Chapter Three
Methods

This chapter provides a brief description of data matching in general and a detailed description of the specific record linkage procedures used for this study.

Data Matching

The term "data matching" may be used to describe three different methods for linking data from multiple sources:

- **Match-merging.** Match of two data files by relying on an exact match of a single common identifier. This method is generally used only when data originate from the same data system or when identifiers (such as SSN) are highly reliable in both data files being matched, or verified in at least one of the data files. Match-merge is the simplest version of deterministic matching.

- **Deterministic matching with multiple identifiers.** Match of two data files using multiple identifiers (e.g., name and date of birth) and a match-merge technique relying on an exact match of identifiers. Multiple identifiers are used when a highly reliable single common identifier is not available.
- **Probabilistic matching.** Probabilistic record linkage is made when the calculated statistical probability of a match exceeds a certain threshold. Matches between two data files are based on comparison of multiple data fields in the two files. Identifiers need not match exactly; identifiers that do not match exactly are assigned a "distance" measure to express the degree of difference between files. Each identifier is assigned a weight (which is data driven) and the total weighted comparison for all identifiers being compared yields a score classifying records as linked, not linked, or uncertainly linked.

FSP and WIC do not share a common information system and do not share a common person ID. FSP and WIC records cannot be reliably linked via a merge on SSN because the SSN may not be equally reliable in the two files: FSP validates SSNs but WIC does not (according to the Phase 1 survey).[45] Because SSNs are not validated by both programs, there is potential for false positive and false negative results from a match on SSN. For this reason, probabilistic matching was the primary approach used for this study, with deterministic matching conducted for sensitivity analyses.

The typical steps in the data matching process are (Clarke, 1995):

- Record selection – selection of a subset of records meeting the population definition for the data that are being matched (described in chapter 2).
- Data standardization – data fields are standardized to impose consistent coding schemes or to parse free-form name and address fields into component parts that are more easily compared.
- Matching – the record-linkage process.
- Inferring – drawing conclusions about the accuracy of the match.

[45] In addition, it was found that WIC certification records are often missing SSNs for infants and young children.

There is a large literature on probabilistic record linkage that is summarized here only briefly.[46] From a theoretical standpoint, probabilistic matching involves the following sequence of events:

- Pairing every record in the base file (file A) with every record in the match file (file B).
- Comparing and scoring the quality of the match for each individual matching variable (e.g., name, SSN, date of birth).
- Applying weights to each matching variable to obtain an overall score.
- For each record in the base file (file A), identifying the matched pair with the highest overall score.
- Determining cutoffs for the overall score to classify matched records into: a) certain matches, b) uncertain matches, and c) certain non-matches.

From a practical standpoint, every record in file A cannot be paired to every record in file B because the Cartesian product is unmanageable.[47] For example, the Cartesian product of Iowa FSP and WIC records for December 2002 would yield a file of over 4 billion records. Typically data are blocked and paired within block (for example, within county); matches and non-matches are identified; and non-matches are re-blocked on different criteria.

Record Linkage Procedures

This study used "Record Linkage Software" from the Census Bureau (U.S. Bureau of Census, undated). The software includes name and address standardization programs as well as the record linkage program. The software is designed to perform one-to-one matching whereby one record in file A is matched to one record in file B. The programs are written in C and Fortran, and the compiled program executables run on personal computers.

The Census record linkage program contains the matching algorithms that are applied to data, using parameters specified by the user. The program leaves considerable decision-making up to the user because matching routines must be data driven, depending on the characteristics and quality of the data being matched. As a result, new users of the program can expect to devote considerable time adapting it to the application at hand.

The main steps in the process are:

1. Data reduction and unduplication
2. Data standardization
3. Determination of blocking variables
4. Determination of matching variables
5. Specification of match parameters
6. Review of match results and specification of score cutoff for certain matches

Each of these steps is discussed below. After processing data in steps 1 and 2, matching was done for: (a) active caseloads in December 2002, and (b) all participants active during the three-year period (Florida and Iowa only). The active caseloads in December 2002 contain one record per person and

[46] The seminal works are Newcombe (1959) and Fellegi and Sunter (1969). See also Winkler (1994 and 1999).

[47] The Cartesian product of two databases is obtained by joining all the records from the first database with all the records from the second database in every possible combination.

tell us the rate of multiple program participation at a point in time (contemporaneous participation). Utilizing all records for the three-year period involves multiple records per person and tells us, for example, the percent of WIC children ever participating in FSP within the three-year period (exposure).

Data reduction and unduplication

Data reduction is the process of reducing the size of data files by eliminating redundant data. The purpose is to enable efficient processing of the large files received from the States. This step was crucial for FSP data files, which were received with one record per participant per month. As discussed in chapter 2, many data fields showed no changes in individuals' information over time. Data were reduced by eliminating records where all data fields were identical except for month of participation; the one "unique" record that was retained was assigned indicators for months of participation from "dropped" records. Data reduction was done both before and after data standardization. The resulting file had multiple records for an individual only if personal identifiers or contact information changed.

Unduplication is the process of ensuring that each individual is associated with only one unique participant ID in each data file. FSP and WIC data from the three States were examined to assess the need for unduplicating prior to matching (for example, we sorted by SSN and participant ID to determine if a single SSN was associated with multiple IDs, and therefore lacked a unique ID). The incidence of duplicates was well below 1 percent and did not warrant the effort to systematically unduplicate the data.[48]

Data standardization

Data were standardized prior to running the record linkage program. Standardization ensures that truly identical data are evaluated as a match across data files. Three types of standardization were applied to the data:

> ***Format standardization***. Data elements were assigned consistent names and formats in all files, imposing consistent coding schemes on categorical variables (race, gender, county) and consistent lengths on character variables.
>
> ***Name standardization.*** First and last name were parsed into separate data fields when a single name field was provided.[49] In all files, first names were standardized, to eliminate nicknames.
>
> ***Address standardization***. A standardization program from the US Census Bureau was used to parse and standardize street address information. The standardization program parses a single street address field into separate fields for: house number, street prefix (e.g., east, west), street

[48] A different "duplicate record problem" was found in the Kentucky WIC data. Kentucky WIC had unique IDs for participants, but there were duplicate records for the combination of participant ID and certification date. These appeared to be due to changes in address associated with transfers to a different local agency. Approximately 34,000 pairs of participant ID and certification date were observed and duplicates were deleted.

[49] The Kentucky FSP file included a single "name" field containing data in the form "last, first".

name, street type (e.g., street, road, avenue), street suffix (e.g., east, west), rural route, and PO box.

Name standardization was implemented using a list of nicknames and corresponding full names that was supplied with the Census standardization programs but not invoked by the PC version of the programs. These nicknames were standardized to full names in SAS. [50] The Census nickname list contained 493 nickname-name pairs, but did not contain a large number of ethnic names, and was found to be more relevant to women than infants and children. For example, in the WIC programs, 7 to 11 percent of women's names were identified as nicknames across States, but only 4 to 6 percent of infants/children's names were identified as nicknames.

Address data pose the greatest challenge for standardization because of the variety of abbreviations and symbols that may be used within an address field. For example, street types may be spelled out (Street, Road, Avenue) or abbreviated (St, Rd, Ave); apartment numbers may be preceded by "#" or "No."; and directional prefixes or suffixes (North, South, East, West) may also appear as street names (e.g., 51 South Street). The Census address standardization subroutine parses address data into 10 component parts:[51]

	Element Name	Examples[52]
1.	House Number	123
2.	Street Name Prefix	N, S, E, W
3.	Street Name	Oak, Main, Martin Luther King
4.	Street Name Suffix	Avenue, Blvd, Place, Court
5.	Within Structure Designator	Bldg, Apt, #
6.	Within Structure Indicator	5-W, 405
7.	Rural Route Designator	RR
8.	Rural Route Indicator	Usually a number
9.	Box Designator	Box, PO, PO Box
10.	Box Indicator	A letter or number

After running the Census standardization subroutine, a standard cleaning algorithm was applied to the parsed data. Street names were cleaned to remove all variations of the following: "unknown", "general delivery", homeless", "don't know", "none", "bad address", "moved", "returned", "undeliverable". Address fields were consolidated by the following rule: if street name (3) was missing, then street name (3) was set equal to "RR designator + RR indicator" (7 + 8) if nonmissing, else street name was set equal to "Box designator + Box indicator" (9 + 10) if nonmissing. This consolidation reduced the number of address fields used in the record linkage routines.

[50] For example, each of the following names was standardized to Katherine: Kate, Kathey, Kathryn, Kathy, Katie, Katrin, Katrina, Katy, Kitty, Catalina, Catherine, Cathey, Cathie, Cathryn, Cathy.

[51] The Census Bureau claims that the address standardizer (written by the Geography Division of the U.S. Census Bureau) provides substantially higher standardization rates than commercial products when applied to the types of addresses that are common in the U.S. (Correspondence from William Winkler.)

[52] Taken from U.S. Bureau of Census (undated).

Parameters for matching

The Census software requires user specification of three items: blocking variables, matching variables, and matching parameters. These items depend on the content and characteristics of the data being matched.

Blocking variables. As specified by the Census software documentation, "blocking is a division of an entire file into mutually exclusive subsets." If data are blocked on first initial of last name, then there will be 26 blocks of data. Every person with last name starting with "A" in file X is paired with every person having last name starting with "A" in file Y, and similarly for all other letters in the alphabet. But a person with last name starting with "A" may get married, take a new last name, and appear in two databases with different last names. This person will not be matched when data are blocked on last name because the records will not even be compared. Data are typically blocked, matched, and re-blocked with different blocking variables to account for errors in blocking variables or changes in those variables over time.

FSP and WIC records were processed separately for women and infants/children. The data were also processed separately for: a) December 2002 active caseloads (Florida, Iowa, and Kentucky), and b) the entire three-year period (Florida and Iowa).

Women and infants/children were each processed through four blocking rounds, defined by the following blocking variables:

- Block 1: Date of birth, 1st initial of last name
- Block 2: 1st five characters of last name, 1st initial of first name
- Block 3: 1st three characters of first name, county[53]
- Block 4a: Social security number (Florida and Kentucky)
- Block 4b: Telephone number, county (Iowa and Kentucky)

Iowa files were not blocked on SSN because SSNs were not available for WIC infants and children. For Iowa, block 4 was defined by phone number and county, with county included because telephone numbers did not include area codes. Kentucky data were processed twice: once with the same specifications as Florida (including SSN) and once with the same specifications as Iowa (SSN not used).

Blocking variables pair up records from two files according to an exact match of the blocking variables. For example, every FSP record with a date of birth equal to January 1, 2000 and a last name beginning with A, gets paired to every WIC record with those characteristics. For Florida infants/children, this block contains 15 WIC records and 10 FSP records resulting in 150 pairs. Every pair is evaluated by comparing all match variables, discussed below.

Over 95 percent of all matches were found in the first round, indicating that 95 percent of all matches matched exactly on date of birth and first initial of last name. Records not matched in the first round were processed through subsequent rounds of blocking.[54]

---

[53] The first four characters of first name were used for Florida, due to the large size of the files.

[54] The December 2002 caseloads were processed through 4 blocking rounds. The three-year caseloads were processed through 7 passes of 4 blocking rounds. The three-year caseloads contained multiple records per person reflecting

Matching variables and parameters. The Census software requires specification of match variables. For each match variable, the type of comparison and probabilities of agreement must be specified. Fourteen variables were evaluated in the match of FSP and WIC files. The variables are shown in table 16. Exact comparisons were used to match gender, race, some components of street address, ZIP code, and telephone number. String comparisons, which are phonetic and do not require exact matches, were used for all other variables.

Specification of match probabilities determines the contribution of individual variables to an overall match score for the paired records. Two probabilities must be specified: P_1, the probability that the variables match when the records truly belong to the same individual, and P_2, the probability that the variables match when the records do not belong to the same individual. A simple example of these probabilities (expressed as fractions) is P_1=1 and P_2=0.5 for gender, assuming no data entry errors for this data field.

Table 16

Matching variables and parameters

Match variable	Comparison type	Probability of agreement, if records match[a] (P_1)	Probability of agreement if records do not match[b] (P_2)
Last name	Special string comparison for last name; inversion option[c]	0.94	0.01
First name	Special string comparison for first name	0.93	0.01
Gender	Exact comparison[d]	0.99	0.50
Race	Exact comparison	0.95	0.32
Date of birth	Ordinary string comparison	0.99	0.01
SSN (Florida, Kentucky)	Ordinary string comparison	0.93	0.01
House number	Special string comparison for numeric address component	0.70	0.28
Street name	Ordinary string comparison	0.59	0.25
Apartment number	Exact comparison	0.45	0.20
Street suffix	Exact comparison	0.84	0.60
City	Ordinary string comparison	0.75	0.37
ZIP code	Exact comparison	0.72	0.28
Telephone number	Exact comparison	0.52	0.01
County	Exact comparison	0.93	0.06

[a] P_1 is expressed as a fraction and was based on the prevalence of matching data fields among Florida FSP and WIC matches identified by the Florida FSP/TANF/Medicaid ID that is present on both files.

[b] P_2 is expressed as a fraction and was based on the prevalence of matching data fields among the Cartesian product of that portion of a sample of Florida FSP and WIC records that excluded the matched sample used to calculate P_1.

[c] When last names do not match, the inversion option compares the last name on file A to the first name on file B, and vice versa, to see if the fields have been inverted.

[d] The exact comparison method requires that the variables being matched agree character-by-character in order to receive the full agreement weight.

changes in individual information. To make this match tractable, an FSP file with multiple records per person was matched, in turn, to 7 WIC files each having one record per person. WIC participants could have up to 7 WIC certification records and thereby appear in up to 7 files. After 7 passes, the highest match score for each WIC participant ID was identified as the best match.

The probabilities shown in table 16 were based on initial estimates obtained from Florida data. Certain matches can be identified in the Florida data for most persons with records in both files. The WIC database contains the State identifier that is used by the FSP/TANF/Medicaid system. This identifier was not used in our matching routines because a purpose of the project was to test probabilistic record linkage for these programs. But the identifier allowed us to check the results of record linkage and to generate estimates of match probabilities for individual variables.

P_1 and P_2 were estimated based on a single month of Florida data. All WIC records with nonmissing FSP/TANF/Medicaid ID were merged to the FSP file, and P_1 was calculated for that subset of matched records. The subset of matched records was then deleted from both FSP and WIC files, and the Cartesian product of unmatched FSP and WIC was used to estimate P_2.

As shown in table 16, the P_1 and P_2 match probabilities vary by data field. The P_1 probabilities reflect the reliability of the data field, the stability of information over time, and the potential for data entry errors. Gender and DOB have P_1 equal to 0.99; first and last name, race, and SSN also have P_1 over 0.90 percent. On the other hand, address fields have P_1 ranging from .45 to .84 and telephone number has P_1 equal to .52. This means that there is only slightly better than a 50 percent probability that a person's telephone number matches on the FSP and WIC data files.

P_2 is the probability that data fields agree when records do not belong to the same person. P_2 is low for items that are unique and/or take on many values within a population – such as SSN, name, and DOB. P_2 is high for items that have few values (race, gender, street suffix), or are not unique to specific persons in a population (ZIP code, city, street name).

The P_1 and P_2 values are used by the record linkage software to assign scores to individual data fields.[55] An overall score for a pair of records is equal to the sum of scores on individual fields. If probabilities P_1 and P_2 are far apart, the matching variable has a large distinguishing power. For example, a match of last name will contribute a large positive score to the overall score, while a non-match of last name will contribute a large negative score to the overall score. A match of gender contributes a *small* positive score, while a non-match of gender contributes a *large* negative score; a match of street name contributes a small positive score, while a non-match contributes a small negative score. Missing data for a data field on one or both files results in no contribution (positive or negative) to the overall score.

The blocking variables and matching variables completely specify the parameters for the Census software. In our implementation, we defined WIC as our base file, or File A. In each round of blocking and matching each record from the WIC file (File A) may be paired with more than one record from FSP (File B). All pairs were evaluated and scored. The WIC file (File A) was sorted by participant ID and overall match score and the pair with the highest score was kept. After each blocking round, data were manually reviewed to determine a cutoff for matches to be designated "certain", and all other records advanced to the next blocking round. After all rounds were completed, results from all rounds were combined, the WIC file (File A) was again sorted by participant ID and overall match score, and the pair with the highest score was kept.

[55] The score when a data field matches is based on an initial value equal to the base 2 logarithm of the ratio of P_1 and P_2. The score when a data field does not match is based on an initial value equal to the base 2 logarithm of the ratio of $(1-P_1)$ and $(1-P_2)$.

While 95 percent of all *matches* were matched in the first round, a large number of records – including ones never to be matched – advanced to subsequent rounds. After the final round, non-matches were categorized as "uncertain matches" or "certain non-matches." Uncertain matches were manually reviewed to develop SAS code for re-categorizing pairs as matches or non-matches.

SAS code was used in place of manual review of every "uncertain" pair. This processing identified pairs with low scores that met the following criteria:

- First name, date of birth, and SSN match.
- First name, last name, and SSN match.
- First name, date of birth, and telephone number match.

These pairs were recategorized as matches. The low scores, in these cases, were due to the "penalty" imposed by a non-matching last name or address.

The distribution of match scores, for all pairs that were determined to match, is shown in Figures 5–7. The matches below the cutoff are pairs that were recategorized. The graphs for Florida and Kentucky show a similar range of scores (up to 32) and two "spikes" above the mean. In these two States, the highest density of matches is located in the rightmost "spike." Iowa matches show a narrower range of scores (up to 25) and two "spikes" of approximately equal density. The lower mean scores in Iowa reflect the absence of SSN as a matching variable.

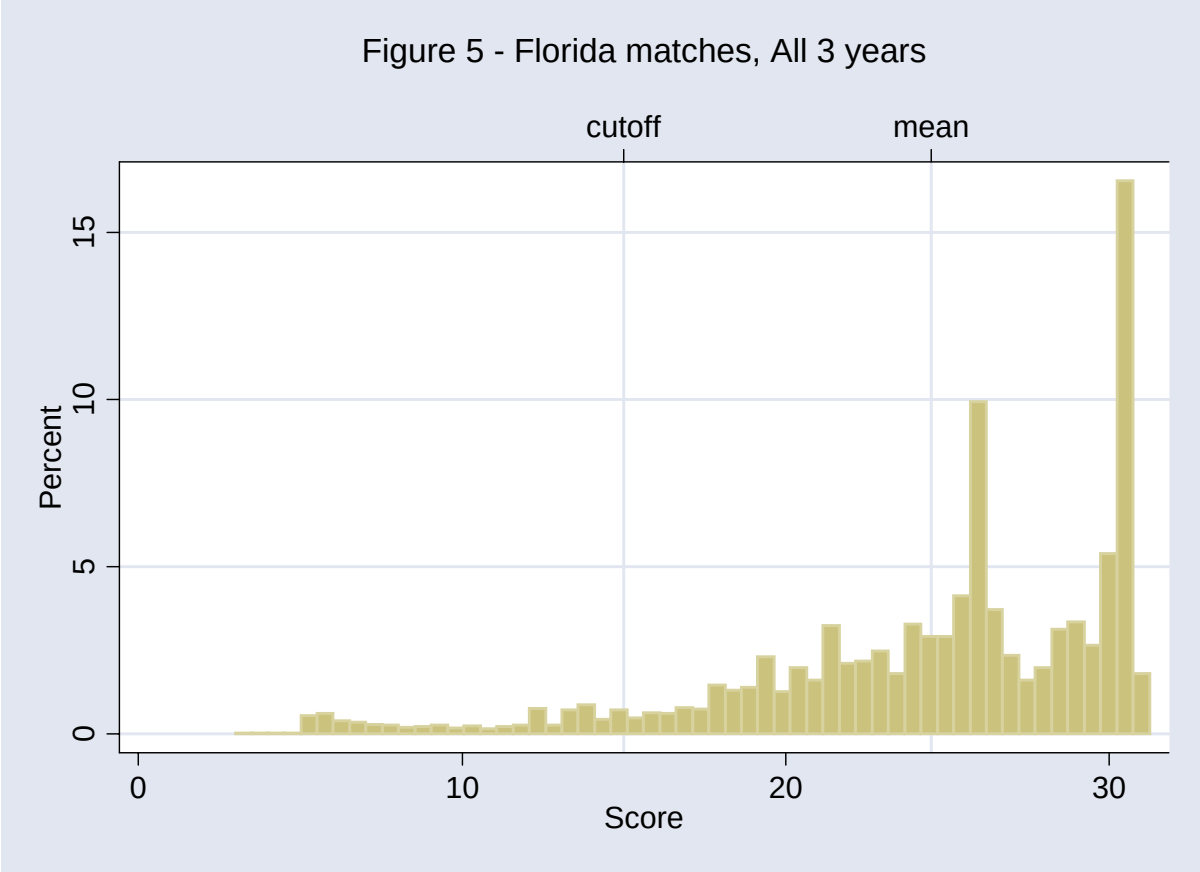

Figure 6 - Iowa matches, All 3 years

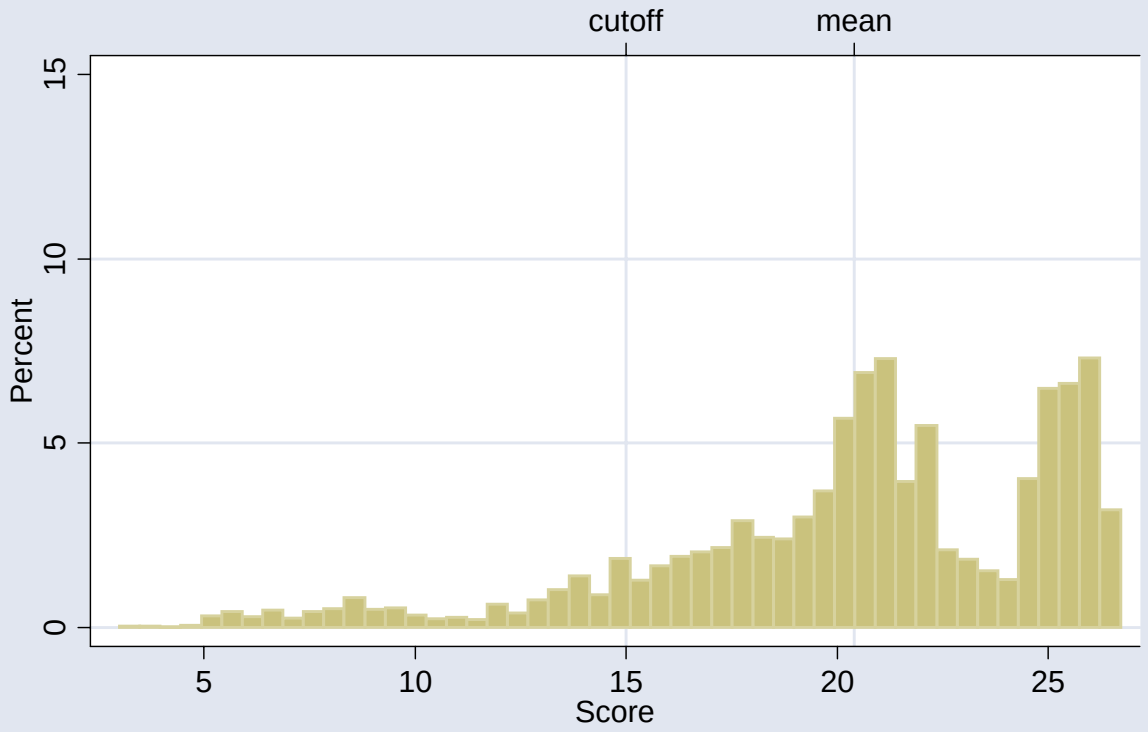

Figure 7 - Kentucky matches, December 2002