



NCHS Data Linkage Program

Incorporating Artificial Intelligence (AI) and Machine Learning (ML) into the Data Linkage Workflow

Cordell Golden

Chief, Data Linkage Methodology and Analysis Branch
National Center for Health Statistics
Centers for Disease Control and Prevention

AI Day for Federal Statistics

April 30, 2026

National Center for Health Statistics



Mission

We collect, analyze, and disseminate timely, relevant, and accurate health data and statistics. Our products and services inform the public and guide program and policy decisions to improve our nation's health.



Vision

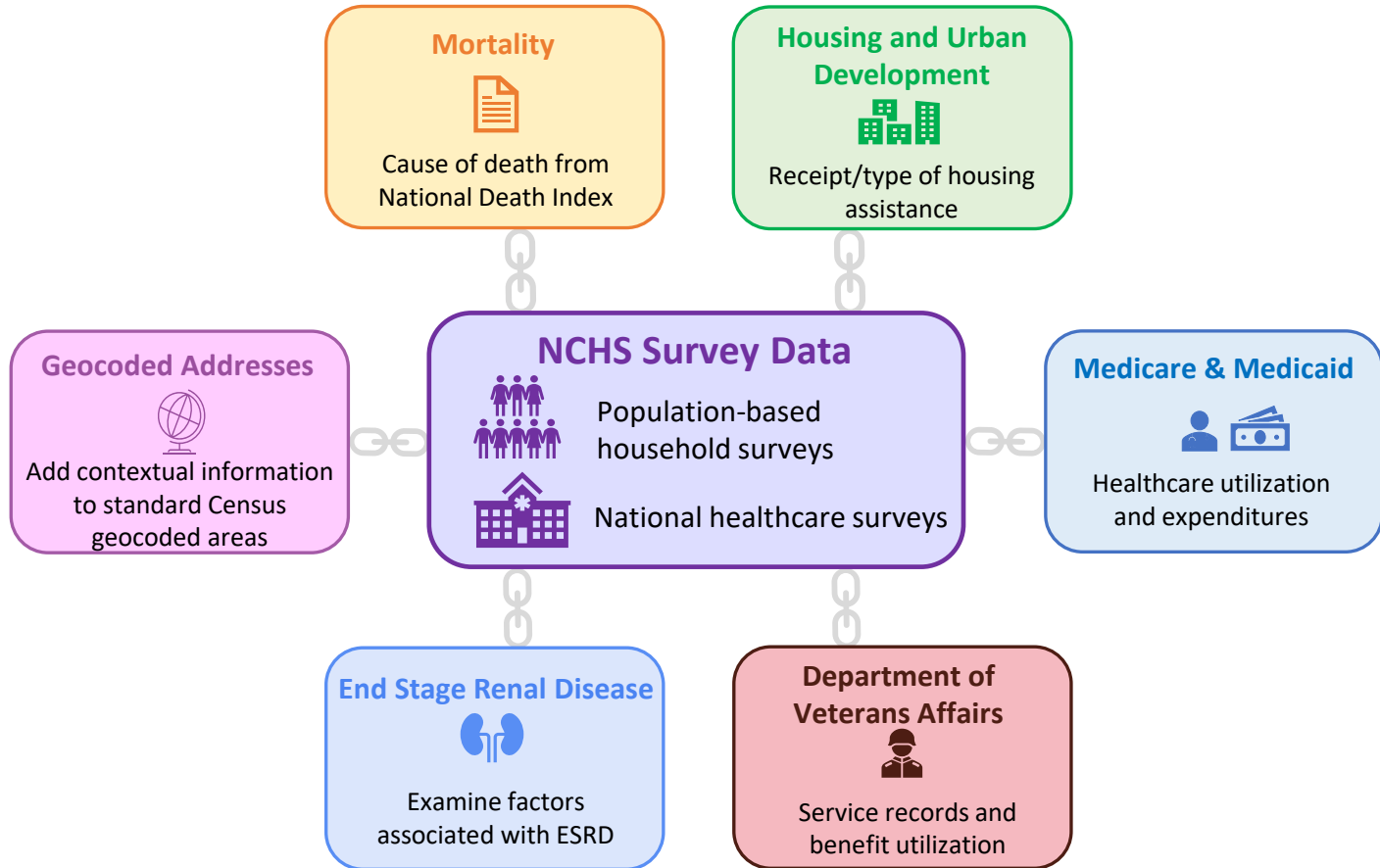
To be a world-class innovator and provider of health data and statistics

NCHS Data Linkage Program

- Create linked data files that support high quality research and program evaluation
- Utilize state of the art linkage methodologies and provide documentation and support for analyzing linked data files
- Explore innovative methods for improving linkage accuracy and efficiency and maintaining researcher access to linked data



NCHS Linkages



AI/ML Resources at CDC



For decades, our agency has pioneered the use of advanced analytics and predictive modeling in public health, from flu forecasting to outbreak detection. The scale and speed of modern AI capabilities continue to transform the way we work to advance public health and protect Americans from health threats. By harnessing AI, we aim to empower staff to effectively use AI to reduce administrative burden, increase efficiency, and drive operational excellence.

AI is a tool that may be appropriate for some, but not all, projects at CDC. This website provides guidance and resources for teams considering using AI in their projects.

AI Guidance & Tools

CDC's AI Strategy

CDC's blueprint for harnessing AI to accelerate disease detection, reduce operational burden, and empower innovation through secure data platforms and an AI-ready workforce.

Generative AI Information and Guidance

Learn about CDC's GenAI capabilities and guidance for use.

CDC's GenAI-powered tools

CDC's GenAI-powered tools can help with tasks such as assisting software development, answering questions, summarizing text, editing or proofreading content, conducting research and data analysis, and improving accessibility.

Glossary

- **Artificial Intelligence (AI):** A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action. Source: [15 U.S.C. 9401\(3\)](#) and [EO 14110](#).
- **Machine Learning (ML):** A set of techniques that can be used to train AI algorithms to improve performance at a task based on data. Source: [EO 14110](#).

General Use / For Everyone

DO

DO NOT

Benefits and Risks of GenAI

Benefits



Increase efficiency of manual or repetitive tasks, such as:

- Synthesizing and summarizing information.
- Drafting and designing communication products for various reading levels, platforms, and audiences, and in various languages.
- Generating and translating programming code.
- Analyzing and deriving insights from data.

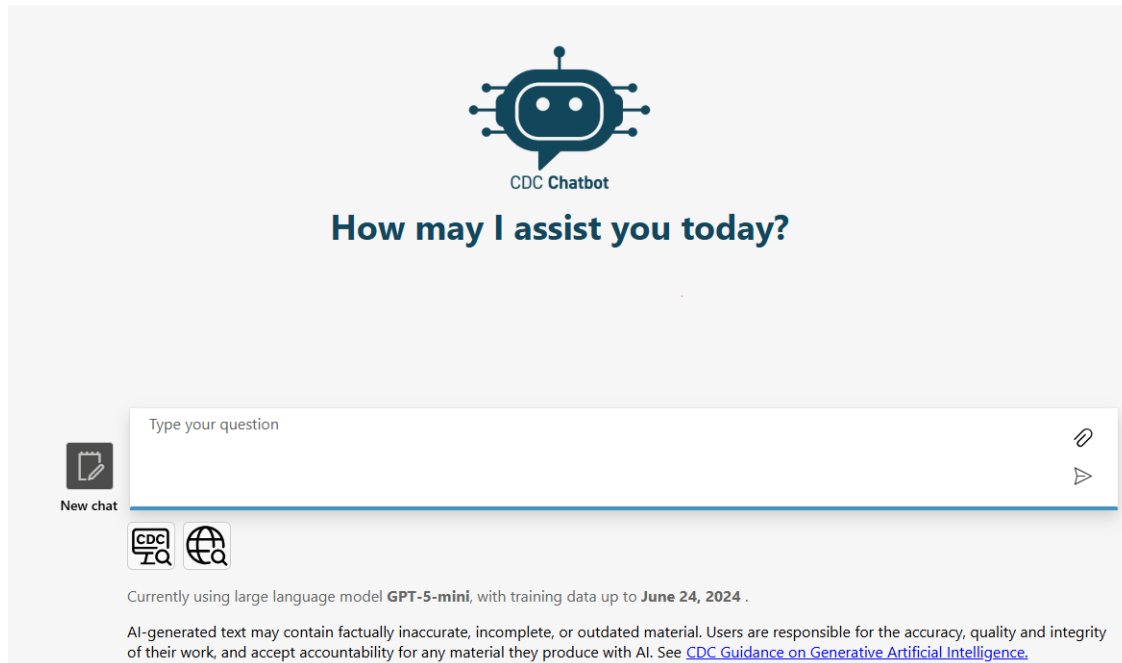
Risks




- May provide fake or false information (sometimes referred to as "hallucinations").
- Could lead to breaches of scientific integrity and quality, among other potential vulnerabilities, if usage of AI-generated material isn't reviewed by a human and corrected where necessary. (No Title)
- May reflect or amplify biases present in a given model's training data in the generated output.
- Possible privacy and data security risks if sensitive data is passed to public interfaces.
- Possible legal or ethical risks, including risks to agency credibility.

Leveraging AI/ML to Improve Data Linkage Workflows

- *How can AI and ML be used to improve the data linkage workflow?*
 - *Let's ask the CDC Chatbot*








CDC Chatbot

How may I assist you today?

Type your question



New chat

Currently using large language model **GPT-5-mini**, with training data up to **June 24, 2024** .

AI-generated text may contain factually inaccurate, incomplete, or outdated material. Users are responsible for the accuracy, quality and integrity of their work, and accept accountability for any material they produce with AI. See [CDC Guidance on Generative Artificial Intelligence](#).

Leveraging AI/ML to Improve Data Linkage Workflows

You

How can artificial intelligence and machine learning be used to improve the data linkage workflow?

Overview

Artificial intelligence (AI) and machine learning (ML) can substantially improve a data linkage (record linkage, entity resolution) workflow by automating and enhancing key steps: blocking/indexing, pairwise comparison, similarity scoring, classification, clustering, active learning, and evaluation. Machine learning can increase accuracy, speed, and scalability while reducing manual effort and repeated tuning.

Below are practical ways ML/AI can be applied, with methods, examples, benefits, and limitations.

Key stages and ML/AI techniques

• Data preprocessing and standardization

- Use ML-based name- and address-parsing models (CRFs, neural sequence models) to extract components (given name, surname, street, unit).
- Use learned embeddings for normalization of text (e.g., character-level or subword models) to handle misspellings and variants.
- Benefit: more robust canonicalization than rule-only normalization.

• Blocking / indexing (to reduce candidate pairs)

- Learn blocking functions using supervised methods or representation learning:
 - Locality-sensitive hashing (LSH) on learned embeddings.
 - Sorted neighborhood or canopy clustering with thresholds learned from data.
 - Learned blocking (e.g., using decision trees or neural models that predict block membership).
- Benefit: reduces comparisons while preserving high recall; ML can optimize tradeoff between efficiency and recall.

• Privacy-preserving linkage

- Use privacy-preserving techniques with ML:
 - Bloom filter encoding and ML classifiers operating on encodings.
 - Secure multi-party computation (SMPC) and homomorphic encryption combined with ML-compatible models.
 - Differentially private model training when releasing models or outputs.
- Benefit: enables linkage across sensitive datasets with reduced disclosure risk.

Motivation for NCHS Use Cases

- Record linkage enables survey data to be integrated with other data sources, expanding the analytic potential of both sources
- Depending on the quality and size of the data sources being linked, the processing time can be prohibitive
- AI and ML-based tools can be incorporated into data linkage workflows to improve efficiency with relatively high accuracy while also preserving privacy

Exploring Use of AI Models in Data Preprocessing

McCarty, F. *Using Natural Language Processing (NLP) in Data Linkage*. 2025 Joint Statistical Meetings

Objective:

Examine the use of Natural Language Processing (NLP) models compared to simple rule-based approaches for identifying non-name text in name fields

Data Source:

- Test data created using R *randomNames* package
- Valid first and last names (n=9,949)
- Supplemented with non-name text (e.g., “pilot study”, “department funded”) (n= 166)

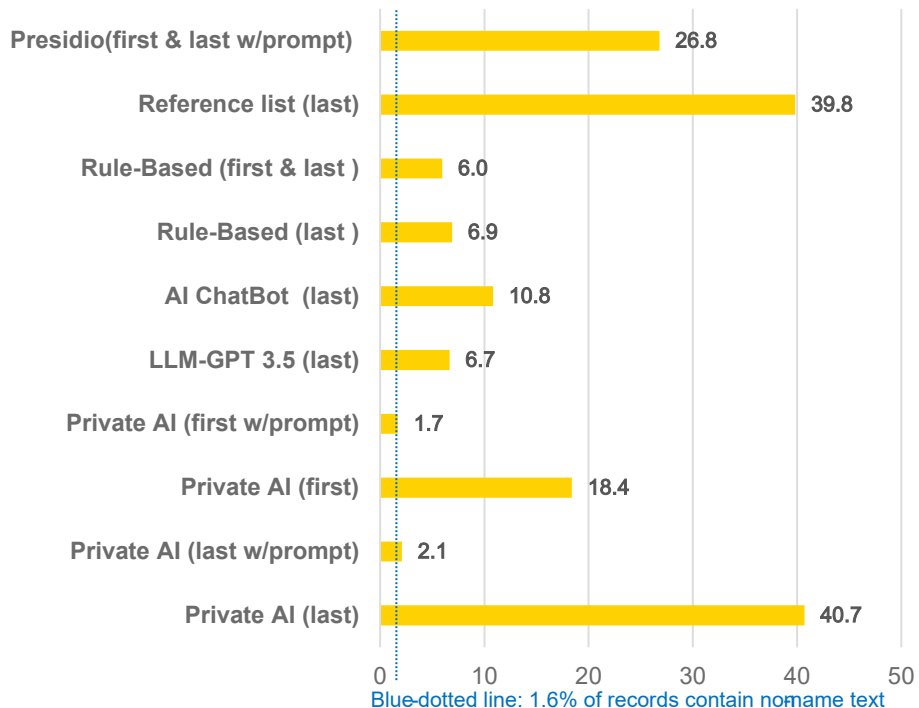
Approaches Evaluated:

- **Named Entity Recognition (NER):** Private AI
- **Large Language Models (LLMs):** GPT-3.5 with few-shot prompting; CDC Chatbot
- **PHI/PII detection-and-anonymization toolkit:** Presidio
- **Rules-based** approached based on number of words and characters
- Comparison with **reference list** of “valid” last names

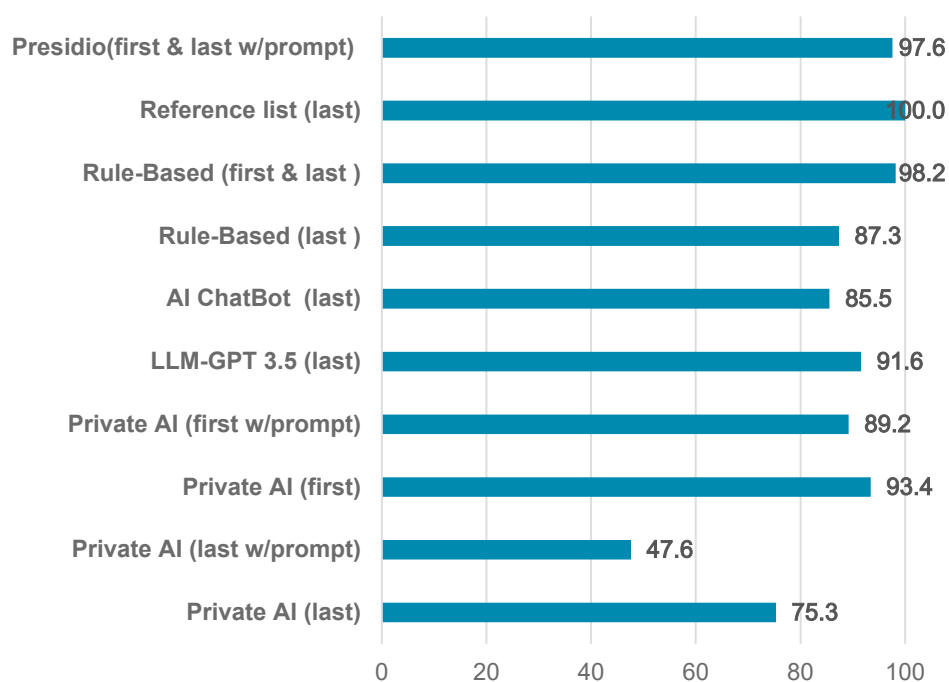
Exploring Use of AI Models in Data Preprocessing

Results

Percent of Records Flagged for Review



Percent of Non-name Records Correctly Flagged for Review



Using ML to Improve Linkage Efficiency - Blocking

Campbell, S.R., D.M. Resnick, C.S. Cox and L.B. Mirel, *Using Supervised Machine Learning to Identify Efficient Blocking Schemes for Record Linkage*. Stat J IAOS, 2021. 37(2): p. 673-680.

Objective:

Examine how a supervised ML algorithm, the Sequential Coverage Algorithm (SCA), can be used to identify efficient blocking schemes for linking large datasets. Compare SCA to traditional blocking methods.

Data Sources:

- 2016 National Hospital Care Survey (n = 5.6 million)
- CMS Medicare Enrollment Database (EDB) (n = 84.6 million)

Methods:

- **Variables available in both datasets:** SSN, Medicare number, name, DOB, Zip code and state of residence
- **Deterministic linkage:** Match on SSN or Medicare number and majority of other non-missing variables used as “*truth deck*” for training/supervising. (n = 1.6 million)
- **Comparison of results from:**
 - Ad-hoc blocking methods – Requires judgment of SME and result in extended processing time
 - Modified version of SCA (supervised ML algorithm) - Designed to learn a set of efficient blocking keys

Using ML to Improve Linkage Efficiency - Blocking

Results

Table 2

Evaluation of efficiency of ad-hoc blocking key previously used by the National Center for Health Statistics (NCHS)

Blocking key	Total cross product (in trillions)	Potential links to be evaluated	Total matches from truth deck	Captured truth deck records
Last 4-digits of SSN or HICN, month birth, day birth, sex	475.8	11,381,076	1,598,511	1,546,710 96.8%

Note: Potential links generated using 2016 National Hospital Care Survey (NHCS) and the Centers for Medicare and Medicaid Services (CMS) Medicare Enrollment Database (EDB). Note: Social Security Number (SSN) is a unique identifier assigned by the Social Security Administration (SSA). Note: Health Insurance Claim Number (HICN) is a unique identifier assigned by CMS.

Table 3

Sequential Coverage Algorithm (SCA) learned blocking scheme in the linkage of the 2016 NHCS to the CMS Medicare EDB

Blocking key	Number of block variables	Number of potential links	Number of truth deck records	Percent of truth deck	Cumulative percent of truth deck coverage
Day of birth, month birth, year birth, ZIP code residence	4	2,288,489	1,167,469	73.0	73.0
First name, last name, month birth, year birth	4	2,314,896	1,137,532	71.1	91.6
First name, day birth, month birth, year birth, sex	5	46,436,412	1,236,422	77.3	94.1
Last name, day birth, month birth, year birth, sex	5	4,538,784	1,363,757	85.3	97.9
First name, last name, state residence	3	39,035,154	1,071,595	67.0	98.6
Middle initial, day birth, month birth, year birth, state residence, sex	6	7,235,868	725,609	45.4	98.8

Note: Potential links generated using 2016 National Hospital Care Survey (NHCS) and the Centers for Medicare and Medicaid Services (CMS) Medicare Enrollment Database (EDB).

Using AI/ML in Privacy-Preserving Linkages

Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies (NCSES/NCHS)
(<https://www.americasdatahub.org/award-pprl1/>)

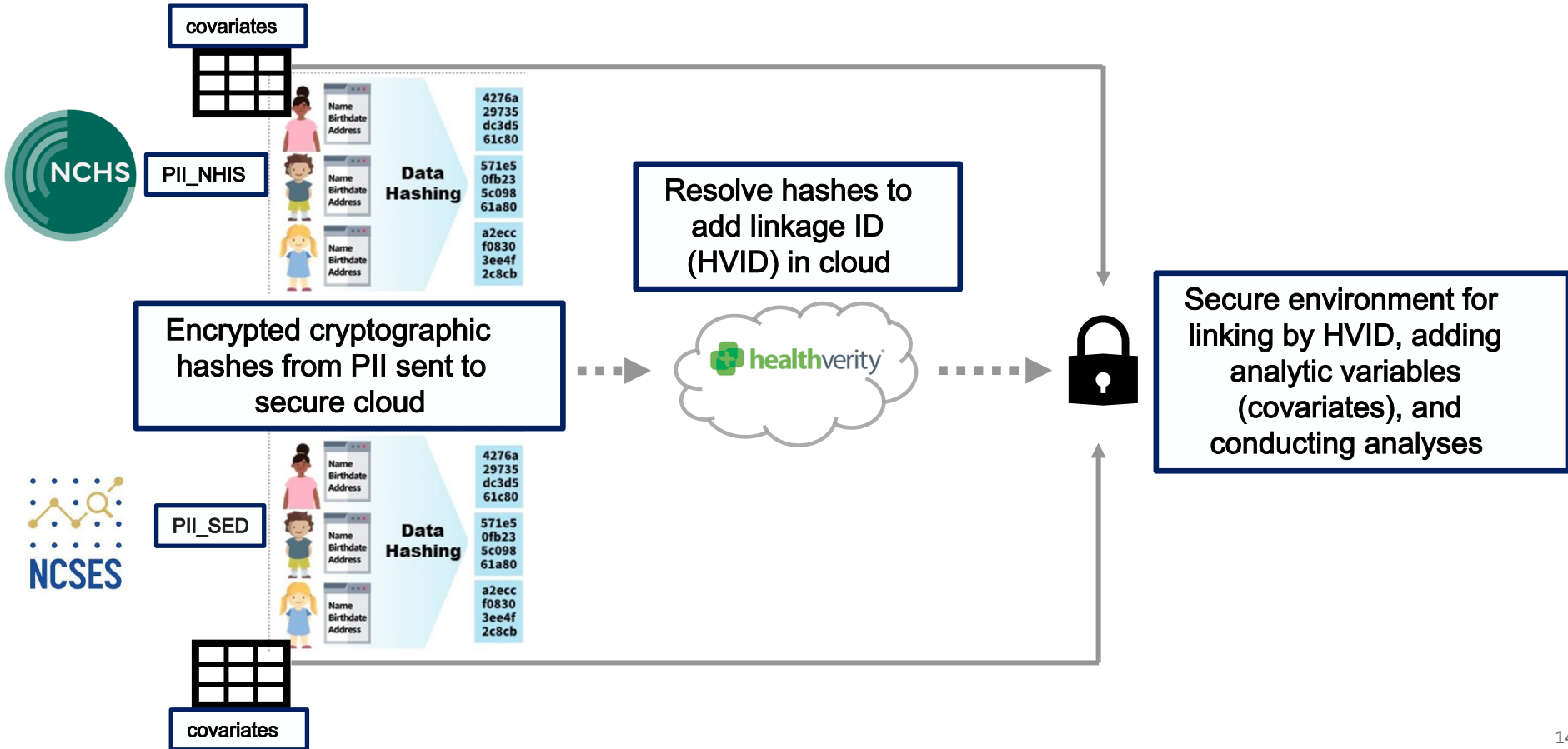
Project Overview:

- Collaboration with the National Center for Science and Engineering Statistics (NCSES)
- Proof of concept for establishing National Secure Data Services data linkage services
- Develop data sharing agreement to link data from two federal statistical agencies
- Establish processes for utilizing commercial PPRL (*HealthVerity*) tool
 - PPRL tool incorporates AI and ML in linkage process
- Resulted in successful linkage of CIPSEA-protected data

Data Sources:

- 2012-2022 National Health Interview Survey – NHIS (NCHS)
- 2012-2022 Survey of Earned Doctorates – SED (NSCES)

NCHS-NCSES PPRL: Data Flow



AI/ML Components: HealthVerity PPRL

Probabilistic Matching

- Multi-target Tracking (MTT)
 - Each piece of data is treated as a noisy observation in a noisy environment
 - All data considered together, without setting a hard threshold at a single data point

Migration Models

- Zero-shot Learning Model
 - People movement/relocation is not random
 - Continuously learns and updates to account for people movement

Supervised vs Unsupervised

- Initially built using supervised data
- Supervised data stripped and migrated to unsupervised model (Expectation-Maximization)

**Data used for this project were isolated from previously trained models and were not used to further train the models ~~Health~~HealthVerity.*

Discussion

- AI and ML can be incorporated to data linkage workflows to improve linkage quality, efficiency, and accuracy
- Data linkages often require use of sensitive data
 - But AI and ML models may retain information to improve performance
- Considerations:
 - Regulations governing the data being linked
 - Limitations of the computing environment
- Transparency and adherence to standards/best practices are essential to ensuring trustworthy application of AI and ML

Thank you!

Cordell Golden

Chief, Data Linkage Methodology and Analysis Branch
National Center for Health Statistics
Centers for Disease Control and Prevention
CGolden@cdc.gov



NCHS Data Linkage Program website:

<https://www.cdc.gov/nchs/linked-data/about/index.html>

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

