Vehicles from manufacturer A– Sample design + NHTSA search tool

Diagram not to scale

Sample Design

- Simple random sample without stratification would not provide reliable state level recall rates
- Did not have the resources to stratify by state
- 3 Sampled Strata
  - Sorted the 51 states by percent of rideshare vehicles that were Manufacturer A vehicles in each state
  - Placed the states into tertiles using the number of Manufacturer A vehicles in each state, keeping each state in 1 tertile exclusively
    - Ensured the smaller states are represented in the sample
    - The percentage of Manufacturer A rideshare vehicles is likely associated with the presence of Manufacturer A repair availability
- Sample size needed for a simple random sample that produces estimates with a 10% MOE at the tertile level was taken from each tertile

National Level Estimates of Rideshare Vehicle Recalls

- Stratified random sample with 4 strata (3 sampled strata + 1 certainty strata)
- The vast majority of the recall data acquired through the CARFAX API consists of the 1 certainty strata
- Only base weights were applied to our 3 sampled strata because there was very little non-response (invalid VINs) that was addressed by inflating our sample size slightly

National Level Results

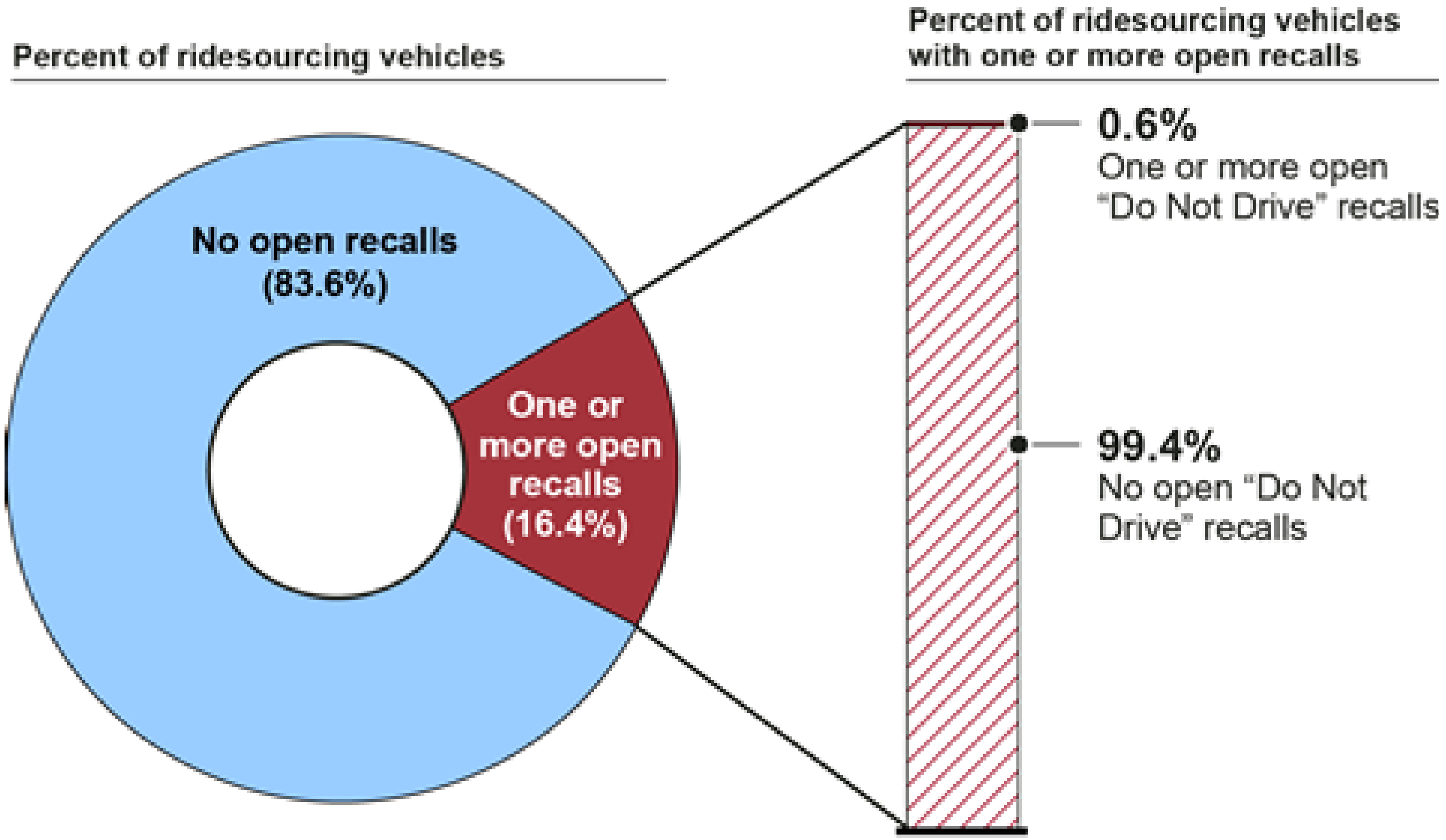


Figure 3 from GAO-23-105996 Ridesharing Vehicle Recalls

Data for ridesourcing vehicles that performed a passenger trip in August 2022. Vehicles' recall status from December 2022. All estimates in this figure have a margin of error within plus or minus 0.4 percentage points at the 95 percent confidence level.

State Level Estimates

- Indirect estimation method was used to produce recall statistics for the cars that were not in the CARFAX VRSS
- Used each strata's sample to calculate the approximate percent of Manufacturer A rideshare vehicles with an open recall in each tertile
- Multiply the tertile-level estimated percent of Manufacturer A rideshare vehicles with an open recall to the number of Manufacturer A rideshare vehicles in each state that are within the tertile to estimate the number of Manufacturer A rideshare vehicles that have an open recall in each state
- Add each state's estimated number of Manufacturer A vehicles with an open recall to each state's CARFAX VRSS derived overall number of rideshare vehicles with an open recall to get a total number of rideshare vehicles with an open recall in each state

How we created a confidence bounds for our State level estimates

- Our sample's results can be reported at a 95 percent confidence interval at the tertile level

- But we reported state level statistics….
  - Noting this we make the upper and lower bounds the total number of manufacturer A rideshare vehicles in the state
  - This produces a 100 percent confidence interval for each state because it captures all sampling error

State Level Results

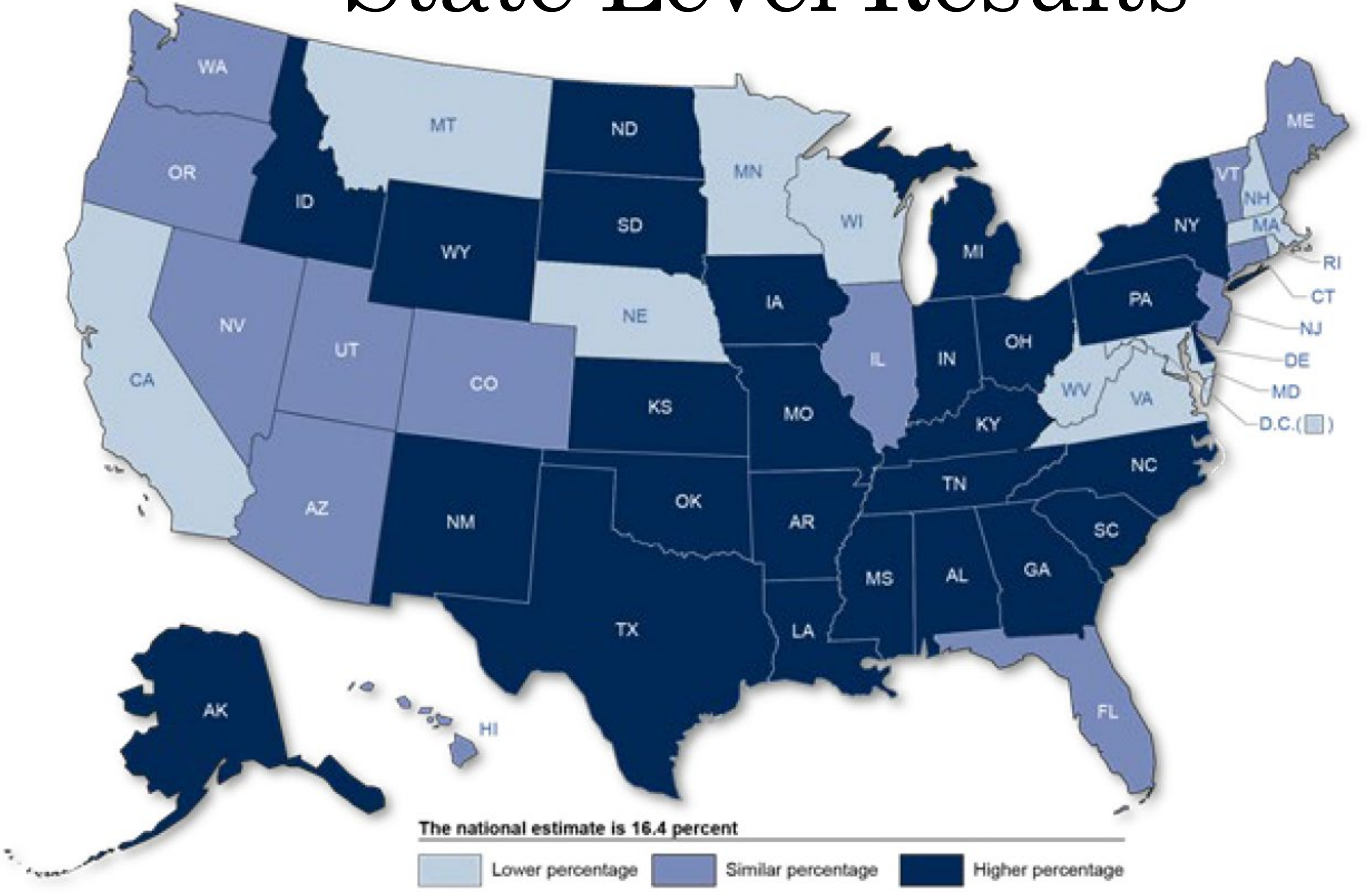


Figure 1 from GAO-23-105996 Ridesharing Vehicle Recalls

Data for ridesourcing vehicles that performed a passenger trip in August 2022. Vehicles' recall status from December 2022.

Final Thoughts

- GAO estimated that nationally, nearly 1 in 6, or about 16 percent, of ridesourcing vehicles that performed a passenger trip in August 2022 had an open safety recall as of December 2022. While CARFAX, reported about 1 in 5, or about 20 percent, of passenger vehicles nationally had an open safety recall in 2022.
- Consider the limitations of your administrative data before forging ahead
- Think of ways that sampling can overcome time and resource restraints
- Don’t feel bound by traditional estimation methods, indirect estimation strategies can sometimes allow you to combine sampled data with administrative data with fewer resources