NCHS Data Linkage Program: Incorporating Artificial Intelligence (AI) and Machine Learning (ML) into the Data Linkage Life Cycle

Cordell Golden
Director, Data Linkage Program
National Center for Health Statistics
Centers for Disease Control and Prevention

AI Day for Federal Statistics
May 2, 2024

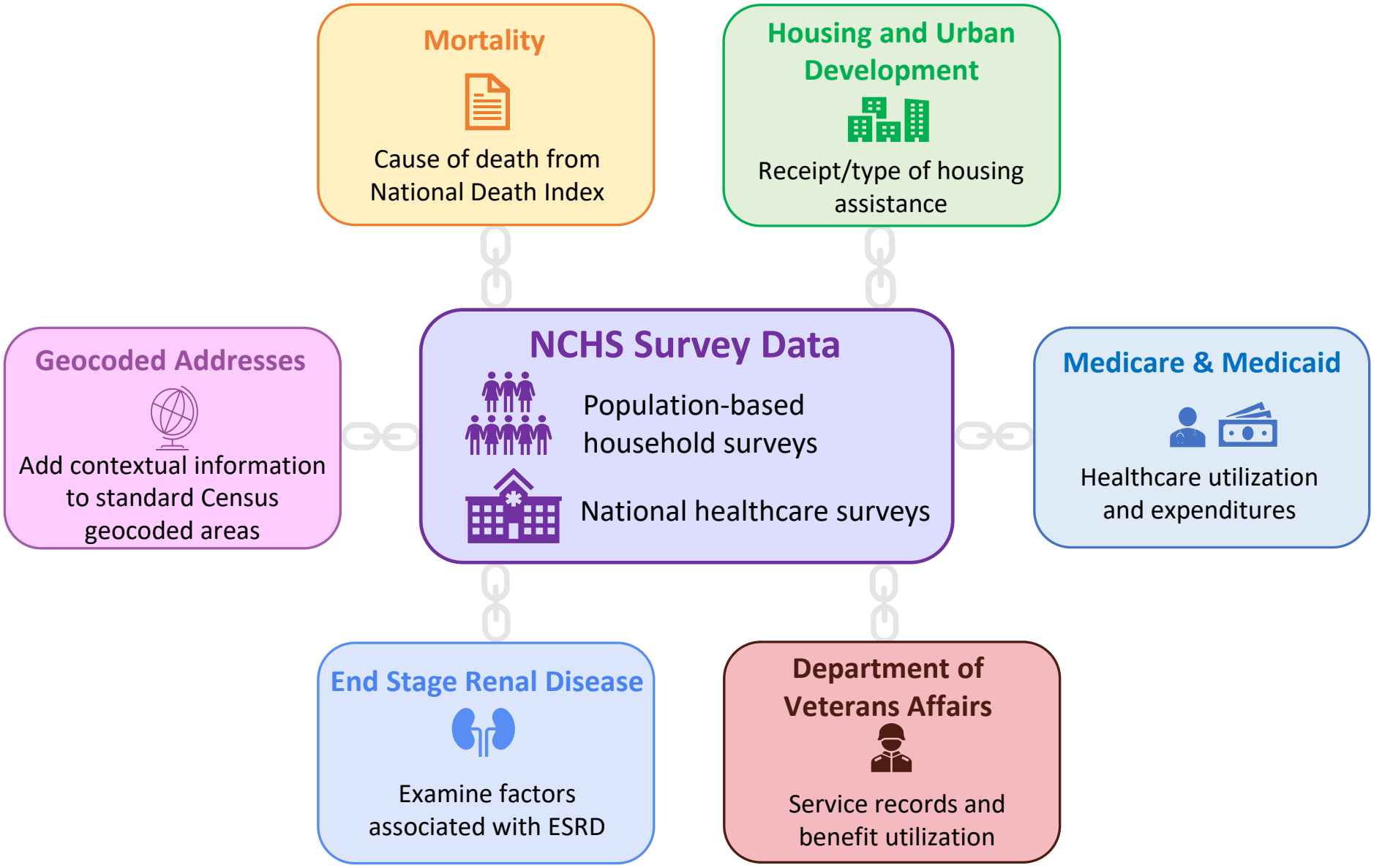
National Center for Health Statistics (NCHS)

- Principal health statistics agency in U.S.
- One of 13 principal federal statistical agencies
- **Mission:** To provide timely, relevant, and accurate health data and statistics that inform and guide programs and policies to improve our nation's health

NCHS Data Linkage Program: Overview

- 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 maintaining researcher access to linked data

NCHS Linkages



AI/ML Initiatives within Federal Government

Initiatives promoting the trustworthy use of AI and ML in the Federal Government:

- AI in Government Act of 2020
- Advancing American AI Act
- Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
- Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government
- OMB Memo M-24-10 on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence

AI/ML Resources at NCHS

Artificial Intelligence (AI) offers tremendous potential for solving complex public health challenges. AI technologies can make processing and analyzing large amounts of complex data more accessible to develop rapid, actionable insights to inform public health action and interventions. CDC is committed to deploying responsible AI tools — developed using best practices, grounded in fairness and health equity principles and trusted by the American public — while managing risks.

Artificial intelligence is a tool that can be considered for some projects, not all, at CDC. This website provides guidance and resources for teams considering the use of AI in their projects.

Basic Terms

Artificial intelligence applied technology to make computers do things that normally require human intelligence. Most AI is based on machine learning.

Machine learning (ML) allows a computer to analyze data to do a task without being explicitly programmed, for example, to find patterns or predict an output based on a set of inputs.

Generative AI is a subset of AI technology that can create content, including text, images, audio or video, when prompted by a user.

Benefits and Risks of GenAI

Benefits	Risks

DO	DO NOT

Generative AI Information and Guidance

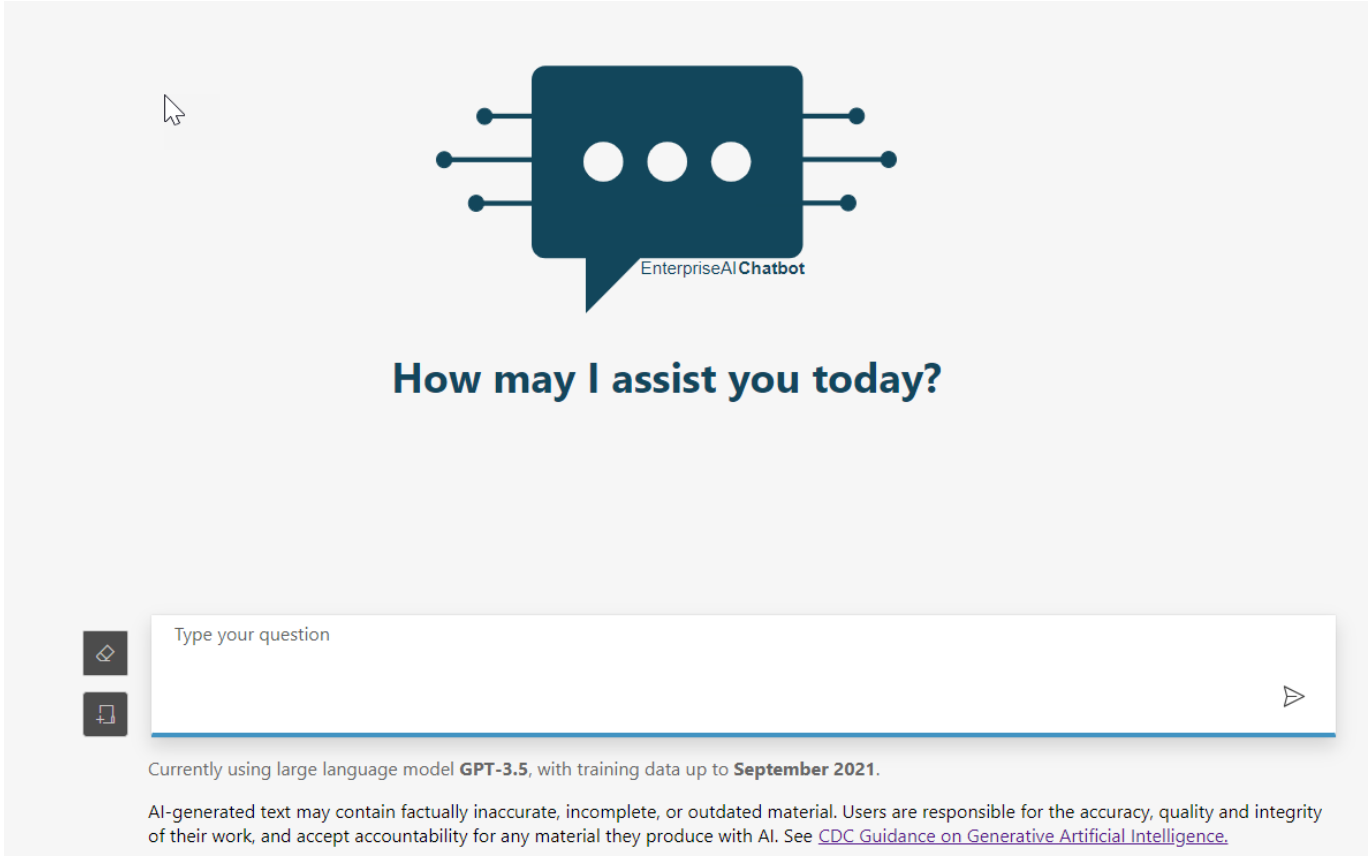
Print

NCHS Artificial Intelligence (AI) and Machine Learning (ML) Release Standards

NCHS AI and ML standards for development of any artificial intelligence or machine learning project. These standards cover the release of a model, or reports that rely heavily on an AI and ML model, in order for the Center to remain transparent in how a model is used, as well as its strengths and limitations

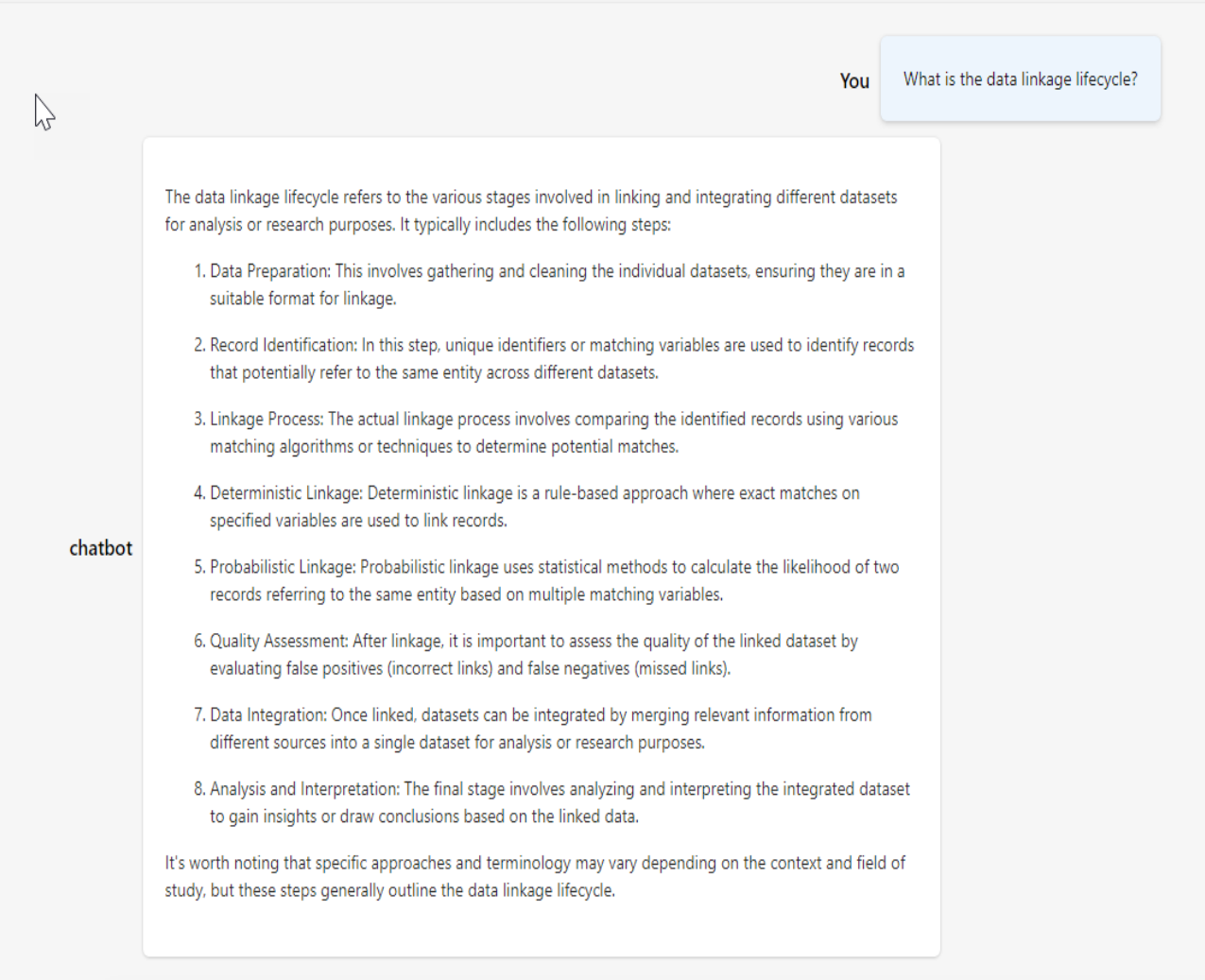
Leveraging AI/ML for Data Linkage Life Cycle

- *How can AI and ML be leveraged to enhance the data linkage life cycle?*
  - *Let's ask the AI Chatbot (cdc.gov)*



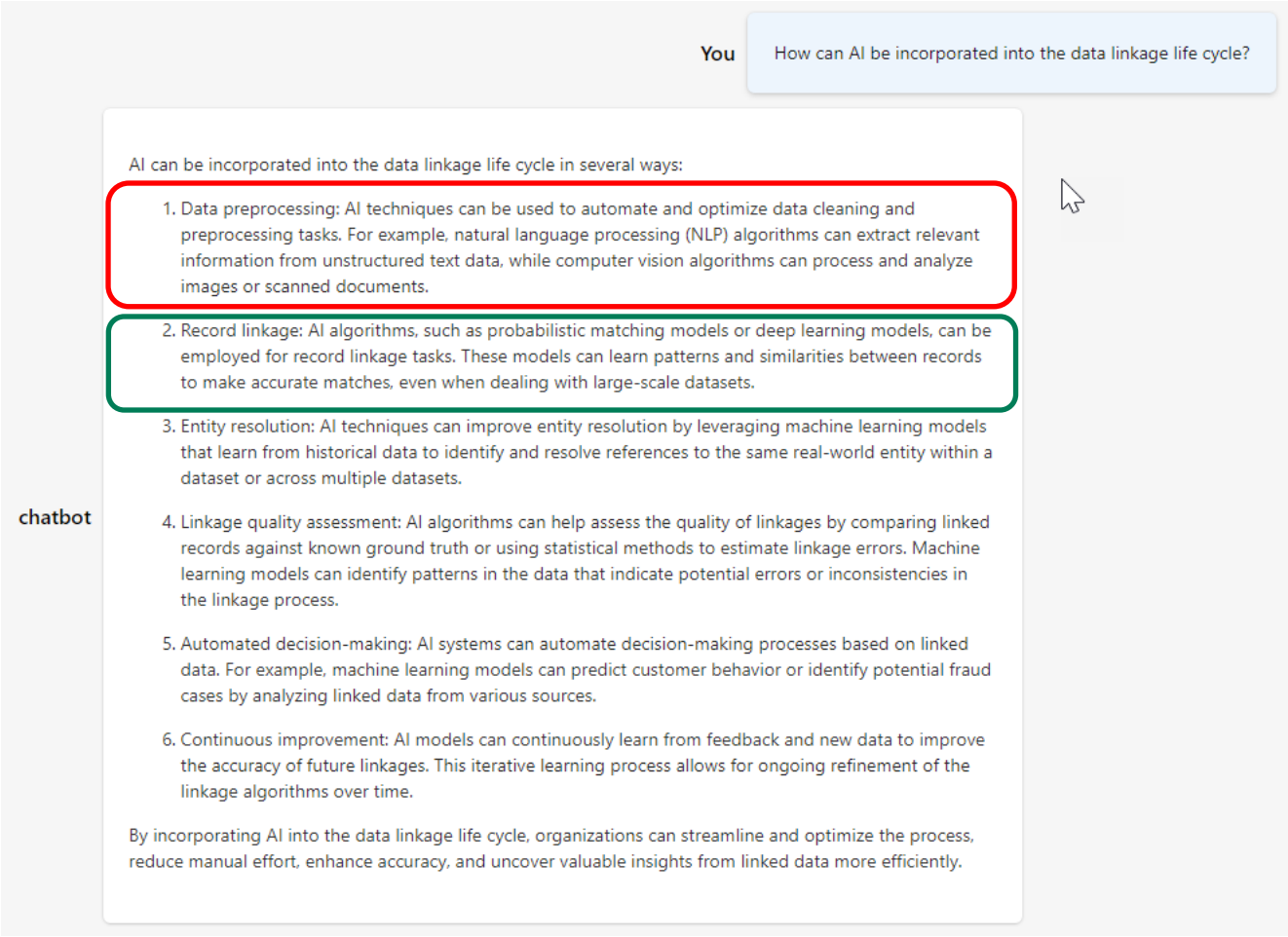
Data Linkage Life Cycle

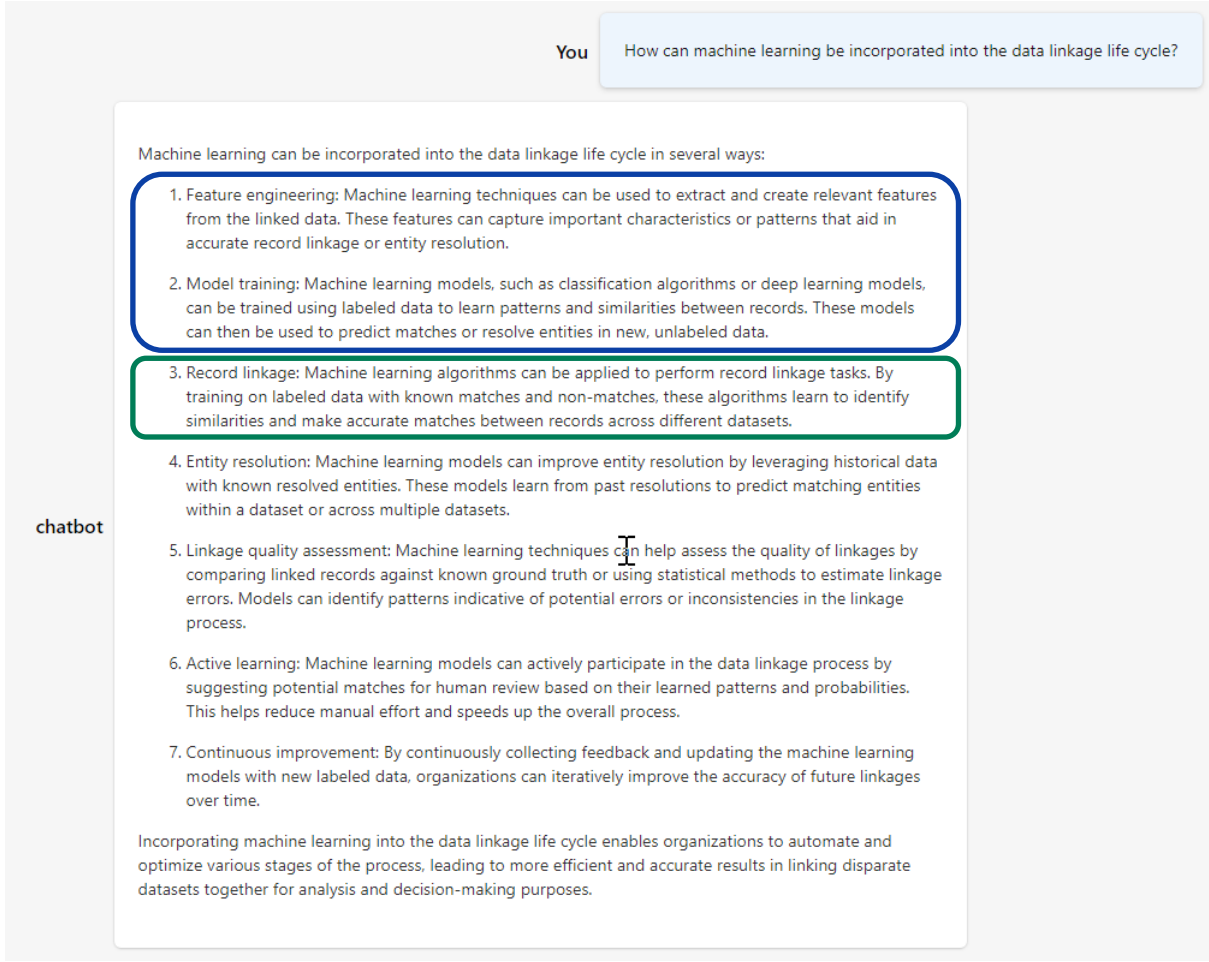
Litmus test



- ✓ Data preparation
- ✓ Record Identification
- ✓ Linkage Process
- ✓ Deterministic Linkage
- ✓ Probabilistic Linkage
- ✓ Quality Assessment
- ✓ Data Integration
- ✓ Analysis and Interpretation

Incorporating AI and ML in Data Linkage Life Cycle





Using AI to Enhance Data Preprocessing

1. Data preprocessing: AI techniques can be used to automate and optimize data cleaning and preprocessing tasks. For example, natural language processing (NLP) algorithms can extract relevant information from unstructured text data, while computer vision algorithms can process and analyze images or scanned documents.

- Data linkage accuracy depends on the quality of the data fields
- Automating the data preprocessing improves efficiency and linkage accuracy
- Type of data field (date vs name) affects the level of effort needed for review
  - Simple rules can be applied to date fields
  - Name fields may contain non-name text that needs to be identified and removed

Using AI to Enhance Data Preprocessing

What's in a Name Field?

Frances McCarty[1], Ben Rogers[2], Jessie Parker[1], Cordell Golden[1]
National Center for Health Statistics, [1]Division of Analysis and Epidemiology, Data Linkage Methodology and Analysis Branch
[2]Divsion of Research Methodology, Office of the Director

Objective:

Examine the use of AI-based large language models (LLM) compared to simple rule-based approaches for identifying non-name text in name fields

Data source:

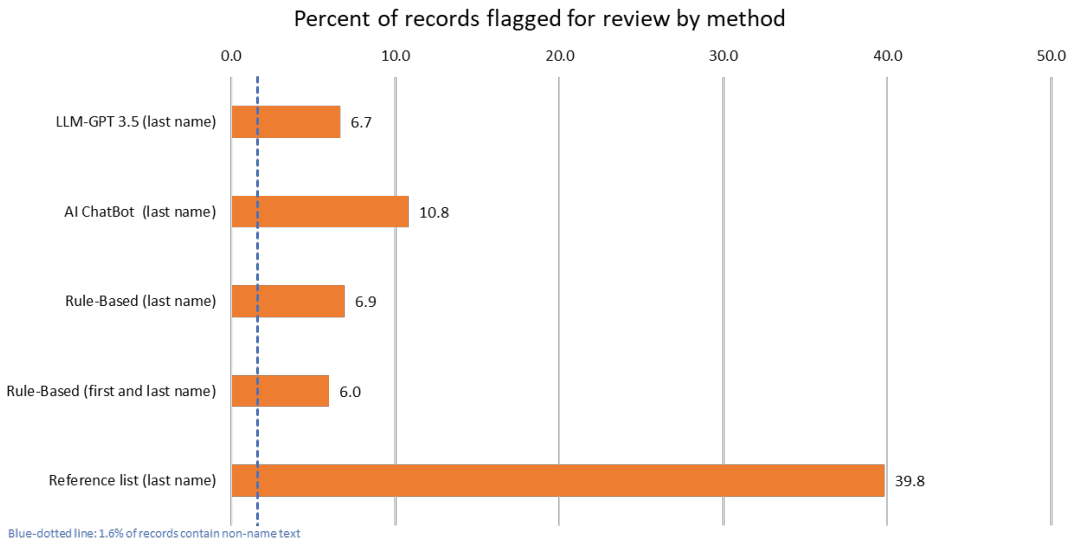
- Test data created using R (*randomNames* package)
- Valid first and last names (n=9,949)
- Supplemented with non-name text (e.g., "pilot study", "department funded") (n= 166)

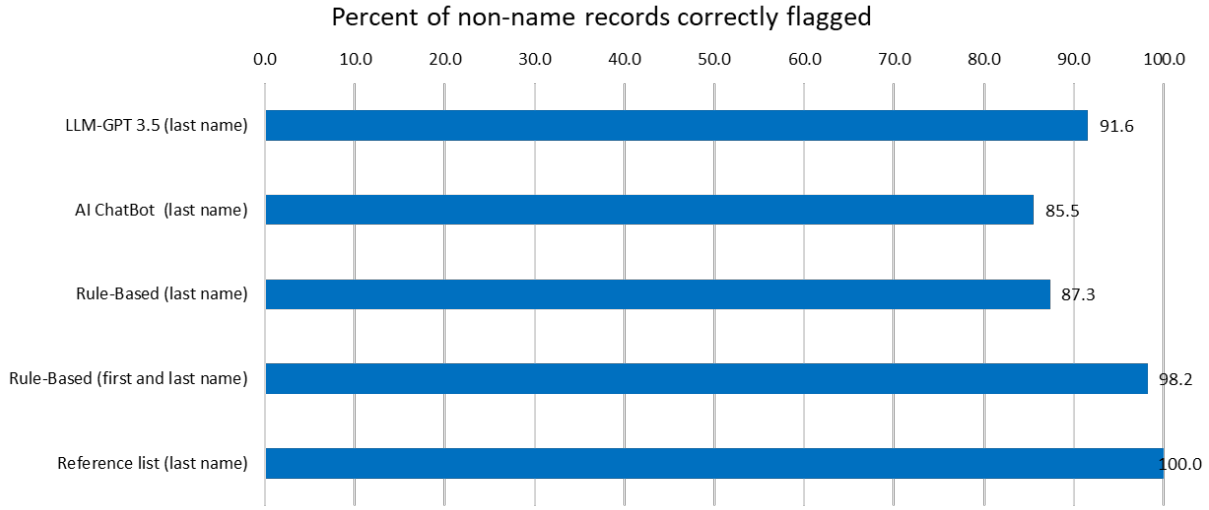
Approaches evaluated:

- LLMs
  - GPT-3.5 with few-shot prompting
  - AI Chatbot (cdc.gov)
- Rules-based approached based on number of words and characters
- Comparison with reference list of "valid" last names

Using AI to Enhance Data Preprocessing

Results





Using ML to Improve Linkage Efficiency

2. Record linkage: AI algorithms, such as probabilistic matching models or deep learning models, can be employed for record linkage tasks. These models can learn patterns and similarities between records to make accurate matches, even when dealing with large-scale datasets.

3. Record linkage: Machine learning algorithms can be applied to perform record linkage tasks. By training on labeled data with known matches and non-matches, these algorithms learn to identify similarities and make accurate matches between records across different datasets.

- Record linkage enables survey data to be integrated with other data sources, expanding the analytic potential of both sources
- Depending on the number of records being linked, the processing time can be prohibitive
- ML algorithms can be incorporated into the record linkage process to improve efficiency with relatively high accuracy

Using ML to Improve Linkage Efficiency

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 machine learning algorithm, the Sequential Coverage Algorithm (SCA), can be used to identify efficient blocking schemes for linking two large datasets. Compare SCA to traditional blocking methods.

Methods:

- 2016 National Hospital Care Survey (n = 5.6 million)

  Center for Medicare & Medicaid Services (CMS) Medicare Enrollment Database (EDB) (n = 84.6 million)
- Data fields 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)
- Ad-hoc blocking methods require judgment of a SME and result in extended processing time
- SCA is a supervised ML algorithm designed to learn a set of efficient blocking keys
  - Modified version of SCA used (Michelson and Knoblock, 2006)
  - Three sequential steps: 1) Learn Block 2) Remove Covered Pairs 3) Optimize

Using ML to Improve Linkage Efficiency

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 Linked Data to Train ML Models

1. Feature engineering: Machine learning techniques can be used to extract and create relevant features from the linked data. These features can capture important characteristics or patterns that aid in accurate record linkage or entity resolution.

2. Model training: Machine learning models, such as classification algorithms or deep learning models, can be trained using labeled data to learn patterns and similarities between records. These models can then be used to predict matches or resolve entities in new, unlabeled data.

- NCHS Data Linkage Program produces high-quality linked data files that can be used for a wide-range of health-related research topics and data science applications
- ML can be used to identify relevant predictors (features) for outcome of interest
- ML prediction models can predict outcomes with high accuracy, but require quality and accurate data for training and validation

Using Linked Data to Train ML Models

Using the National Health Interview Survey Linked Mortality Files to Train and Assess the Performance of Machine Learning Prediction Models to Predict All-Cause Mortality and Interpret Model Predictions using Explainable AI

Orlando Davy, MPH; Frances McCarty, PhD; Yulei He, PhD; Cordell Golden, MPS
Centers for Disease Control and Prevention, National Center for Health Statistics

Objectives:

Evaluate selected ML prediction models using linked data as the training data and validation source to assess model performance for predicting all-cause mortality. Use Explainable AI to interpret model results.

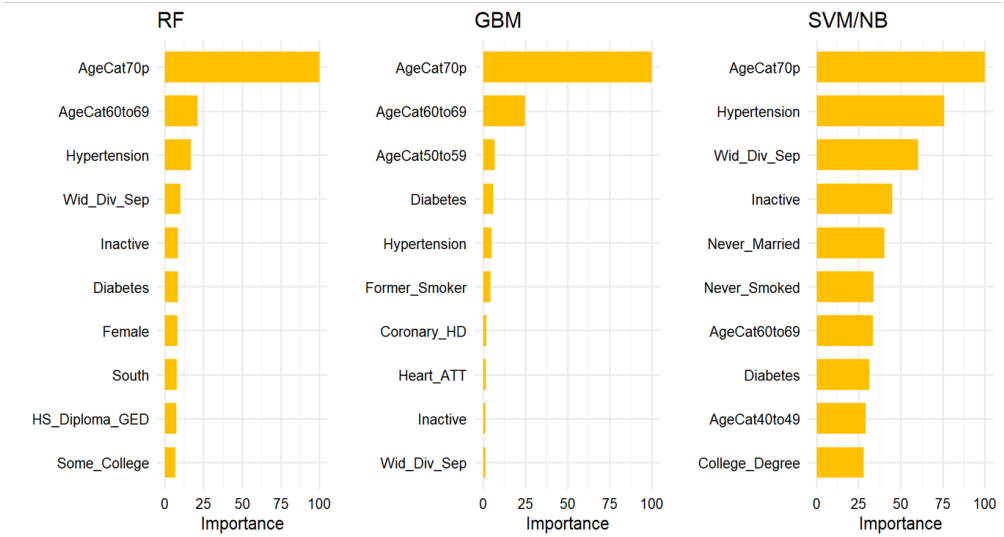
Methods:

- Public-use 2000 - 2001 National Health Interview Survey (NHIS) Linked Mortality Files (LMF) with mortality information through 2019
- Linkage eligible sample adult respondents with complete information for selected predictor variables (n = 46,949)
  - Training set: 2000 NHIS LMF (n = 23,210)
  - Validation set: 2001 NHIS LMF (n = 23,739)
- 19 selected predictors (socioeconomic status, health behaviors and health conditions)
- Selected ML prediction models: Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), Naïve Bayes (NB)

Using Linked Data to Train ML Models

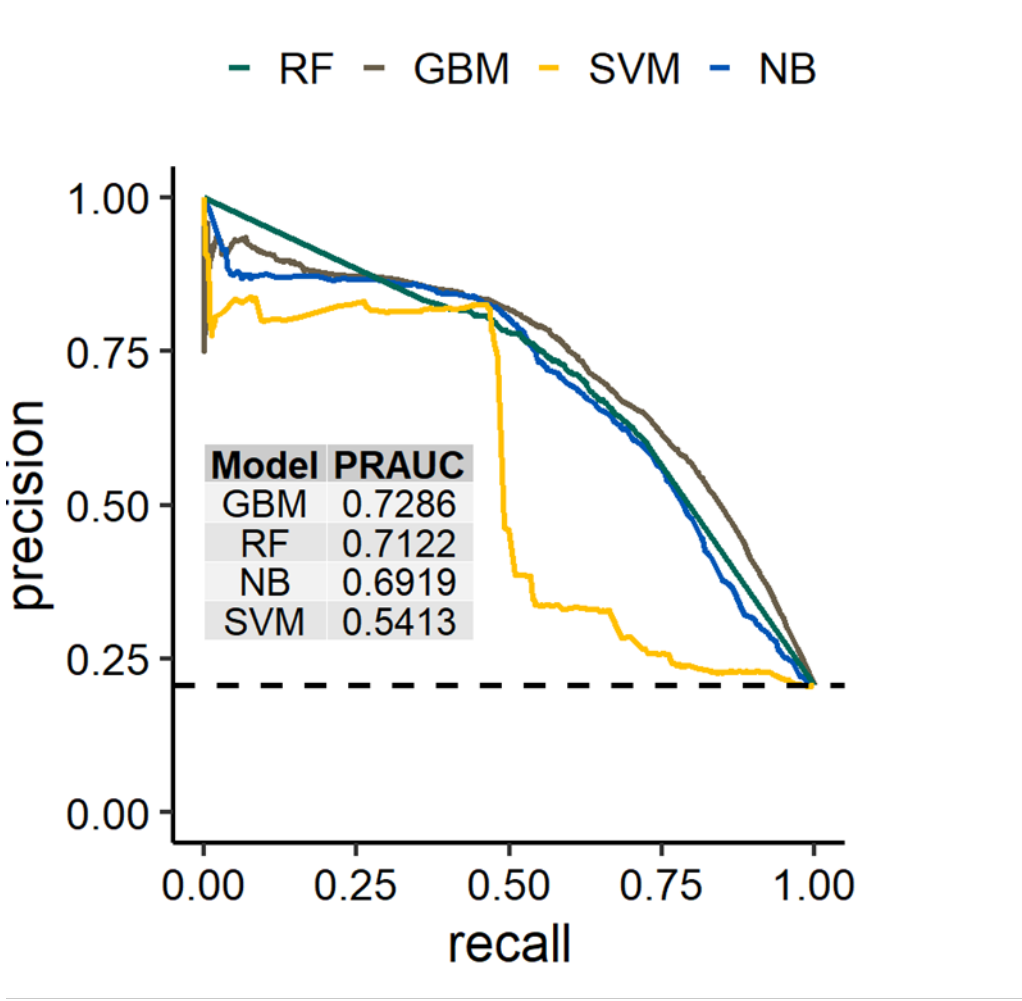
Results

Variable Importance



Performance Measures

	RF	GBM	SVM	NB
Misclassification Error	0.1305	0.1239	0.1320	0.1560
Precision	0.7304	0.7752	0.8262	0.6044
Recall	0.5826	0.5957	0.4562	0.7051
Balanced Accuracy	0.7633	0.7723	0.7156	0.7926
F1-Score	0.6481	0.6648	0.5878	0.6509
Process Time(min)	3.20	0.58	1.00	0.02



Summary and Considerations

- The trustworthy application of AI and ML supports several federal initiatives
- AI and ML can be incorporated to the Data Linkage Life Cycle to improve quality, efficiency, and accuracy
- AI and ML models/algorithms require high-quality data for training and validation
  - But biases in training data can impact models/algorithms
- Data linkages often requires use of sensitive data
  - But AI/ML may retain information to improve performance
- Transparency and adherence to standards/best practices are essential to ensuring trustworthy application of AI/ML

Thank you!

Cordell Golden
CGolden@cdc.gov

Visit our website: www.cdc.gov/nchs/data-linkage

Contact the Data Linkage Program: datalinkage@cdc.gov

Subscribe to the NCHS Data Linkage Program LISTSERV to receive updates! Email a message to list@cdc.gov. Leave the subject line blank. In the body of the message, type:

- SUBSCRIBE NCHS-DATA-LINKAGE-PROGRAM last name, first name

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 report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.