# Linking Human Resources and Child Welfare Data

## What is Data Linking?

Data linking involves pairing observations from two or more data files and identifying the pairs that belong to the same entity. For example, if you wanted to know how caseload is associated with the length of time someone stayed with the agency, this would typically require linking Human Resources (HR) data for an individual with data for cases assigned to the same individual pulled from a child welfare case-level data system. Another type of linking problem might involve tying training data from a learning management system to case-level data to determine whether training resulted in better practice, such as more accurate risk assessments.

## Linking in Child Welfare Agencies

The main data sources needed to answer questions related to employee performance, tenure, and turnover are HR data, training system data, and child welfare case data.

Within each of these large buckets, there also may be separate databases that need to be linked to answer the question you are interested in. For example, applicant information may be stored in a database that is separate from other HR data on those hired. In state-supervised, county-administered systems, there might be a state-level system in place for case-level information, but HR data are held by individual counties that have different systems in place.

In addition to the existence of multiple databases housing data of interest, challenges may be encountered when identifying variables within these databases to use as linking keys. A common ID code across systems provides a key that may be used to clearly link records from different databases, through what is known as a deterministic match.

However, in many jurisdictions, different ID codes are used by HR and child welfare data systems. Learning management systems or other training databases may use yet another means of identification. Making an accurate deterministic linkage also requires that ID codes be unique to an individual. IDs are not unique in systems that use a position ID rather than a person ID. More than one individual may be associated with a given position code, making it unclear when linkages are between the same or different entities. Another issue arises when a new ID is assigned when a person is rehired after leaving the agency. In this case the full picture of the individual's tenure would require using other methods of linking. A further complication arises when an ID code is reused after someone leaves the agency. Last, but not least, a match may fail, or different individuals may be linked erroneously, even when a common unique identifier is available, due to inaccurate or missing data in the ID field.

## Data Preparation

Before beginning any data linking, a first step is data cleaning and standardization. Work done up front to correct issues such as missing data and entry errors and to reconcile differences in how the same data appear in different files (e.g., names in lower case versus upper case or full names versus initials) will maximize successful

matches. The table below (Dusetzina et al., 2014) shows some common issues that could affect a successful linkage.

| Field | Type of Issue | Examples |
|---|---|---|
| Names | Case<br>Nicknames<br>Synonyms<br>Prefixes<br>Suffixes<br>Punctuation<br>Spaces<br>Digits<br>Initials<br>Transposition | John Smith \| JOHN SMITH<br>Charles \| Chuck<br>William \| Bill<br>Dr. John Smith<br>John Smith, III<br>O'Malley \| Smith-Taylor \| Smith, Jr. John Smith Jr<br>J2ohn Smith<br>AM \| A.M. \| Anne Marie<br>John Smith \| Smith John |
| Address | Abbreviations | RD \| Road \| DR \| Drive |
| Dates | Format<br>Invalid values | 01012013 \| 01-01-2013 \| 01JAN2013<br>Month = 13 \| Day = 32 \| Birth year = 2025 \| Date = 29FEB2013 |
| Social Security Number | Format | 999999999 \| 999-99-9999 \| 999 99 9999 |
| Geographic Locations | Abbreviations<br>FIPS codes<br>SSA codes<br>ZIP Codes<br>Concatenation of state and county codes | NC \| North Carolina<br>North Carolina = 37<br>North Carolina = 34<br>99999 \| 99999-9999<br>Mecklenburg County, NC \| 37119 |
| Sex | Format | Male / Female \| M / F \| 1 / 2 |

FIPS = Federal Information Processing Standards; SSA = Social Security Administration

Dusetzina and colleagues suggest several strategies to handle these issues, including:

1. Ensure all fields used as keys for linking data are in the same format (e.g., convert names to uppercase, remove periods and spaces).
2. Parse identifiers to maximize the available information. For example, separate date of birth or other date fields into separate fields for month, day, and year. Separate addresses into separate fields for street, city, state, and zip, and separate name into first name, middle initial, and last name. Doing this when using a probabilistic linking software allows credit for a partial match when you do not have letter-for-letter agreement. Partial matches, when combined with other indicators, may be enough to make a clear deterministic match (see below for more information on probabilistic and deterministic matches). This type of strategy is also recommended when looking at data over time (e.g., when a name changes due to marriage or divorce).
3. Minor misspellings or typos in data strings (e.g., names) may be helped by conversion to phonetic codes using a method such as Soundex, which is included in many database packages. The National Archives and Records Administration maintains the current rule set for the official implementation of Soundex used by the U.S. government.

## Types of Record Linkage

**Deterministic Matching**

Deterministic matches link individuals across databases by making an exact match on a common identifier or set of identifiers. The common identifier could be a unique ID code that is the same across all systems or a set of variables that describe someone sufficiently to constitute an exact match (e.g., name, DOB, gender, office).[1] Alternatively, cross-walking can be used to connect data using other data that they have in common. For example, if social security number links to personnel number in HR data and also links to a child welfare worker ID in case data, the HR and child welfare ID codes can be linked deterministically by cross-walking using the social security number. For a successful deterministic match, everything needs to match exactly. Even with a common ID code, matching still may not be perfect due to missing data, duplicate records, or entry errors that prevent a match.

**Probabilistic Matching**

Probabilistic matching compares records from each system on one or more common fields (e.g., name) and uses a set of formal decision rules to assign a probability that the records are a match. The Centers for Disease Control recommends probabilistic matching when there are duplicate records, missing data, or coding errors. Probabilistic matching would also be used when no linking variables (such as a common, unique ID code that could be used for a deterministic match) are available. The success of probabilistic approaches depends on the uniqueness of the combinations of variables used to make the match.

*How probabilistic matching works*

Software calculates a score for a pair of records based on how likely they are to be the same person. The matched pairs are then sorted in score order, and a threshold is set for accepting the score as a match. A second threshold is set for clearly rejecting the score as a match.[2] In between is a gray area where potential matches are manually reviewed.

*Steps in the probabilistic matching process*[3]

1. Blocking—Blocking makes matching more manageable by defining a few key matching variables and sorting files on those variables so that the matching is done within blocks, where matches are more likely and very unlikely matches are not considered. Typical blocking variables are name, date of birth, gender, race/ethnicity, and address.
2. Matching—Potential matches are identified based on field-specific or value-specific matching within the blocking variables, and a probability weight is assigned to each pair of records.
    a. Field-specific linking matches records on a variable value. Agreement on a more specific field like date of birth generates a higher score than matching on a less specific field like gender. Agreement on more of the blocking variables generates a higher probability score.
    b. Value-specific linking matches on specific values of the blocking variables. Agreement on an uncommon value is stronger than agreement on a common value. For example, a name match on John Smith would be less definitive than a name match on Quentin Higginbotham.
3. Scoring—A total probability weight is assigned to a pair of records based on a combination of the probabilities of a match generated by comparisons on each blocking variable. Linking software also typically assigns a score for sensitivity (the probability that a match is a true

positive) and specificity (the probability that pairs of records identified as a non-match are true negatives). Programs may use other metrics as well.

4. Decision Making—The range of probabilities is examined to set two cutoff scores: 1) an upper bound that defines the point at which a score would be considered a match and 2) a lower bound that defines the point below which a pair of records would be considered a non-match. In the gray area between these two cut scores, matches would undergo a manual review and additional variables may be considered.

*Probabilistic matching software*

There are several options to choose from, both paid and open source. Those listed below are recommended by the U.S. Agency for Healthcare Research and Quality. For smaller agencies or linking tasks, Microsoft Excel also offers a fuzzy matching plug-in (limited to databases of no more than 79,000 rows). What you choose will depend on your agency size, budget, and the complexity of your project needs.

Conducting the matching process may require the assistance of your IT department or other people not currently part of your team. They may have experience with one or more of these options or be aware of others.

Free/open source linking software

1. Link Plus (or Link Plus Beta), developed by the Centers for Disease Control. The package provides a graphical user interface that is straightforward and easy to use, requiring only beginner-level knowledge of the linkage process.
2. Link King, developed by Washington State's Division of Alcohol and Substance Abuse. Like Link Plus, this software provides a graphical user interface that is described as straightforward and easy to use, requiring only beginner-level knowledge of the linkage process. The Link King is capable of handling larger datasets. The software is free, but it requires having a SAS license.
3. Choice Maker 2, developed by ChoiceMaker Technologies. The program is used by health care researchers and handles large datasets. It uses machine language technology that requires data to train the algorithm. The program can be used with Java and runs against Oracle, MS SQL Server, and XML databases. It provides web services and batch interfaces (Goldberg & Borthwick, 2004).
4. Febrl, developed by the ANU Data Mining Group is also used by health care researchers. The website says it is intended for advanced users and works on Windows 10. It also is described as efficiently handling very large data sets.

Commercial software

1. LinkageWiz, developed by LinkageWiz Software. This is a standalone product that doesn't require additional software such as SAS or Microsoft Access. Offers a free trial version.
2. G-Link, developed by Statistics Canada. Click here for more information.
3. LinkSolv, developed by StrategicMatching. This is an Access application and requires purchase of Microsoft Access. Data do not have to be in Access tables.

## Resources/Additional Reading


Ansolabehere, S., & Hersh, E. D. (2017). ADGN: An algorithm for record linkage using address, date of birth, gender, and name. *Statistics and Public Policy, 4*, 1–10.

Christen, P., & Churches, T. (2005). *Febrl - freely extensible biomedical record linkage*. http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/manual.html

Dusetzina, S. B., Tyree, S., Meyer, A. M., Meyer, A., Green, L., & Carpenter, W. (2014). *Linking data for health services research: A framework and instructional guide* (AHRQ Publication No. 14-EHC033-EF). Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK253313/

Goldberg, A., & Borthwick, A. (2004). *The ChoiceMaker2 record matching system.* ChoiceMaker Technologies. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.2691&rep=rep1&type=pdf

Hoopes, M. (n.d.). *Record linkage concepts* [PowerPoint slides]. https://www.hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or19.pdf

Putnam-Hornstein E. (2013, August). *Linking records to advance child protection: Recent examples from California*. Invited Presentation, Los Angeles County Department of Public Health: Los Angeles, CA.


---

[1] Ansolabehere & Hersh (2017) found that matching by all or any combination of address, name, gender and date of birth produced a rate of matches comparable to using social security number in a large-scale study linking Texas voter registration data, military ID, passport number, and state ID data (e.g., driver's license, concealed carry permits)

[2] https://www.hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or19.pdf

[3] Hoopes, M., Record Linkage Concepts, (PowerPoint Presentation) retrieved from https://www.hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or19.pdf, 6/13/2020.




Funded through the Department of Health and Human Services, Administration for Children and Families, Children's Bureau, Grant #HHS-2016- ACF-ACYF-CT-1178. The content of this publication does not necessarily reflect the view or policies of the funder, nor does mention of trade names, commercial products or organizations imply endorsement by the US Department of Health and Human Services.