Using APCD Unique IDs for Data Linkage

Kenley Money
2018 NAHDO Annual Conference
October 11, 2018

ARKANSAS APCD
All-Payer Claims Database

ADMINISTERED BY ACHI
Arkansas Center for Health Improvement

Introduction

Kenley Money

Director of Information Systems Architecture

Kenley.Money@achi.net

Kanna Lewis

Microsimulation Architect

KLewis@achi.net

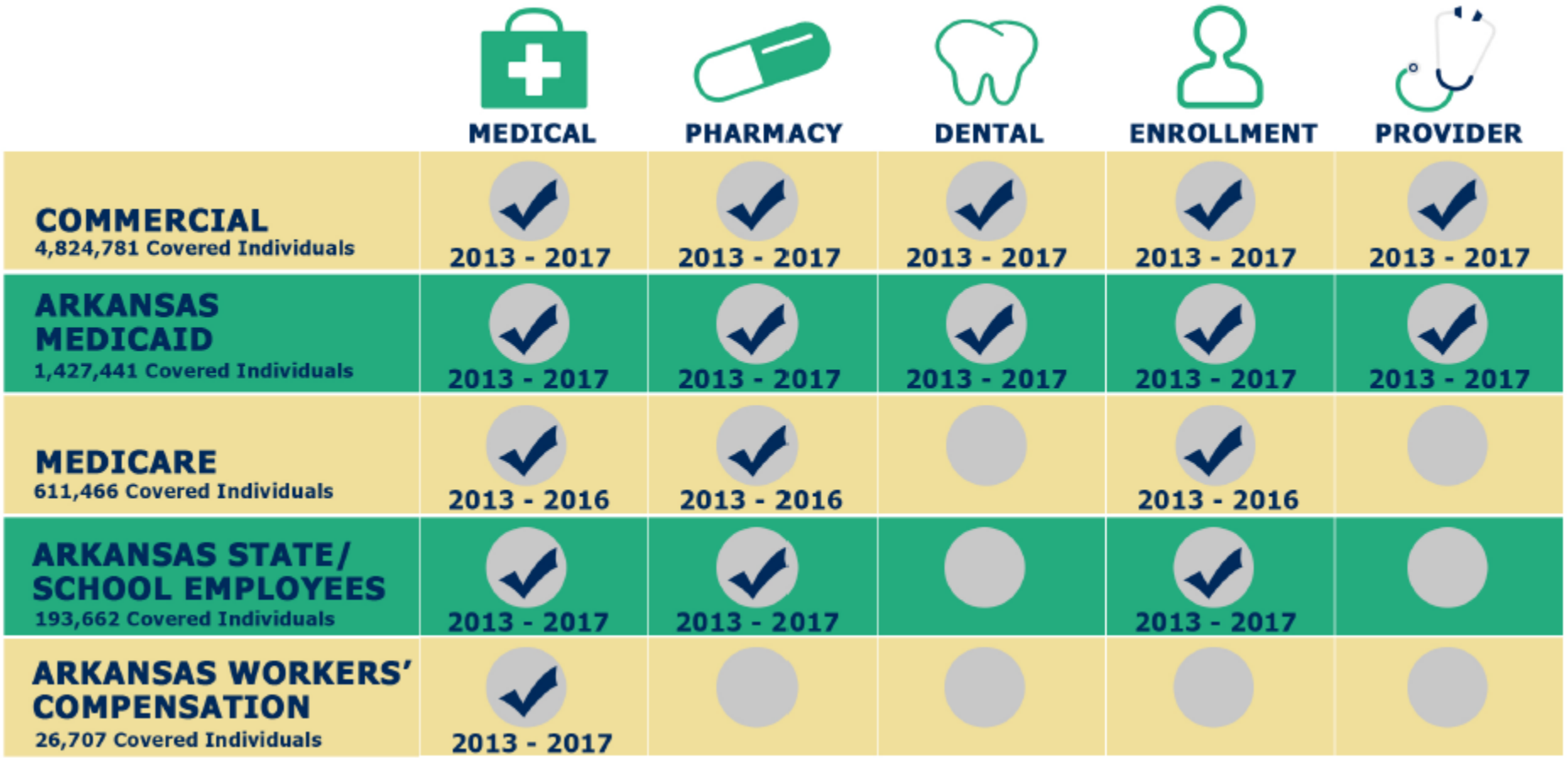
About ACHI

- The Arkansas Center for Health Improvement (ACHI) is a nonpartisan, independent health policy center dedicated to improving the health of Arkansans
- Established in 1998, creating a much needed intersection between research and policy
- ACHI has proven experience in the management and integration of health data to support research and health policy

Arkansas All-Payer Claims Database

- Established with Act 1233 of 2015, naming ACHI as the APCD Administrator
- Collects member/enrollment, medical claims, pharmacy claims, dental claims, and provider data
- Requires data submission from all carriers with 2,000 or more members/enrollees in Arkansas
- Does not allow the collection of personal identifiers (i.e., name, address, city, and SSN)

Arkansas All-Payer Claims Database



Challenges Faced Without Personal Identifiers

- Unable to track individuals longitudinally within and across carriers
- Unable to execute comprehensive analyses, including but not limited to:
  - Understand cost and impact of healthcare coverage on the Arkansas population
  - Track disease burden and other social determinates of health across the Arkansas population
- Significantly reduces the value of Arkansas APCD data to data requestors

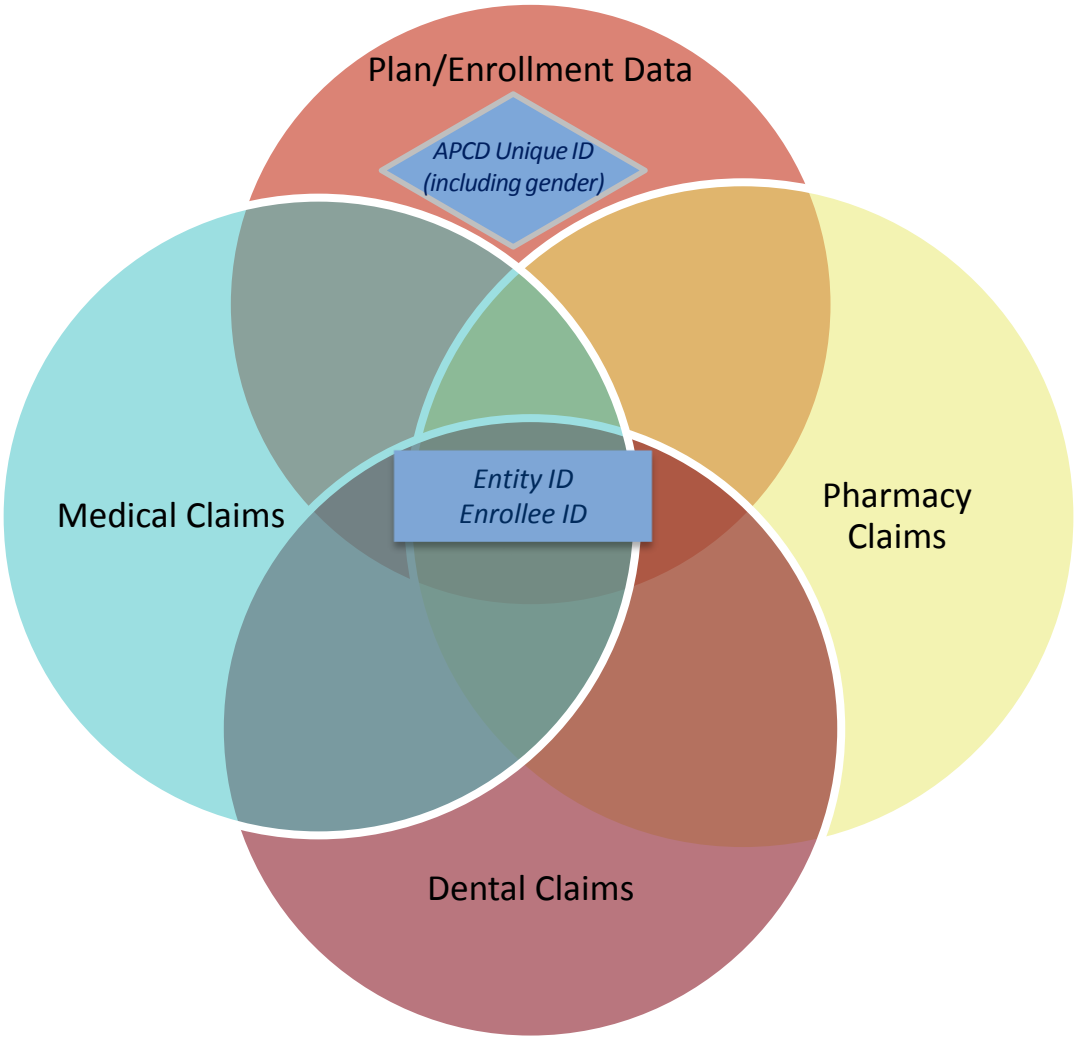
APCD Unique ID

- The APCD Unique ID can be used as a proxy to identify members/enrollees across carriers
- The APCD Unique ID is a hashed version of the last name and date of birth for each member/enrollee
- Each member/enrollment record contains the enrollee's APCD Unique ID
- All carriers are required to create the APCD Unique ID using the same hashing methodology to ensure consistency
- Used with gender, the APCD Unique ID can identify unique enrollees across carriers with high accuracy

APCD Unique ID

Within Carriers:
The Entity ID (representing the carrier) + Enrollee ID (sometimes called the Member ID) are used to associate member/enrollee records with claims records.

Across Carriers:
The APCD Unique ID plus gender code are used to find members/enrollees in different carriers, and sometimes in different plans within a carrier's submission.



APCD Unique ID Accuracy

- Combining last names — especially common last names — with date of birth can link data from different individuals to a single individual when these are the same; this is called a **Collision**
- Adding gender to the APCD Unique ID helps identify them as different individuals; Collisions can still happen

How often do individuals have the same last name, date of birth, and gender?

High-Level Examples

Entity – In this framework, an entity is an individual

Reference – (APCD Unique ID, gender) pair

First Name	Last Name	Date of Birth	Gender
John	Smith	10/31/1985	Male
Mike	Smith	10/31/1985	Male

APCD Unique ID	Gender
pm5XL/6OKZ	Male
pm5XL/6OKZ	Male

Reference matching
(Collision)

Linkage

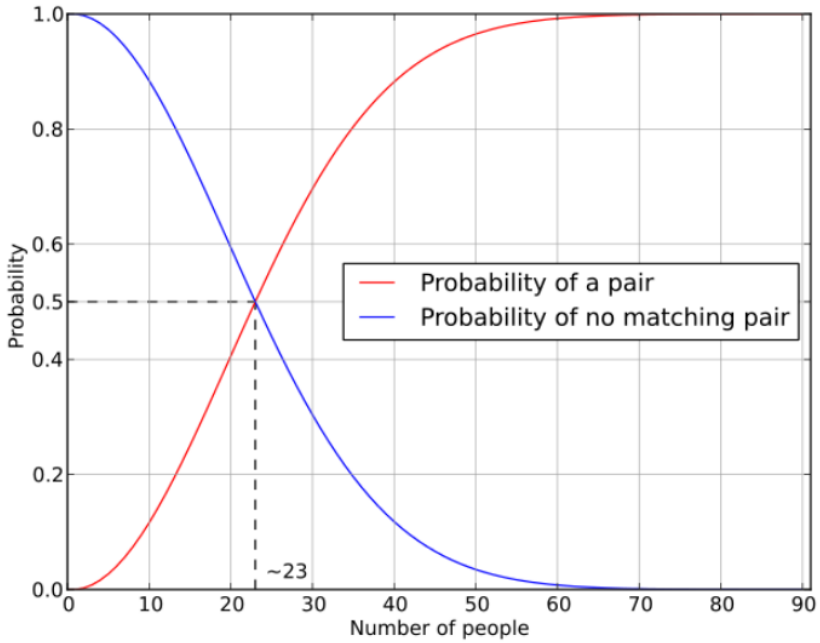
Arkansas APCD has very limited demographics information. Here, the focus will only be on deterministic linkage using (APCD Unique ID, gender).

Goals

- Data quality validation by quantifying the expected reference matching rate
- Linkage accuracy improvement by removing APCD Unique IDs with a high probability of a false positive

The Birthday Problem

If there are 23 people in one room, there is 50% probability that at least two people in the room share the same birthday (not day of the week).



$p(n)$ is the probability of at least two of the n people sharing a birthday. According to the pigeonhole principle, $p(n) = 1$ when $n > 365$. When $n \leq 365$:

$$p(n) = 1 - \frac{n!\binom{365}{n}}{365^n}$$

Last Name Distribution Over Time

Diversification of last name distribution

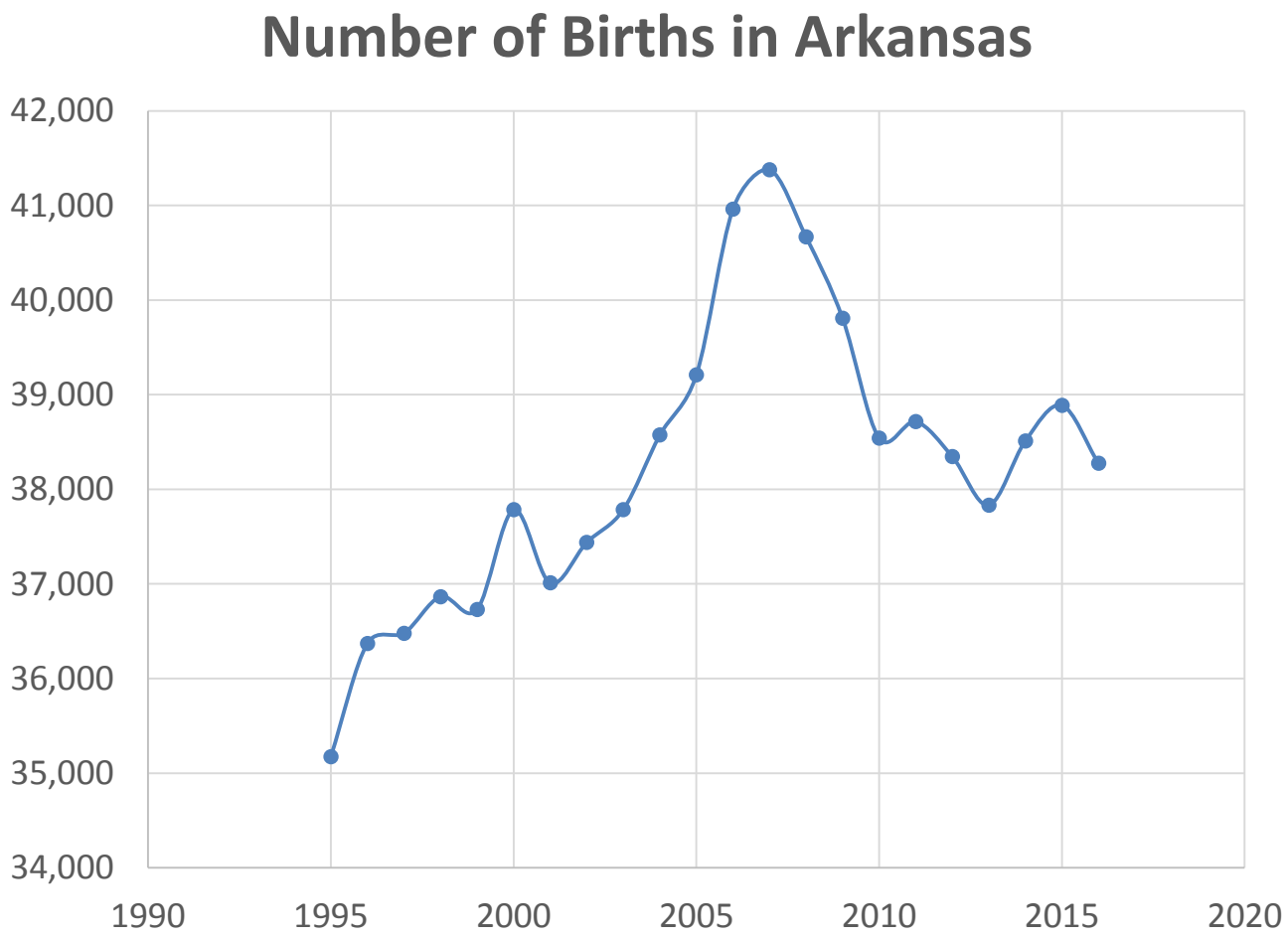
Top 5 Last Names in Arkansas

1990		2015	
Last Name	Rate (%)	Last Name	Rate (%)
SMITH	1.6	SMITH	1.2
WILLIAMS	1.2	WILLIAMS	1.0
JOHNSON	1.1	JOHNSON	0.9
JONES	1.1	DAVIS	0.8
BROWN	0.9	BROWN	0.7

GARCIA

1990 0.05% (Rank 267) → 2015 0.28% (Rank 21)

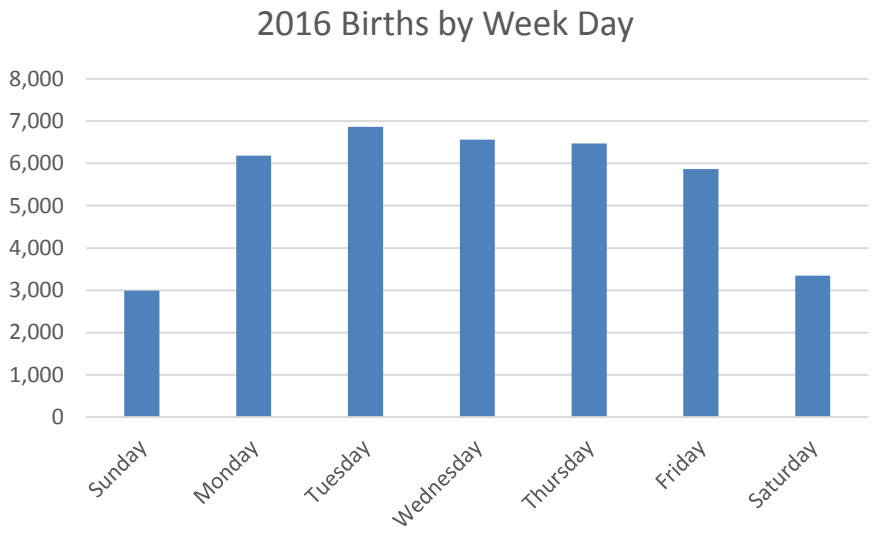
Birth Year Distribution

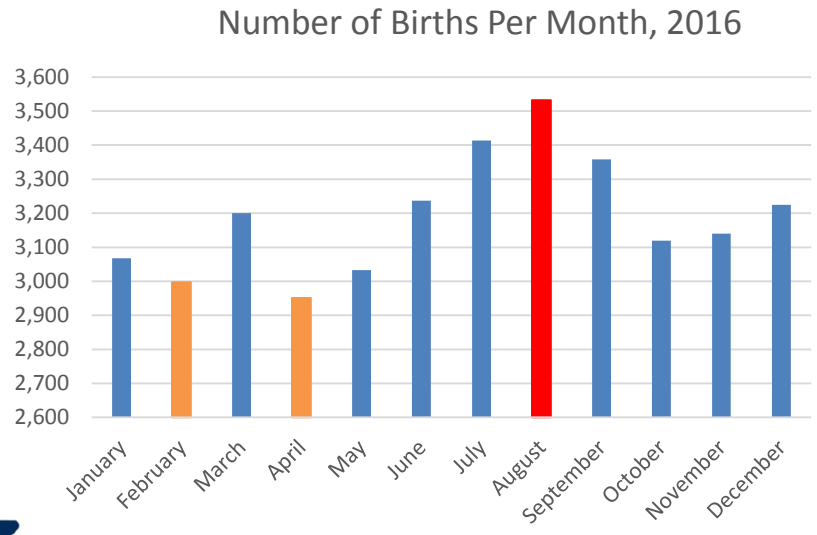


Birthday Distribution in Arkansas

"Weekday effect" is especially prominent in recent years

August — high number of births
February & April — low numbers of births





Model

Calibrate birthday distribution conditional on (last name, birth year) and use a combinatorial argument to obtain the expected value and variance of reference matching rate.

Data

(1) Arkansas birth certificate data: birth year 1989-present (Arkansas Department of Health)

(2) Arkansas voter roster (public record)

Modeling steps

(1) Estimate p(birthday | last name, birth year) while utilizing a smoothing variable maximizing the likelihood of observing the empirical data under the model.

- Tested a use of empirical (counting) model, as well as the smoothed model, and found the smoothed model to produce a more accurate result in randomly generated files.

(2) Use (1) to compute the expected number of reference matches and variance.

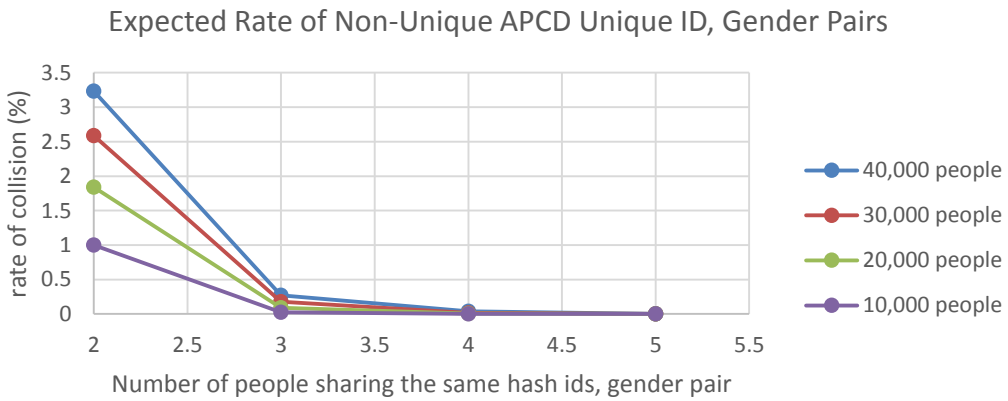
- The formula is derived by induction on the number of references sharing the same birthday. The majority of reference matches are due to exactly two references sharing the birthday. And the number of pairs sharing a birthday grow as

$$\propto \binom{N}{2} \sum_{1 \le i \le 365} p_i^2$$

where $p_i = 1/365$ for uniform birthday distribution, for example.

Model Results

The expected rate of (APCD Unique ID, gender) collisions for the population in Arkansas is approximately **3.5%**. The more references there are in the file, the higher rate of reference matching.



A **probability score** will be recorded alongside APCD Unique ID and gender to improve on the accuracy of the record linkage. For example, an APCD Unique ID corresponding to "SMITH born on Wednesday" has a high record matching score and the corresponding record may not be used to improve on specificity.

One caveat: In our model, there was no special consideration given to twins. The number of expected twins in a dataset is not dependent on the number of references. The rate of twins of the same gender in Arkansas for all datasets is estimated to be **1.6%.**

Proof of Concept 1: School BMI-Diabetes Linkage

School BMI-APCD Record Linkage

Dataset Validation

40,000 distinct (APCD Unique ID, gender) pairs in the BMI dataset per birth year on average. Around 1,400 (3.6%) per birth year have at least one reference match within the BMI dataset, passing the data validation test.

Creation of an Analyzable Dataset

Those who found a reference match within the BMI dataset were removed prior to linkage with APCD to improve upon specificity. Since it is very unlikely that more reference matches would be found in the APCD than those found in the BMI dataset alone, the linkage can be justified with **99%** accuracy.

Proof of Concept 2: Birth-Death Certificate Record Linkage

- Infant mortality study
  - To support efforts to reduce infant mortality, ACHI has conducted analyses to identify infants who died within the first 12 months of life and generate a profile of their healthcare service utilization.
  - Death certificate of population deceased before age 1 was linked with birth certificate to determine the cause of death.
  - APCD Unique ID validation model renders a solid quantitative guideline for when it is appropriate to link records by APCD Unique ID alone. In this study, **134 records out of 1,014 had reference matches due to a high rate of death among multiples** (twins and triplets). This exceeds a model derived collision threshold. Thus, other measures such as PII were utilized for a better linkage accuracy.

Conclusion

The reference matching rate in a randomly created analytic dataset is expected to be **3.5%** and lower for a smaller set size. However, the context of study set needs to be closely monitored.

APCD Unique ID combined with gender (and other data as needed) can represent unique individuals for many types of analyses in lieu of full personally identifiable information.