

Description of Matching Method for Linking Datasets in LINKAGE

Thomas E MaCurdy^{1, 2}

December 2024 (updated)

¹ *Stanford University*; ² *Acumen, LLC*

Abstract

This white paper summarizes the method implemented in the LINKAGE System to link individuals across data sources, comprised of medical information and services datasets (MI-Data) (e.g., claims, medical records, etc.) and Data Holder (DH) datasets (e.g., surveys, clinical trials, registries, etc.). Each data source contains privacy-protected data accessible within the LINKAGE System. The linkage algorithm derives from a probabilistic approach that scores matches of individuals' characteristics across diverse data structures, allowing for flexible forms of finder files with varying overlaps in identifying variables and differing reliabilities of these variables.

Acknowledgements

This research has been supported by contract GS10F0133S/140D0421F0687 funded by the National Institute on Aging (NIA) of the National Institutes of Health (NIH). Any opinions in this document do not necessarily represent the official position or policy of NIA or NIH. This paper greatly benefited from discussions and comments from Jonathan Skinner and Mir Ali as NIA reviewers. Thomas DeLeire, Dan Moran, and Sonam Sherpa contributed extensive comments and, along with Clint Cummins and Clifton Wolak, provided expert research assistance producing the software implementing algorithms and the empirical results reported in the paper.

Table of Contents

1. Introduction	1
2. Overview of Record Linkage Approaches	2
2.1 Framework Underlying Probabilistic Record Linkage	2
2.2 Characterization of Common Record Linkage Approaches	3
3. Description of the Probability Uniqueness-Based Linkage Method (PUBLIM)	4
4. Description and Categorization of Variables Included in Finder Files	6
5. Step 1: Designating Variable Categories and Match Quality Bands	7
6. Step 2: Identifying Pools of Candidate Matches	11
7. Step 3: Scoring and Selecting Unique Matches	12
7.1 Selecting Candidate Linkages Based on Highest Match Score	12
7.2 Specifying Score Thresholds for Assigning Linkages	13
7.3 Accounting for Missing Data for Variables in Finder Files	13
7.4 Illustration of Setting Type 1 Error Score Thresholds with Some Missing Data	14
7.5 Clerical Review and Assigning Linkages Match Confidence Levels	14
8. Information Delivered to Researchers About Linkage Files	16
References	17
Appendix A: Example of Scoring and Linkage Selections	20
Appendix B: Estimating the Probabilities of False Matches	22
Appendix C: Example of Setting Thresholds for Linkage Scores to Bound Probabilities of False Match Rates for Finder Files with Complete and Missing Data	24
Appendix D: Characterization of PUBLIM Approach in Fellegi and Sunter Framework	29

1. Introduction

The task of linking individuals across data sources involves a research analysis consisting of a sequence of steps aimed at (i) determining the persons in two datasets constituting candidate matches for being the same individual, and (ii) rating these candidates to assess the most likely selection for linkage deemed to be the same person. The linkage method developed in this paper, termed the Probability Uniqueness-Based Linkage Method (PUBLIM), does not adopt a conventional deterministic or probabilistic approach to linkage, which relies on pre-described or imputed weighting schemes to balance tradeoffs in discrepancies between individuals' finder-file identifying variables and prospective candidate characteristics. Instead, the linkage method summarized here implements a sequential investigative analysis integrating a variant of probabilistic matching that can incorporate knowledge about data quality in various steps, thus permitting the application of sophisticated criteria tailored to handle problematic cases.

This linkage approach implements three analysis steps to assign matches across data sources based on individuals' identifying variables provided in finder files:

Step 1: Designating variable categories and match qualities

Step 2: Identifying a pool of candidate matches

Step 3: Scoring and selecting unique linkages.

The approach provides transparency to researchers in understanding the reliability of linked IDs and their appropriateness for use in various aspects of their data analyses.

The linkage applications addressed in this white paper focus on the record-matching method implemented in the LINKAGE System to integrate individuals' circumstances across various privacy-protected data sources. The National Institute of Aging (NIA) supports LINKAGE in providing confidential data resources to approved research groups, enabling them to conduct a broad range of studies related to health and aging. The data sources in LINKAGE include administrative medical information and services datasets (MI-Data) (e.g., claims, medical records, etc.) and datasets (e.g., surveys, clinical trials, registries, etc.) maintained by Data Holders (DHs) in the LINKAGE Environment to make their confidential data accessible to the scholarly and policy communities while satisfying minimum data necessary criteria.

The output of this research process consists of DH linkage files that relate individual IDs across datasets, with each ID allotted a linkage score and one of three overall match confidences: *Strong*, *Fair*, and *Weak*. The determination of these confidence levels encapsulates a novel approach in PUBLIM for measuring the commonality and trustworthiness of the variables identifying individuals in the different finder files. The specific weightings of factors in the linkage research steps depend on the contents of finder files and the assessed accuracy of the various finder-file variables in the datasets. Documentation accompanies linkage files explaining the criteria implemented to assign linkages and to convey the degree of confidence in the match.

The following discussion describes the PUBLIM process for producing linkage files. Section 2 presents an overview of probabilistic matching approaches to provide context for the linkage method developed in this paper. Section 3 describes the PUBLIM algorithm for scoring the uniqueness of characteristic matches based on probabilistic methods. Section 4 presents an overview of the typical types of variables included in finder files. Section 5 discusses the elements of Step 1 in the linkage process; Section 6 outlines the criteria governing the execution of Step 2; and Section 7 covers the implementation of Step 3. Finally, Section 8 describes the information provided with the linkage files delivered to researchers.

2. Overview of Record Linkage Approaches

The following discussion outlines features of prominent probabilistic matching and linkage approaches proposed in the literature. These approaches specify algorithms for comparing the characteristics of persons in Dataset A and Dataset B by "scoring" candidate match person-pairs with the aim of deeming whether pairs represent the same individuals in the two datasets. To conduct this linkage analysis, approaches analyze finder files for Datasets A and B , providing information on K characteristics for individuals $i \in A$, $i = 1, \dots, N_A$ and individuals $j \in B$, $j = 1, \dots, N_B$. The finder-file characteristics $k = 1, \dots, K$ include such variables as persons' name, age, gender, and residence location. For each individual i in Dataset A , matching algorithms implement a probabilistic approach that produce measures of the closeness of the characteristics of candidates j in Dataset B most comparable to individual i . Depending on the values of closeness measures, analysts assign a linkage for individual i when a measure achieves a sufficiently high match confidence to justify registering individuals i and j as the same person.

2.1 Framework Underlying Probabilistic Record Linkage

The foundational work of Fellegi and Sunter (1969) (hereafter F&S) specifies the optimal scoring/matching rules when using probabilistic methods as the basis for linkage algorithms. F&S designate a comparison vector $\gamma_{ij} = \gamma_{ij}(\dots, \gamma_k(i, j), \dots)$ as gauging the closeness of the characteristics of individuals i and j in the finder files, with the elements $\gamma_k(i, j)$ measuring the closeness of the k -th characteristic. F&S interpret γ_{ij} as a randomly distributed vector across the (i, j) pairs, with the distributions of the element $\gamma_k(i, j)$ determining the properties of this multivariate distribution.

To specify an optimal matching rule, F&S define two key probability parameters, motivated by a Bayesian approach for reaching decisions based on available evidence. To specify these parameters, denote $Y_{ij}=1$ when $i \in A$ and $j \in B$ are the same person, and $= 0$ otherwise. The set $\{Y_{ij}=1\}$ includes the (i, j) pairs constituting true matches (i.e., the same person), and the set $\{Y_{ij}=0\}$ contains those pairs not true matches. F&S define the two probabilities:

$$m_{ij} = \Pr(\gamma_{ij} | Y_{ij} = 1) \quad u_{ij} = \Pr(\gamma_{ij} | Y_{ij} = 0) \quad . \quad (1)$$

The " m " and " u " parameters play a major role in computing posterior odds in Bayesian decision theory. In particular, a fundamental Bayesian formula yields the following relationship:

$$\left[\text{posterior odds of } (i, j) \in \{Y_{ij} = 1\} \right] = \left[\text{prior odds of } (i, j) \in \{Y_{ij} = 1\} \right] \times \frac{m_{ij}}{u_{ij}} \quad . \quad (2)$$

The m_{ij}/u_{ij} ratio corresponds to the Bayes Factor, and its information fully determines the posterior probability that the (i, j) pair is a true match. F&S prove that ranking the ratios m_{ij}/u_{ij} from their highest to the lowest values constitutes the optimal matching of individual pairs, with values falling below a lower threshold indicating the evidence is too weak to justify assigning this pair a linkage.

In practice, linkage algorithms rank the m/u ratios substituting estimates for these parameters. The set $\{Y_{ij}=1\}$ is unknown, and linkage methods replace this set with their inferred match set,

denoted here by $\{y_{ij}=1\}$ —comprised of those (i, j) pairs deemed linked by the match process.¹ Using this match set $\{y_{ij}=1\}$ and its complement $\{y_{ij}=0\}$, linkage approaches estimate the probabilities:

$$m_{ij} = \Pr(\gamma_{ij} \mid y_{ij} = 1) \quad u_{ij} = \Pr(\gamma_{ij} \mid y_{ij} = 0). \quad (3)$$

Sorting the ratio m_{ij}/u_{ij} estimates from highest to lowest, a matching algorithm selects the highest values as linkages. This "scoring" procedure produces the relationship:

$$\left[\text{posterior odds of } (i, j) \in \{y_{ij} = 1\} \right] = \left[\text{prior odds of } (i, j) \in \{y_{ij} = 1\} \right] \times \frac{m_{ij}}{u_{ij}}, \quad (4)$$

which corresponds to the probability of an (i, j) pair being a member of the inferred match set—in sharp contrast to being a true match.

Additionally, the optimality theory of F&S does not specify the form of comparison vectors and takes a form as predetermined. Consequently, rank orders of the ratios m_{ij}/u_{ij} across the different specifications of γ_{ij} invariably do not provide a basis for inferring a "better" or preferred match set.

2.2 Characterization of Common Record Linkage Approaches

The most widely used linkage methods construct the $\gamma_k(i, j)$ elements of the comparison vector γ_{ij} as discrete variables that classify similarities of characteristics into groupings. These approaches (e.g., Winkler (1988, 1989, 2002, 2014), Herzog et al. (2007), Christen et al. (2017), Enamorado et al. (2019), Splink (2024a)) distinguish a k -th characteristic of an (i, j) pair as falling into different groups ℓ (e.g., identical, similar, different) and then interpret characteristics within a given grouping as equivalent.

Viewed in the F&S framework, this formulation implies γ_{ij} possesses a multivariate multinomial distribution, with the $\gamma_k(i, j)$'s having their own univariate multinomial distributions.² Additionally, these algorithms typically assume mutual independence of the elements $\gamma_k(i, j)$. Relying on these assumptions, the approaches decompose the m/u ratios into the product

$$\frac{m_{ij}}{u_{ij}} = \frac{m_{\ell_1}}{u_{\ell_1}} \dots \frac{m_{\ell_k}}{u_{\ell_k}} \dots \frac{m_{\ell_K}}{u_{\ell_K}}, \quad (5)$$

where the subscript combinations (ℓ, k) designate $\gamma_k(i, j) = \ell$ with ℓ denoting the grouping assigned to the k -th characteristics for the (i, j) pair.

Relationship (5) motivates the scoring of match pairs to be calculated separately for each characteristic k of the match (i, j) pair and then aggregated to produce a total score for this pair. This popular scoring method allows for calculating the separate contribution of the closeness of each characteristic to the total score for a match (i, j) pair.

A major shortcoming of these approaches concerns the loss of identifying information when simply allocating characteristic comparisons to groupings with no distinctions accounted for within a

¹ This assumption mirrors settings in hypothesis testing whereby inference presumes estimates emulate the null hypothesis.

² Multinomial distributions encompass Bernoulli distribution, consistent with only two classifications such as agree and not agree.

group. Knowing individuals match on a common last name of Smith is less informative than knowing they match on the name Krzysztof, which is likely to be quite rare in an entire population. Knowing individuals match on residence in ZIP5 08701, the 5-digit zip code with the largest population located in New Jersey, is less informative than knowing they match on ZIP5 99762, the 5-digit zip code for residence in Nome, Alaska. In both cases, popular linkage algorithms classify these comparisons as "identical" and make no further distinctions.

Recognizing this shortcoming, various linkage approaches incorporate frequency-based adjustments to compensate for the rarity of characteristics within groupings (e.g., Xu et al. (2022), Splink (2024b), Gill et al. (1997), Yancey (2000), Zhu et al. (2009)). Generally, these approaches modify $m_{\ell k}/u_{\ell k}$ ratios by factors depending on the rarity of characteristics within groups. For example, Xu et al. (2022) propose conducting a two-step procedure that adjusts $m_{\ell k}/u_{\ell k}$ ratios by factors depending on probability ranges reflecting the agreement of matched pairs on their values of characteristic k . In essence, this approach introduces subgroups within groups to distinguish characteristics uniqueness and then conducts a conventional application of the F&S framework. Another example, Splink (2024b) proposes an approach involving expansion of match classifications to distinguish particular values of characteristics, such as creating separate categories for individual names and features of administrative ID values. Such adaptations of Splink substantially increase the number of cells and burden of calculating match scores which can lead to significant complexities in the stability of its computational algorithm.³ To avoid these computational complications, Splink offers an alternative method of altering scores after determining the linkage set by creating subgroups within classifications and increasing scores with more unique matches in the population; these ex-post adjustments refine Splink scores without changing linkage determinations. Still other examples introduce continuous similarity measures for comparing string data, but then transfer comparison values to categorical values representing different levels of agreement (e.g., Herzog et al., 2007; Sadinle, 2017; Enamorado et al., 2019), which again fall into the conventional F&S framework with multiple agreement classifications.

The adjustments of match weights implemented by existing frequency-based approaches typically do not result in substantial alterations in linkages. This outcome, combined with the computational burden incurred by these approaches, has led to them not commonly being used in linkage applications.

3. Description of the Probability Uniqueness-Based Linkage Method (PUBLIM)

The record linkage approach proposed in this paper operationalizes a probabilistic frequency-based matching approach that exploits all the identifying information contained in finder files to score matches and assign linkages. This approach distinguishes the extent of uniqueness of characteristic identifiers in datasets, which motivates naming this approach the Probability Uniqueness-Based Linkage Method (PUBLIM).

PUBLIM scores match pairs based on the comparison vector $\gamma_{ij} = \gamma_{ij}(\dots, \gamma_k(i, j), \dots)$ where $\gamma_k(i, j) = p_k(i, j)$ with

$$p_k(i, j) = \text{probability (share of) individuals } j \in B. \text{ in Dataset } B \text{ have characteristics } k \quad (6)$$

³ Such complexities can arise, for example, when increasing the number of cells in the multivariate multinomial distribution underlying Splink leads to cells with insufficient observations to estimate cell probabilities reliably, including having no observations resulting in parametric identification issues.

matching those of individual $i \in A$ in Dataset A .⁴

This method computes the total score of the candidate match pairs for individual i , applying the formula:

$$P_{ij} = \prod_{k=1}^K p_k(i, j) \quad P_i = \text{Min}_{j \in N_B} \{P_{ij}\} . \quad (7)$$

The probabilities $p_k(i, j)$ measure the degree of uniqueness of individual i 's characteristic k in the population of Dataset B . The rarer the characteristic, the smaller the $p_k(i, j)$. The more unique the characteristics of individual i in the overall population of Dataset B , the smaller the P_i .

Transforming this totaled probability into a log odds ratio produces a match score that increases the more unique a candidate match, with odds defined as:

$$O_i = \frac{1 - P_i}{P_i} = \text{ratio of the chances of mismatching to matching individual } i \text{ to the selected individual } j \text{ in Dataset } B.$$

Taking common logs produces:

$$S_i = \log(O_i) = \text{match score for individual } i. \quad (8)$$

A higher score signals a higher chance that the linked Dataset B member corresponds to the individual i in Dataset A .

As with most probabilistic linkage approaches, PUBLIM translates into the classic F&S framework. The elements $\gamma_k(i, j) = p_k(i, j)$ of the comparison vector γ_{ij} possess a continuous distribution with a domain comprised of the unit interval. The vector γ_{ij} possesses a multivariate distribution with its elements each having a continuous distribution over the unit interval. The totaled probability P_i follows a continuous univariate distribution over the unit interval. Finally, the match score S_i possesses a continuous univariate distribution with a domain comprised of positive numbers. The F&S theory encompasses comparison vectors and scores with continuous distributions. The PUBLIM approach does not need nor assume any independence of the elements $\gamma_k(i, j)$ of the comparison vector γ_{ij} . Appendix D presents the details interpreting the PUBLIM algorithm in the F&S framework and its optimality match score ranking.

An important advantage of PUBLIM concerns its computational efficiency. Its algorithms compute scores with one pass through the dataset, and this computation readily accommodates parallel processing. Alternatively, computation of estimates for other scoring methods—e.g., Fastlink developed by Enamorado et al. (2019) and implemented in Splink (2024a)—requires maximization of a likelihood function using an expectation-maximization (EM) algorithm or multiple iterative passes through the data to compute equilibrium parameter estimates. Such methods encounter heavy computational burden when linking individuals in large datasets.

The following sections explain the three analysis steps enabling efficient computation of match scores and using these scores to select linkages given the information on individuals' identifying variables provided in finder files. Section 4 characterizes variables included in finder files. Section 5 (Step 1) describes organizing finder-file information to create a systematic method for computing match probabilities $p_k(i, j)$. Section 6 (Step 2) outlines how to narrow the pool of viable candidates j in Dataset B to compute the lowest value of P_i . Finally, Section 7 (Step 3) describes using match scores to select linkage sets that infer which individual j in Dataset B most likely represents individual i in Dataset A .

⁴ This analysis presumes that Dataset B is larger than Dataset A (i.e., $N_B > N_A$).

4. Description and Categorization of Variables Included in Finder Files

As noted in the Introduction, the linkage applications constituting the focal point of this paper involve integrating privacy-protected data sources in the LINKAGE System. Accordingly, Datasets *A* typically represents DH datasets (e.g., surveys, clinical trials, registries, etc.) in LINKAGE, and Datasets *B* normally correspond to medical information and services datasets (MI-Data) (e.g., claims, medical records, etc.) available in LINKAGE to supplement and enhance the information available in DH datasets.⁵ The following discussion freely interchanges the terms Datasets *A* and DH datasets and the terms Datasets *B* and MI-Data datasets.

Table 4.1 lists the types of variables contained in DH finder files intended to match the identification of individuals across datasets, with varying subsets of these variables incorporated in the finder files provided by different DHs. The most common categories of variables include Name, Gender, Administrative IDs, Age Information, and Residence Information, with a file giving various versions of the data elements shown in these categories. Occasionally, DH finder files include variables in the Health Status and Other Variables categories, which can provide efficient linkage information when a DH dataset requires individuals to satisfy specific criteria, such as having a particular health condition (e.g., heart condition, dementia) or being classified in a certain socioeconomic status (e.g., low income).

Table 4.1: Finder-File Variables for Linking Individuals Across Data Sources

Category	Variable	Description
Name	LN	Last Name
	FN	First Name
	MI	Middle Initial or Name
Gender	Gender	Gender
Administrative IDs	SSN	Social Security number
	ITIN	Individual Taxpayer Identification Number
	HIC	Health Insurance Carrier
	MBI	Medicare Beneficiary Identifier
	HIID	Health Insurance ID
	MID	Medicaid ID
Age Information	AGE	Reported age
	YOB	Year of Birth
	MOB	Month of Birth
	DOB	Day of Birth
	YOD	Year of Death
	MOD	Month of Death
	DOD	Date of Death
Residence Information Residence Information	State	State of Residence
	County	County
	City	City
	Address	Reported address
	ZIP9	9-digit ZIP Code
	ZIP5	5-digit ZIP Code
	ZIP3	3-digit ZIP Code
Health Status	DIAG	Specific diagnoses
	CD	Chronic disease
	DIS	Disability Indicator
	PREG	Pregnancy Indicator
	PROC	Procedures (e.g., hospitalization)

⁵ A prominent source of the MI-Data includes the universe of Medicare and Medicaid claims and assessment data maintained by the Centers for Medicare & Medicaid Services (CMS).

Category	Variable	Description
Other Variables	OV1	Email
	OV2	Phone number
	OV3	Primary language code
	OV4	Marital status
	OV5	Household size
	OV6	English proficiency code
	OV7	Income code
	OV8	Immigration status
	OV9	Location Service Delivery

Given a finder file including the variables and associated categories in Table 4.1, linkage algorithms measure the commonality of the variables identifying persons in the different datasets to select unique matches representing the same individuals in these datasets. The processes followed in linkage steps depend on the trustworthiness and discriminating power of finder-file variables, which can result in dynamic and hierarchical analyses to discover the most credible linkages.

5. Step 1: Designating Variable Categories and Match Quality Bands

The first step in implementing a linkage algorithm consists of organizing the finder file variables into the categories used to check for commonality among individuals and specifying the notion of closeness in each category deemed to assign a match. This structure establishes the framework for measuring the closeness of the characteristics of Dataset *B* members with those supplied for any finder-file individual.

To explain the PUBLIM approach, Table 5.1 illustrates a typical list of variables included in DH

Table 5.1: Common Finder-File Variables for Linking Individuals Across Data Sources

Category	Variable	Description
Name	LN	Last Name
	FN	First Name
Administrative IDs	SSN	Social Security Number
	HIC	Health Insurance Carrier
	MBI	Medicare Beneficiary Identifier
Gender	Gender	Gender
Age Information	YOB	Year of Birth
	MOB	Month of Birth
	DOB	Day of Birth
	YOD	Year of Death
Residence Information	State	State of Residence
	ZIP5	5-digit ZIP Code

finder files intended to match the identification of individuals across datasets. The second column lists the finder-file variables, and the first column presents our classification of these variables into variable categories used to measure comparisons. In practice, the formulation of variable categories involves data cleaning procedures and can include multiple finder-file variables, such as several types of Admin IDs (e.g., SSNs and Health Insurance IDs), more than one value of the same Admin ID (e.g., three SSNs), several names, and a list of residence locations.

Table 5.2 (page 9) specifies varying stringencies of closeness used in our linkage algorithm to register a match in a variable category, comparing the characteristics of an individual in a finder file and a member of Dataset *B*. In this table, the term "match quality band" designates the different measures of closeness, with a lower-numbered band constituting a closer match and a higher-

numbered band relying on weaker criteria to assign a match. A higher band typically produces a larger number of Dataset *B* members as matches for a variable category. Constructing match quality bands can incorporate external knowledge about the accuracy of information provided for variable categories in finder files.

Inspection of Table 5.2 reveals that for the Name category, Match Quality Band 1 requires exact matching on both first and last names, while Band 2 also allows for nicknames.⁶ The remaining bands for the Name category allow for wrong or missing names. Band 3 allows for fuzzy matches of last names, measured by either the Levenshtein distance (Levenshtein, (1965)) or the Jaro–Winkler distance (Jaro (1989)). The following examples illustrate the computation of Levenshtein distances equaling 1 (on the left) and 2 (on the right):

1. Harnry → Harry
 - Delete "n" in the middle of Harnry

1. Harnry → Hanry
 - Delete "r" in the middle of Harnry
2. Hanry → Henry
 - Replace "a" with "e" in the middle of Hanry.

For these two examples, the Jaro–Winkler distance, a string metric measuring an edit distance between two sequences, computes the difference values:

- Harnry → Harry
- Jaro-Winkler distance is 96.1%

- Hanry → Henry
- Jaro-Winkler distance is 76.5%

Band 4 also admits names that sound similar, as measured using either Soundex or Metaphone codes. The accompanying table presents three example first names, with the corresponding Soundex and Metaphone codes. Table 5.2 assumes use of the Jaro–Winkler distance and Soundex.

Name Matches Based on Soundex and Metaphone

Scenario	Name	Soundex Code	Metaphone Code
(1) Both Soundex and Metaphone match	Derek	D62	TRK
	Derrick	D62	TRK
(2) Soundex matches, Metaphone does not match	Leanne	L5	LNN
	Lean	L5	LN
(3) Soundex matches, Metaphone does not match	Haslam	H245	HSLM
	Heislen	H245	HLN

Finally, Band 5 ignores any information about names in matching, which results in the inclusion of all members of the Dataset *B* as matches—i.e., anyone in the population could be a match. Assignment to this Band penalizes mismatches in probabilistic scoring by setting $p_{ki}=1$, the worst possible probability in scoring.

For the Administrative ID category in Table 5.2, Match Quality Band 1 requires an exact match on any of the three Admin IDs (SSN, HIC, MBI). Band 2 designates a match if any of the Admin IDs for a finder-file individual falls within a Levenshtein distance of 1 for a comparable ID of a Dataset *B* member, and Band 3 records a match if the Admin IDs differ by no more than a Levenshtein distance of 2. Finally, Band 4 nullifies Admin ID information by including all Dataset *B* members as matches.

For the Gender category, the match quality bands either require an exact match (Band 1) or ignore any information about gender (Band 2).

For the Age category, Band 1 requires exact agreement on the date of birth to assign a match. Band 2 allows for only the year of birth to agree to assign a match. Band 3 relaxes these criteria by designating matches when years of birth fall within one year. Band 4 further relaxes these criteria by allowing matches when years of birth are precisely one decade of one another. Lastly, Band 5 assigns matches ignoring Age information, lowering scoring to its worst contribution.

Finally, for the Residence category, Band 1 requires an exact agreement of 5-digit zip codes to

⁶ Datasets providing lists of nicknames include “en.wiktionary.org/wiki/Appendix: English_given_names” and “github.com/carltonnorthern/nicknames.”

designate a match. Band 2 relaxes this rule and requires only state of residence to match exactly. Lastly, Band 3 assigns matches irrespective of residence information, as if the effective finder-file observation has no information on Residence.

Table 5.2: Specification of Match Qualities for Variable Categories

Category	Variables	Match Quality Band						
		1	2	3	4	5	6	7
Name	First Name (FN), Last Name (LN)	FN and LN both match exactly	FN (including nicknames) and LN match	LN matches with Jaro–Winkler distance $\geq 94\%$. FN – including nicknames matches	FN Soundex match and LN exact match	Name mismatch	N/A	N/A
Administrative IDs (Admin ID)	Social Security Number (SSN), Health Insurance Carrier (HIC), Medicare Beneficiary Identifier (MBI)	Any Admin ID (SSN, HIC, or MBI) exact match	Any Admin ID matches within a Levenshtein distance ≤ 1	Any Admin ID matches within a Levenshtein distance ≤ 2	Admin ID mismatch	N/A	N/A	N/A
Gender	Gender	Gender match	Gender mismatch	N/A	N/A	N/A	N/A	N/A
Age Information	Birth Date	Birth date match	Year of Birth (YOB) and Month of Birth (MOB) match	YOB matches	Day of Birth (DOB) and MOB match exactly, YOB is within ± 10	YOB is within ± 1	YOB is ± 1 decade exactly	Age mismatch
Residence Information	5-Digit Zip Code (ZIP5), County, State	ZIP5 match	County	State	Residence mismatch	N/A	N/A	N/A

6. Step 2: Identifying Pools of Candidate Matches

For each individual i included in a DH finder file, constructing the candidate pool of potential matches consists of identifying those members in the MI-Data with characteristics matching those of individual i . Given a candidate pool for each variable category, selecting a single Dataset B member constituting individual i —i.e., the linkage for individual i —amounts to evaluating the best extent of agreement across all variable categories between these persons.

Keeping the candidate pool as small as possible improves efficiency in the minimization computation specified in (7) calculating P_i . For a variable category, a MI-Data member constitutes a match candidate for individual i when this person's attributes satisfy the criteria for a match quality band for that category, as specified in Table 5.2. Construction of a candidate pool for a quality band includes all individuals falling into all the smaller numbered bands for that variable category, resulting in an ever-expanding pool for the higher numbered bands. An analysis generates a smaller pool using the most restrictive (best) quality band in a variable category that produces a non-empty set of candidates.

Establishing candidate match pools in this way for each variable category does not guarantee that any single member of Dataset B appears in every category set. In such instances, an algorithm must strategically relax quality bands for some variable categories until at least one MI-Data member occupies every category pool.

To construct candidate match pools efficiently, the PUBLIM algorithm performs the following three actions for each individual j included in a DH finder file:

- (1) Start with the lowest number of match quality band for each variable category and search for matches in the MI-Data. If none are found, move to the next lowest band number and stop when the algorithm produces a non-empty candidate set for the variable category.
- (2) Determine whether any MI-Data member appears in every category-candidate set. If yes, stop and move to Step 3 of the linkage process (i.e., the selection step).
- (3) If no intersection of members occurs across category candidate sets, then relax match quality bands until one achieves a non-empty intersection, focusing first on those categories most discriminating in identifying unique MI-Data members.

Note that relaxing a quality band in a variable category and reaching the highest number makes the candidate pool for that category the entire MI-Data population, eliminating information about this category in selecting a linkage for individual i .

Determining the hierarchy for performing Action 3 depends on the reliability of the variables included in a DH finder file and their corresponding MI-Data counterparts in identifying unique matches in the various variable categories. Strong linkages typically do not go beyond Action 2, with most assigning the single MI-Data member in the intersection of the minimal-sized candidate sets constructed for the categories associated with individual i . In those cases with no member in the intersection across sets, relaxation of quality bands in Action 3 should first occur for those categories (i) less prone to providing inaccurate information and (ii) most discriminating in limiting the number of viable match candidates. Consider, for example, a finder file with the variables listed in Table 5.1 and the match criteria presented in Table 5.2. Relaxing match quality in the Name category beyond Band 4 would ensure low credibility in any assigned linkage, whereas relaxing match criteria to Bands 3 or 4 for Age would imply less diminishment of credibility with known unreliability of precise birth dates reported in either a DH finder file or a MI-Data dataset. Similarly, for the Residence category, relaxing its match quality bands may have little effect on linkage credibility if one knows

that individuals' locations are fluid in data sources.

7. Step 3: Scoring and Selecting Unique Matches

The completion of Step 2 produces a candidate pool for each individual i in a DH finder file comprised of members in the MI-Data source that appear in the candidate sets for every variable category constructed for this individual. Linkage algorithms balance various discrepancies in characteristics, weighting the effectiveness of variable categories in detecting the MI-Data member with identifying characteristics consistent with those of individual i , which might include adjusting for external knowledge about the accuracy of information provided for variable categories in the data sources used for linkage.

7.1 Selecting Candidate Linkages Based on Highest Match Score

Applying the probabilistic scoring framework outlined in Section 3 offers an approach for identifying which candidate in MI-Data has the highest likelihood of being a match to individual i in the DH finder file. In our application, Dataset B corresponds to that partition of the MI-Data population tailored to represent the same population from which the DH dataset (Dataset A) drew its sample. Scoring, then, measures the likelihood of a member of the MI-Data source having the same combination of characteristics as individual i in the DH finder file.

To calculate values for the $p_k(i, j)$ probabilities in Section 3 measuring the likelihood of a candidate j 's characteristics k matching those of individual i , define k_q as the quality band (q) used to designate the match for the candidate. A measure of $p_k(i, j)$, denoted shorthand by p_{ki} , becomes:

$$p_{ki} \equiv p_k(i, j) = \text{share of individuals } j \in B \text{ in Dataset } B \text{ with } k \text{ characteristic falling into same band } k_q \text{ candidate pool as individual } i \text{ in Dataset } A.$$

In situations where Dataset B comprises an entire population, such as when B comprises the entire Medicare population, the above measure is the actual probability, not an estimate, that individual i has the same characteristics as members of the overall population.

Scoring the overall match quality of a candidate amounts to computing p_{ki} for each characteristic k using the most stringent quality band encompassing this person. Weakening a match quality band often generates a higher number of 1-to-many matches for the category, which increases p_{ki} . Probabilistic scoring assigns a candidate the lowest possible p_{ki} for each variable category, which increases this candidate's match score. For complicated cases, a linkage algorithm can entertain candidate pools for multiple quality bands in a variable category to ensure credible candidates remain in linkage consideration due to missing and inaccurate data and weak matches in other variable categories. Given the p_{ki} 's calculated in this way, the formula in Section 3 for P_i calculates the probability that a member of the MI-Data dataset has the same combination of match characteristics as the candidate, which in turn matches the combination of characteristics identifying individual i in Dataset A finder file. The formula in Section 3 for S_i computes the match score for individual i , interpretable as the log odds of selecting the correct Dataset B member as being individual i .

A common linkage approach assigns a match to Dataset A individual i by selecting the Dataset B person with the highest match score. The Dataset B candidate group for finding the highest score initially comprises those individuals included in the intersection of the match pools constructed by the actions described above. The highest-scored person in this group sets the threshold (i.e., minimum) value for considering other candidates. For any other Dataset B person, outside the initial intersection group, to have a score potentially exceeding this threshold, the candidate must be in a higher quality band than the threshold person for at least one characteristic, typically comprising a relatively small

set of individuals. Scoring these candidates may involve relaxation of bands for some characteristics to include them in a new intersection of all category sets, but this individual may be a very high-quality match for other categories, yielding a score exceeding the initial threshold and setting a new higher threshold.

Appendix A presents a simple example of the scoring and matching approach described above, considering the linkage of four individuals in one dataset against five in another using a finder file containing information on three variable categories.

7.2 Specifying Score Thresholds for Assigning Linkages

Arraying calculations of the match scores for each Dataset *A* individual *i* from the largest to the smallest establishes the hierarchy for assigning linkages. Designating a level as the minimally acceptable score, all individuals with scores exceeding this value acquire a linkage in the Dataset *B*.

A natural approach for setting an acceptable score threshold relies on bounding the probability of false matches (Type 1 errors) associated with this threshold. Appendix B outlines an estimation method for calculating the probabilities of Type 1 errors using the finder file variables implemented in the linkage of DH dataset. The percentiles of this estimated Type 1 error distribution specify the score cutoffs corresponding to the probabilities of assigning a false match. For example, the 95th percentile of this distribution associates scores above this threshold with a 5% Type 1 error rate. Using these estimates, a linkage operation can set the score cutoff for assignments by bounding its probability of a Type 1 error.

Appendix C illustrates an application of the estimation framework outlined in Appendix B to set score thresholds for assigning linkages to achieve desired Type 1 error rates.

7.3 Accounting for Missing Data for Variables in Finder Files

Finder files with missing (absent) data for some variables and individuals require setting the linkage threshold values separately for score distributions reflecting the pattern of missing finder file data. The Type 1 error score distribution for scores computed using missing data for finder file variables need not be the same as one estimated using complete data.

The linkage approach described above readily accommodates accounting for missing data patterns in finder file variables. Given a DH finder file, first determine the distinct distributional patterns of missing data. For each pattern, then calculate the score distributions for DH observations characterized by this pattern. One may then estimate the Type 1 error score distribution consistent with this pattern using the procedure described in Appendix B. Finally, using the percentiles of this constructed Type 1 error score distribution, one can set threshold values for the score distribution targeted to attain the desired false match error rate.

Performing this exercise for each pattern and calculated score distribution by applying a common Type 1 error percentile yields a linkage algorithm that achieves a false match error rate corresponding to this percentile. In sharp contrast to other linkage methods, this approach accounts for missing finder-file data without relying on distributional assumptions relating missing data and other characteristics information available in the finder file. Linkage methods in the literature typically assume that missing finder-file data occurs independently of the informational content of finder-file characteristics with complete data, generally recognized as a dubious assumption.

7.4 Illustration of Setting Type 1 Error Score Thresholds with Some Missing Data

Appendix C explains the separate calculations needed to estimate score thresholds for assigning linkages to achieve a targeted Type 1 error rate for a finder file with complete data for a segment of the DH population and missing finder-file data for another segment of the DH population. This appendix summarizes results produced by the PUBLIM approach applied to linking members of the Health and Retirement Study (HRS) to individuals in the entirety of the Medicare enrollment population; HRS constitutes a prominent DH dataset currently maintained in the LINKAGE Environment. The HRS finder file includes the variable categories in Table 5.1, and the linkage exercise uses the match quality brackets in Table 5.2. The HRS finder file has complete data for virtually all observations, except for about 16% missing data on Admin IDs.

Table C.1 in Appendix C presents the breakdown of HRS matches to Medicare classified by candidate pools for the different match quality bands, providing a sense of the specific types of matches and their corresponding distribution of *S* scores measured by percentiles. As seen in the first row of Table C.1, nearly 60% of the observations in the HRS finder file—18,911 out of a total 31,833—have exact matches for all variable categories. Not surprisingly, this group receives high match scores, ranging from 21.2 for the 1st percentile to 26.2 for the 99th. (As noted above, differences in *S* score arise among exact matches due to differences in the likelihood (rarity) of matches within the same categories, such as matching on a common last name of Smith versus the rare name of Krzysztof.) In contrast, categories with multiple mismatched variables typically exhibit *S* scores falling below 10.

Tables C.2 and C.3 in Appendix C present the score thresholds aligned with achieving specified Type 1 error rates, recognizing differences in the informational content between the complete and missing data cases. Because missing data arises only for Admin IDs for a nontrivial share of HRS observations, Table C.3 provides separate calculations for Type 1 Error thresholds for those segments of the HRS finder file with and without data on Admin IDs. As expected, the score cutoff thresholds attaining a given Type 1 error rate are invariably higher for the complete data cases than the missing data cases. As revealed in rows 1 and 2 of Table C.3, selecting a 5% Type 1 error rate implies setting the score threshold at 11.6 for the complete data cases and 10.7 for the incomplete data cases; selecting a 1% rate yields score thresholds of 12.9 for complete data and 12.0 for incomplete. While these threshold differences may appear small, Appendix C explains how these differences impose linkage penalties on cases with mismatches on AdminIDs compared to cases with missing AdminIDs. Using the 5% Type 1 error thresholds, mismatched-data cases are over 20 percentage points less likely to be linked than missing-data cases.

7.5 Clerical Review and Assigning Linkages Match Confidence Levels

Linkage operations invariably conduct clerical reviews to assess the credibility of analytical approaches to assign linkages. These reviews may reassign some questionable linkages, with clerical intervention determining the match status. As part of this review, our group typically assigns a confidence level to each linkage, classifying them as "*Strong*", "*Fair*", or "*Weak*" to guide researchers in using linkage files produced by the above approach.

As an initial construction of these classifications, we categorize PUBLIM scores falling in the range with lower than a 1% false match rate (Type 1 error) as Strong; scores with a false match rate between 1-5% as Fair; and scores with a false match rate between 5-10% as Weak.

The clerical review can result in some linkages being reclassified, with scores assigned to Strong, Fair, and Weak falling outside the above initial ranges. This review consists of face-validity checks of

match cases with scores near the thresholds to assess their credibility. In instances where the selection appears optimal, the assigned confidence level remains unchanged. If the confidence level appears misassigned, it is adjusted accordingly. If an alternative match candidate with a competing score seems to constitute a more credible selection, the linkage is switched to the alternative candidate with an assessed confidence level. Researchers can use these confidence levels in configuring samples appropriate for their studies or designate their own score thresholds to formulate their samples.

Examples where clerical reviews may result in reclassification include matching an Admin ID or a Name attribute unique in the entire Dataset *B* population, which produces a high score, even when some other key attributes are mismatched (e.g., name is entirely different). In another example, matching on the exact address if it were available in a finder file and MI-Data would be highly distinguishing in identifying individuals, and can select a candidate based on scoring even when names do not match, and over an alternative candidate with a matching name and missing address. In yet another example, probabilistic scoring aims to penalize a "mismatch" in a characteristic more severely than the penalty associated with a missing value for this characteristic. This assignment presumes in the mismatch case that this designation represents errors in the entries of variables in either the DH or MI-Data databases; in some cases, such an error could trigger a clerical review. Finally, scoring relies on multiplying probabilities to calculate joint probabilities, justified by assuming independence in the informational content and reporting accuracy across characteristics. This assumption is violated for highly correlated finder file variables, such as those grouped into variable categories listed in Table 2.1. Our approach mitigates this criticism by matching variable categories and not on the individual variables making up these categories.

Before finalizing linkage files, our group shares findings with the DH supplying a finder file to seek their assessments of the criteria used to select linkages and assigned confidence levels. Our linkage analysis provides output information to DHs summarizing the overall match rate (e.g., % of finder file observations linked to unique matches), listing the scores and confidence levels for linked individuals, and identifying problematic linkage selections. As needed, our team breaks out matching statistics based on finder file ID characteristics (e.g., IDs with age greater than 65). Additionally, for assignments requiring significant judgment in selecting options, our group provides statistics in meetings with DHs showing the impacts of alternative options on matching rates and assignments. Any unresolved or problematic matches receive further investigation and discussions with DH authorities to determine how to handle these cases. In our experience, DH authorities often can provide additional finder file variables or assessments of the reliability of specific finder-file variables that permit the resolution of any Strong and Fair candidate cases (e.g., non-unique matches). DHs may also customize the criteria for selecting unique matches among candidates, and have the option to drop cases with matches too close to select, or to add unmatched IDs into a pool for re-runs to see whether any existing matches improve.

Our team interacts closely with each DH to learn about the reliability of variables in DH finder files and features of the data sources relevant for gauging the certainty of ID candidate linkages. Other candidates for addition in this linkage classification may include matches provided in historical DH linkage files that failed to achieve the same matches in the current file, which would allow researchers to use the earlier linkage files to update findings maintaining time-consistent linkages; such matches typically receive a Weak confidence level designation to signal inconsistencies across historical linkages. At their discretion, DHs may decide to include IDs with only Strong or Fair in linkage files, or not even distinguish between these classifications.

8. Information Delivered to Researchers About Linkage Files

Upon completion of the matching process and any customization, the LINKAGE team produces a DH Linkage File and documentation describing the assignment of the match confidence levels recorded in the file for each ID. This information provides researchers transparency in understanding the confidence levels assigned to linked IDs.

Table 8.1 presents the structure and contents of a DH Linkage File. Researchers wanting to integrate data with only a Strong confidence level of individual identification across data sources would select only those observations with a "1" in the second column of the table; researchers

Table 8.1: Structure of DH Linkage File

DH Individual ID	Confidence Level	Linkage Score	Type I Error Percentile
ID #	1, 2, or 3	0+	0+

choosing a Fair (or better) confidence level would select those observations with a "2" in the second column; and, finally, those choosing the broadest criteria (i.e., including Strong, Fair, and Weak) for linkage would select observations with a "3" in the second column, which comprises all observations in the Linkage File. The final column includes the percentile score of the Type 1 error associated with the corresponding ID, calculated consistent with population composition and patterns of missing finder-file values used to compute the false match score distribution for this ID.

The documentation accompanying DH Linkage Files also contains summaries of Tables 5.1 and 5.2 applicable for specifying the confidence of linkage designations. Knowledge of these tables entirely explains the criteria for assigning the three linkage groups designated in the file. Researchers can refer to these specifications to determine which linkage confidence level best applies to their various studies.

References

- Aleshin-Guendel, S. & Sadinle, M. (2023). Multifile partitioning for record linkage and duplicate detection. *Journal of the American Statistical Association*, 118(543), 1786–1795.
- Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, 17(18), 6937.
- Belin, T. R. & Rubin, D. B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430), 694–707.
- Bezin, J., Duong, M., Lassalle, R., Droz, C., Pariente, A., Blin, P. & Moore, N. (2017). The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. *Pharmacoepidemiology & Drug Safety*, 26(8), 954–962.
- Buckles, K., Haws, A., Price, J., & Wilbert, H. E. B. (2023). Breakthroughs in historical record linking using genealogy data: The Census Tree Project (No. w31671). National Bureau of Economic Research.
- Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '08*. (pp. 151–159). Association for Computing Machinery.
- Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.
- Christen, P., & Winkler, W. E. (2017). Record linkage. In *Encyclopedia of machine learning and data mining*. (pp. 1066–1075). Springer.
- Copas, J. B. & Hilton, F. J. (1990). Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3), 287–320.
- Delluc, A., Tromeur, C., Le Ven, F., Gouillou, M., Paleiron, N., Bressollette, L., Nonent, M., Salaun, P., Lacut, K., Leroyer, C., Le Gal, G., Couturaud, F., & Mottier, D. (2016). Current incidence of venous thromboembolism and comparison with 1998: A community-based study in Western France. *Thrombosis and Haemostasis*, 116(5), 967–974.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series: B (Methodological)*, 39(1), 1–38.
- Enamorado, T. (2018). Active learning for probabilistic record linkage. Social Science Research Network (SSRN).
- Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2), 353–371.
- Fellegi, I., & Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210.
- Fatemi, B., Kazemi, S. M., & Poole, D. (2018). Record linkage to match customer names: A probabilistic approach. *arXiv preprint arXiv:1806.10928*.
- Fortini, M., (2020). An improved Fellegi-Sunter framework for probabilistic record linkage between large data sets. *Journal of Official Statistics*, 36(4), 803–825.
- Goldstein, H., Harron, K., & Cortina-Borja, M. (2017). A scaling approach to record linkage. *Statistics in Medicine*, 36(16), 2514–2521.
- Gill, L. (1997) "OX-LINK: The Oxford medical record linkage system. In *Record Linkage Techniques 1997: Proceedings of an International Workshop and Exposition* (pp.15–33) Washington, DC: National Academy Press.

- Grannis, S., Overhage, J., Hui, S., & McDonald, C. (2003). Analysis of a probabilistic record linkage technique without human review. *AMIA Annual Symposium Proceedings*, 259–263.
- Hejblum, B., Weber, G., Liao, K., Palmer, N., Churchill, S., Shadick, N., Szolovits, P., Murphy, S., Kohane, I., & Cai, T. (2019). Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Scientific Data* 6, 180298.
- Herzog, T., Scheuren, F., & Winkler, W. (2007). *Data Quality and Record Linkage Techniques*. Springer-Verlag, New York.
- Hof, M., & Zwinderman, A. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine* 31(30), 4231–4242.
- Hofert, M., & Mächler, M. (2016). Parallel and other simulations in R made easy: An end-to-end study. *Journal of Statistical Software*, 69(4), 1–44.
- Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association*, 84(406): 414–420; and Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 354–359.
- Kim, G., & Chambers, R. (2012). Regression analysis under incomplete linkage. *Computer Statistics & Data Analysis*, 56(9), 2756–2770.
- Lahiri, P., & Larsen, M. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469), 222–230.
- Larsen, M., & Rubin, D. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453), 32–41.
- Lee, D., Zhang, L., & Kim, J. (2020). Maximum entropy classification for record linkage. *arXiv:2009.14797*.
- Li, P., Dong, X., Maurino, A., & Srivastava, D. (2011). Linking temporal records. In *Proceedings of the VLDB Endowment*, 4(11), 956–967.
- Mamun, A., Aseltine, R., & Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. *PLoS ONE* 11(4), e0154446.
- Meng, X., & Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80(2), 267–278.
- Noboa, S., Mottier, D., Oger, E., & EPI-GETBO Study Group (2006). Estimation of a potentially preventable fraction of venous thromboembolism: a community-based prospective study. *Journal of Thrombosis and Haemostasis*, 4(12), 2720–2722.
- Prindle, J., Himel, S., & Putnam-Hornstein, E. (2023). An open-source probabilistic record linkage process for records with family-level information: Simulation study and applied analysis. *PLoS ONE* 18(10), e0291581.
- Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), 600–612.
- Sayers, A., Ben-Shlomo, Y., Blom, A., & Steele, F. (2015). Probabilistic record linkage. *International Journal of Epidemiology*, 45(3), 954–964.
- Splink. (2024a). Data linkage at scale. *Splink Documentation*. <https://moj-analytical-services.github.io/splink/>.
- Splink. (2024b). Term-frequency adjustments. *Splink Documentation*. https://moj-analytical-services.github.io/splink/topic_guides/comparisons/term-frequency.html.

- Steorts, R., Hall, R., & Fienberg, S. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516), 1660–1672.
- Tancredi, A., & Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics* 5(2B), 1553–1585.
- Tuppin, P., Pestel, L., Samson, S., Cuerq, A., Rivière, S., Tala, S., Denis, P., Drouin, J., Gissot, C., Gastaldi-Ménager, C., & Fagot-Campagna, A. (2017a). Poids humain et économique des cancers en France en 2014, les données du Sniiram. *Bulletin du Cancer* 104(6), 524–537.
- Tuppin, P., Rudant, J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., de Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., Gissot, C., Weill, A., & Fagot-Campagna, A. (2017b). Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Revue Épidémiologie et de Santé Publique* 65(4), 146–167.
- Levenshtein, V. I. (1965) Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4), 845–848. English translation (1966) in *Soviet Physics Doklady*, 10(8), 707–710.
- Winkler, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667–671.
- Winkler, W. E. (1989). Frequency-based matching in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 778–783.
- Winkler, W. E. (2002). Methods for record linkage and Bayesian networks, *Proceeding of the Section on the Survey Research Methods*, American Statistical Association, 3743-3748.
- Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics* 6(5), 313–325.
- Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* 11(1), 95–103.
- Xu, H., Li, X., Shen, C., Hui, S., & Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *The Annals of Applied Statistics* 13(3), 1753–1790.
- Xu, H., Li, X., & Grannis, S. (2022). A simple two-step procedure using the Fellegi–Sunter model for frequency-based record linkage. *Journal of Applied Statistics* 49(11), 2789–2804.
- Yancey, W. (2000). Frequency dependent probability measures for record linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 752–757.
- Zhang, L., & Tuoto, T. (2020). Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184(2), 522–547.
- Zhu, V. J., Overhage, M. J., Egg, J., Downs, S. M., & Grannis, S. J. (2009). An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. *Journal of the American Medical Informatics Association* 16(5), 738–745.
- Zhu, Y., Matsuyama, Y., Ohashi, Y., & Setoguchi, S. (2015). When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. *Journal of Biomedical Informatics*, 56, 80–86.

Appendix A: Example of Scoring and Linkage Selections

To illustrate the scoring and linkage approach described in this paper, consider the artificial datasets in Table A.1. Dataset *A* includes four records, with a finder file containing the information reported for three variable categories: Name, DOB, and ZIP5. In the lower portion of the table, Dataset *B* includes five records with corresponding information on the variable categories for the individuals in this dataset.

Table A.2 presents the scores for each record in Dataset *A* against the individuals in Dataset *B*. Scoring uses the quality bands specified in Table 5.2.

For record A1 scored against B1, name falls in Band 3 of the Name category. Record B2 also falls in this Band when compared to A1, yielding a component score $2/5 = 0.4$ (i.e., the p_{1i} probability). For DOB, both B1 and B2 fall into the Band 1 classification, yielding $p_{2i} = 2/5 = 0.4$ as the component score. Finally, in the ZIP5 category, record B1 exactly matches A1's location information, producing a component score of $p_{3i} = 1/5 = 0.2$.

For record A1 scored against B2, names fall in Band 1 of the Name category. This yields a component score of $p_{1i} = 1/5 = 0.2$ probability. For DOB, record A1 falls into the Band 1 classification, along with B1 and B2. This yields $p_{2i} = 2/5 = 0.4$ as the component score. Finally, in the ZIP5 category, B2's location information does not match that of A1, implying classification in the lowest quality band yielding a component score of $p_{3i} = 1$.

The column labeled P in Table A.2 computes the product of the component scores, calculating the joint probability $P_i = p_{1i} \cdot p_{2i} \cdot p_{3i}$, measuring the uniqueness of the match. The score S equals the log odds $S_i = \log((1 - P_i)/P_i)$.

Comparing total scores for A1 indicates selecting B1 as the linkage for this individual. The key factor producing this highest score involves the uniqueness of the location information for individual A1.

Calculations of match scores for individuals A2, A3, and A4 mirror the computations summarized above. All component scores for these individuals fall into either the highest or lowest quality band for the relevant variable category. In the highest category, a value of $p_{ki} = 0.2$ indicates that only a single

Table A.1: Illustrative Example of Datasets

	Name		DOB	ZIP5
	First	Last		
Dataset A				
A1	James	Smith	1943	59921
A2	Robert	Martines	1951	77449
A3	Dion	Wooden	1947	77449
A4	Maya	Delgado	1930	11368
Dataset B				
B1	Jim	Smyth	1943	59921
B2	James	Smith	1943	77449
B3	Robert	Martines	1952	77449
B4	Dion	Wooden	1947	11368
B5	Mary	Delapaz	1930	77449

Table A.2: Scoring of Linkage Candidates

	Variable Category Scores			Total	
	Name	DOB	ZIP5	P	S
Score A1					
B1	0.4	0.4	0.2	0.032	1.48
B2	0.2	0.4	1	0.08	1.06
B3	1	1	1	1	
B4	1	1	1	1	
B5	1	1	1	1	
Score A2					
B1	1	1	1	1	
B2	1	1	0.6	0.6	-0.18
B3	0.2	1	0.6	0.12	0.87
B4	1	1	1	1	
B5	1	1	0.6	0.6	-0.18
Score A3					
B1	1	1	1	1	
B2	1	1	0.6	0.6	-0.18
B3	1	1	0.6	0.6	-0.18
B4	0.2	0.2	1	0.04	1.38
B5	1	1	0.6	0.6	-0.18
Score A4					
B1	1	1	1	1	
B2	1	1	1	1	
B3	1	1	1	1	
B4	1	1	0.2	0.2	0.60
B5	1	0.2	1	0.2	0.60

record in Dataset B has matching information. A value of $p_{ki} = 0.6$ indicates that three single records in Dataset B have matching information. Classification into the lowest quality band yields a component value of $p_{ki} = 1$.

The bolded scores in the S column identify the highest scores and the primary linkage selection. Selecting a threshold of 0.8 as the cutoff for establishing a linkage would imply no linkage for individual A4.

Expanding the number of individuals in Dataset B invariably reduces the match probabilities p_{ki} 's for the highest quality bands, increasing the scores of the most distinctive matches. For example, the ZIP5 data for individual A1 in Table A.2 indicates residence in Lake McDonald, Montana, which has a population of just one resident. Consequently, matching in this circumstance makes zip code information as discriminating as matching a unique and correct Admin ID, implying a component probability $p_{ki} = 1/(\text{number records in Dataset B})$.

Appendix B: Estimating the Probabilities of False Matches

Calculating the probability of false matches—termed Type 1 errors in the linkage literature—offers a reliability measure of a linkage approach. This appendix outlines a procedure for constructing such estimates for the PUBLIM linkage algorithm summarized in this paper.

Theoretical Framework for Computing Probability of False Matches

To translate the linkage method described in this paper into the notation of the F&S framework, designate A and B as two datasets with entities denoted as $a \in A, b \in B$. The set $A \times B = \{(a, b) \mid a \in A, b \in B\}$ includes all possible matches, which decomposes into the union of two disjoint sets: $M = \{(a, b) \mid a = b, a \in A, b \in B\}$ and $U = \{(a, b) \mid a \neq b, a \in A, b \in B\}$. The set M contains all the true matched pairs, and U includes all unmatched entities. The elements of a and b consist of an array of comparable attributes identifying the entities a and b .

A linkage operation infers membership of the sets M and U by specifying a comparison vector:

$$\gamma = \gamma(\alpha(a), \beta(b)) = \{\gamma_1(\alpha_1(a), \beta_1(b)), \dots, \gamma_K(\alpha_K(a), \beta_K(b))\},$$

where $\alpha(a)$ represents the matching variables for entity a in file A , $\beta(b)$ for entity b in file B , and the vector elements $\gamma_k(\alpha_k(a), \beta_k(b))$ measure the agreement pattern between attribute k recorded for entities a and b . When errors occur in reporting attribute k for entities, $\alpha_k(a)$ and/or $\beta_k(b)$ fail to depict the true information about the k -th attribute. A linkage algorithm observes $\gamma(\alpha(a), \beta(b))$ and assigns either $y = 1$ selecting (a, b) as a matched pair (i.e., $(a, b) \in M$), or $y = 0$ selecting (a, b) as an unmatched pair (i.e., $(a, b) \in U$).

Given a finder file containing the elements $\alpha(a)$ describing the attributes of each individual i in Dataset A , the linkage algorithm compares these attributes to the corresponding attributes $\beta(b)$ of the entities in Dataset B to determine which, if any, entity represents individual i . The algorithm produces the comparison vectors $\gamma = \gamma(p) = \{p_1, p_2, \dots, p_K\}$, where p_k equals the probability entity a in Dataset A has the same k attribute(s) as entity b in Dataset B . Given the calculated linkage score $S_i = S(\gamma_i)$ for individual i , $y_i = 1$ when $S_i \geq \theta$ and $y_i = 0$ otherwise, with θ a prescribed threshold.

Designate the distribution of the linkage score by $S \square f(S) = f(S(\gamma)) = f(S(p))$, where $f(\cdot)$ denotes a continuous univariate density. The underlying assumption presumes the p_k 's follow continuous distributions on the $[0, 1]$ space, with values measuring the comparisons of entities $a \in A$ with their most likely options $b \in B$.

The probability of a false match equals

$$(A.1) \quad P(y = 1 \mid U) = P(S \geq \theta \mid U) = \int_{\theta}^{\infty} f(S \mid (a, b) \in U) dS.$$

Estimation of the Probabilities of False Matches

Constructing estimates of the probabilities of invalid matches involves specifying sets knowing that no matches exist in these sets. Accordingly, partition Dataset B into two disjoint Datasets G and H , with entities $g \in G$ and $h \in H$. In this specification, Dataset G corresponds to A in the above theoretical framework, and Dataset H corresponds to B . With Dataset B containing no duplicate individuals, the unmatched set $U = G \times H = \{(g, h) \mid g \in G, h \in H\}$, and the matched set $M = \emptyset$ (null

set).

Applying the linkage algorithm described in this paper produces match scores S_i for each individual i in Dataset G and assigns a linkage $y_i = 1$ in Dataset H when $S_i \geq \theta$. An estimate of the probability of false matches equals

$$(A.2) \quad P(y = 1 | U) = \frac{\sum_{i=1}^{N_G} y_i}{N_G} = \Phi .$$

where N_G = size of the G population. The estimate Φ represents the share of the G population matched to an entity in the H population using threshold θ . By construction, all these matches are false.

One can determine a specified value of Φ (size of the Type 1 error) by varying the value of the threshold θ , which alters the number of $y_j = 1$ in Dataset H . Lowering θ increases this number, and raising it decreases it. The calculated match scores ordered from highest to lowest do not change.

The computation of Φ can accommodate the clerical-review adjustments in linkage assignments described in Section 7.4, which need not rely on strict use of the threshold θ to assign values of δ_j . These clerical reviews can lead to changes in $y_j = 1$ for some individuals in Dataset G . To improve accuracy, the linkage operation must implement the same judgments in assigning linkages to Datasets G and H as applied to Datasets A and B .

Computation of a reliable estimate Φ requires formulating the informational content of the matching variables $\alpha(g)$ to emulate the content of $\alpha(a)$ as close as feasible. This involves constructing dataset G to have a population composition analogous to Dataset A , and applying simulation methods to replicate the known distributional patterns of missing data and reporting errors in Dataset A in building Dataset G . An analysis would simulate the $\alpha_k(g)$ data to emulate the probabilistic combinations of missing data and reporting errors in the $\alpha_k(a)$ elements occurring in file A . In some instances, DH authorities may know of deficiencies in the personally identifiable information (PII) available in their Dataset A sources, advising on the types of errors and the likelihood of occurrence. In most other instances, little will be known beyond missing values, leading to Φ being an approximate estimate of the probability of false matches.

Appendix C: Example of Setting Thresholds for Linkage Scores to Bound Probabilities of False Match Rates for Finder Files with Complete and Missing Data

This appendix illustrates computing score thresholds for assigning linkages to achieve desired Type 1 error rates. This example covers the cases of a finder file with data available for all variables and cases with missing data for some variables for some individuals.

The analysis presents scoring results that match finder file information on observations in the Health and Retirement Study (HRS) and the entire Medicare enrollment population. Section 7.4 describes the contents of the HRS finder file and the match quality bands used in conducting this linkage exercise.

Overview of HRS-Medicare Matches

To provide context for the nature of the matches in this exercise, Table C.1 (pages 26-27) quantifies the extent of exact, partial, and mismatches (or missing) across the five variable categories available in the HRS finder file. For each row designating a match classification, the row lists the number of observations in the classification group and the mean and percentiles of the score distribution associated with this group. The HRS finder file has virtually complete data for about 84% of its observations, and missing data on Admin IDs for about 16% of its finder-file population.

Distributions of Match Scores

To maintain generality in the discussion below, interpret the finder file of HRS as a Dataset A containing data for all variables for segment $A1$ of its population and some missing data on AdminIDs for segment $A2$, with the union of these segments comprising the entire A population. PUBLIM scores the matching of Dataset A to Dataset B , with B constituting the universe of Medicare enrollment data. Let $f(S)$ denote the score distribution computed for population segment $A1$, and $g(S)$ denote the score distribution computed for segment $A2$.

Table C.2 presents the percentiles of the distributions $f(S)$ and $g(S)$. Row 1 reports the percentiles of the score distribution computed for those finder-file observations with complete data matched against the universe of Medicare data; this row reports the $f(S)$ distribution. Row 2 presents the percentiles of the score distribution computed for those finder-file observations with incomplete data matched against the universe of Medicare data; this row reports the $g(S)$ distribution.

Table C.2: Distributions of HRS Match Scores

Score Distribution	Score Percentiles										
	0.1 th	1 st	5 th	10 th	25 th	50 th	75 th	90 th	95 th	99 th	99.9 th
<i>Distributions with Complete and Missing Data on AdminIDs</i>											
1. Complete data on AdminIDs (f)	8.7	9.7	13.2	18.9	22.1	23.5	24.6	25.2	25.5	26.1	26.9
2. Missing data on AdminIDs (g)	4.1	7.2	8.2	8.6	9.7	14.0	15.8	16.7	17.0	17.7	18.4

Distributions of Type 1 Error Scores

Appendix B explains the computation exercise producing Type 1 error score distributions. The HRS dataset corresponds to Dataset A in Appendix B, and the Medicare population corresponds to Dataset B . Following Appendix B, split the Medicare population in the Datasets G and H . Designate

$f_{T1}(S)$ as the Type 1 error score distribution computed for the Dataset H segment of the Dataset B population assuming complete finder-file data, and designate $g_{T1}(S)$ as the Type 1 error score distribution calculated for the Dataset H segment of Dataset B assuming missing finder-file data on Admin IDs.

Table C.3 presents the percentiles for the Type 1 error score distributions associated with complete and missing data in the Dataset A finder file. Row 1 shows the Type 1 error score distribution associated with using complete finder-file data, estimating the $f_{T1}(S)$ percentiles. Row 2 shows the Type 1 error score distribution associated with utilizing a finder file with missing data on AdminIDs, estimating the $g_{T1}(S)$ percentiles.

Table C.3: Distributions of Type 1 Error Match Scores

Score Distribution	Score Percentiles										
	0.1 th	1 st	5 th	10 th	25 th	50 th	75 th	90 th	95 th	99 th	99.9 th
<i>Type 1 Error Distributions with Complete and Missing Data on AdminIDs</i>											
1. Complete data on AdminIDs (f_{T1})	6.8	7.4	7.9	8.2	8.6	9.2	9.9	10.9	11.6	12.9	18.3
2. Missing data on AdminIDs (g_{T1})	4.1	4.9	7.3	7.7	8.3	8.8	9.5	10.2	10.7	12.0	13.5
<i>Comparison of Mismatched and Missing AdminIDs</i>											
3. Missing data on AdminIDs*	4.1	7.2	8.2	8.6	9.7	14.0	15.8	16.7	17.0	17.7	18.4
4. Mismatched - Complete data on AdminIDs	8.1	8.5	8.8	9.1	9.7	11.3	14.8	16.3	16.9	17.6	18.7

* Same as row 2 in Table C.2.

Setting Score Thresholds to Achieve Selected Type 1 Error Rates

To construct a score threshold for those observations with complete finder-file information (segment $A1$), row 1 of Table C.3 reports the percentiles of its associated Type 1 error distribution f_{T1} . This row reveals that 5% of falsely matched observations with complete data have scores of 11.6 or greater. Accordingly, assigning linkages to those complete-data observations with scores exceeding this threshold yields approximately a 5% false match rate.

Setting this threshold in row 1 of Table C.2 falls at the 3.9th percentile of the f distribution, producing a linkage rate of 96.1% for segment $A1$ of the A population. Lowering the Type 1 error rate to 1% would decrease the linkage rate, requiring scores to exceed 12.9 to assign linkage.

To construct a score threshold for those observations with incomplete finder-file information (segment $A2$), row 2 of Table C.3 lists the percentiles of its associated Type 1 error score distribution g_{T1} . This row reveals that 5% of falsely matched observations with missing data have scores of 10.7 or greater. Lowering the Type 1 error rate to 1% requiring scores to exceed 12.0 to assign linkage.

Setting this threshold in row 2 of Table C.2 results in a linkage rate of 68.8% for segment $A2$ of the A population, corresponding to the threshold at the 31.2th percentile of the g distribution.

Setting these thresholds for the f and g distributions to the 95th percentiles of their associated Type 1 error score distributions yields an overall 5% false match rate for assignment of linkage in the

Dataset *A* population.

The results in Table C.3 demonstrate that implementing a 5% false match across finder file segments *A1* and *A2* induces a penalty on mismatched versus missing Admin ID cases. The close proximities of the f_{TI} and g_{TI} percentiles in rows 1 and 2 might, at first impression, suggest small penalties since only minor differences exist in Type 1 error thresholds for mismatched and missing cases. However, applying the different Type 1 error thresholds to the score distributions associated with missing and mismatched Admin IDs cases adversely impacts the linkage rates of mismatched compared to missing data cases. Specifically, row 3 in Table C.3 (which replicates row 2 in Table C.2) shows the percentiles for the missing data scores, and row 4 presents the percentiles for only those scores in the f distribution assigned to the mismatch Admin ID cases (i.e., in Quality Band 4 in Table 5.2). As noted above, setting a 5% Type 1 error rate for the missing-data distribution (row 3) yields a 68.8% linkage rate for cases with missing Admin IDs. For the mismatched Admin ID population, the 11.6 score threshold associated with a 5% Type 1 error rate for the complete-data distribution corresponds to the 52.0th percentile for the mismatched-data distribution (row 4), implying a 48.0% linkage rate for cases with mismatched Admin IDs. Consequently, mismatched-data cases are over 20 percentage points less likely to be linked than missing-data cases—interpretable as a sizable penalty assuming similar population compositions of these cases.

Comparing the summary statistics in rows 3 and 4 of Table C.3 also reveals important distributional properties relating to the finder file variables of missing and complete data observations. In sharp contrast to assumptions maintained in much of the linkage literature, the differences in these statistics indicate that missing finder-file data are not independent of other matching variables.⁷

⁷ For example, the scoring algorithms implemented in the popular Fastlink and Splink linkage approaches invoke the assumption of conditional independence of missing data and other matching variables.

Table C.1: Summary of HRS-Medicare Matches Qualities by Variable Categories

Variable Categories					Count*	Score Mean	Score Percentiles								
Admin ID	Residence	Age	Gender	Name			p1	p5	p10	p25	p50	p75	p90	p95	p99
Exact Match	Exact Zip5 Match	Exact Match	Exact Match	Exact Match	18,911	23.9	21.2	21.8	22.2	23.0	24.0	24.8	25.3	25.6	26.2
Missing	Exact Zip5 Match	Exact Match	Exact Match	Exact Match	2,566	15.6	12.8	13.5	13.9	14.7	15.7	16.6	17.0	17.3	17.9
Exact Match	Exact County Match	Exact Match	Exact Match	Exact Match	1,465	22.7	19.9	20.5	20.9	21.8	22.8	23.6	24.1	24.5	25.1
Exact Match	Exact Zip5 Match	Partial Match	Exact Match	Exact Match	1,334	21.8	18.7	19.6	20.0	20.8	21.8	22.7	23.5	23.8	24.4
Exact Match	Exact Zip5 Match	Exact Match	Exact Match	Mismatch	709	17.0	16.2	16.5	16.6	16.7	17.0	17.3	17.6	18.0	18.5
Exact Match	Exact Zip5 Match	Exact Match	Exact Match	Partial Match	640	23.6	20.8	21.4	21.9	22.7	23.7	24.6	25.1	25.5	26.1
Exact Match	Exact State Match	Exact Match	Exact Match	Exact Match	632	21.4	18.8	19.3	19.8	20.5	21.6	22.4	22.8	22.9	23.2
Missing	Exact Zip5 Match	Partial Match	Exact Match	Exact Match	392	13.3	9.6	10.8	11.3	12.1	13.4	14.4	15.3	15.6	16.1
Exact Match	Mismatch	Exact Match	Exact Match	Exact Match	343	19.8	17.1	17.9	18.3	19.0	19.8	20.8	21.2	21.3	21.4
Missing	Mismatch	Partial Match	Exact Match	Exact Match	229	9.1	7.9	8.2	8.3	8.7	9.0	9.5	10.0	10.4	10.9
Missing	Exact County Match	Exact Match	Exact Match	Exact Match	215	14.0	10.8	11.8	12.2	12.9	14.2	15.1	15.7	16.0	16.8
Partial Match	Exact County Match	Partial Match	Exact Match	Mismatch	199	10.9	8.9	9.4	9.7	10.5	11.0	11.4	12.1	12.4	12.8
Missing	Exact Zip5 Match	Exact Match	Exact Match	Mismatch	185	8.9	8.2	8.2	8.3	8.5	8.9	9.2	9.5	9.8	10.2
Missing	Exact State Match	Partial Match	Exact Match	Exact Match	165	9.8	8.3	8.6	8.8	9.2	9.6	10.2	11.2	11.8	12.8
Partial Match	Exact State Match	Partial Match	Exact Match	Mismatch	164	10.4	9.0	9.2	9.3	9.8	10.3	11.0	11.4	11.5	11.9
Partial Match	Exact Zip5 Match	Partial Match	Exact Match	Mismatch	159	11.7	10.2	10.6	10.7	11.0	11.6	12.3	12.6	13.0	13.7
Partial Match	Exact Zip5 Match	Exact Match	Exact Match	Exact Match	142	21.9	17.9	18.7	19.2	20.7	22.1	23.5	24.1	24.5	24.9
Missing	Exact Zip5 Match	Exact Match	Exact Match	Partial Match	141	15.4	13.1	13.6	13.8	14.6	15.4	16.3	16.9	17.0	17.6
Exact Match	Exact County Match	Partial Match	Exact Match	Exact Match	134	20.7	17.3	18.3	18.7	19.7	20.7	21.7	22.3	22.5	23.5
Missing	Exact County Match	Partial Match	Exact Match	Exact Match	114	10.6	8.5	8.9	9.3	9.6	10.3	11.2	12.9	13.7	14.2
Missing	Exact Zip5 Match	Partial Match	Exact Match	Mismatch	113	7.7	6.8	6.8	7.1	7.3	7.6	8.1	8.3	8.4	8.7
Partial Match	Exact County Match	Partial Match	Mismatch	Mismatch	113	11.0	8.9	9.9	10.1	10.4	10.9	11.5	12.0	12.5	13.1
Mismatch	Exact Zip5 Match	Exact Match	Exact Match	Exact Match	100	15.5	12.7	13.3	14.0	14.6	15.6	16.4	17.0	17.3	17.7
Partial Match	Exact Zip5 Match	Partial Match	Mismatch	Mismatch	93	11.4	10.0	10.1	10.3	10.8	11.6	11.9	12.4	12.5	13.1
Missing	Exact Zip5 Match	Missing	Exact Match	Exact Match	92	11.1	8.6	9.0	9.2	10.2	11.4	12.2	12.6	12.7	13.2
Missing	Mismatch	Partial Match	Exact Match	Partial Match	85	9.1	8.2	8.4	8.5	8.7	9.1	9.5	9.9	10.3	10.5
Missing	Exact State Match	Exact Match	Exact Match	Exact Match	84	12.9	10.7	10.9	11.2	11.8	12.8	13.9	14.5	14.6	14.9
Exact Match	Missing	Exact Match	Exact Match	Exact Match	82	19.9	17.2	17.8	18.3	19.0	20.1	21.1	21.2	21.3	21.4
Partial Match	Exact State Match	Partial Match	Mismatch	Mismatch	81	10.2	8.9	9.2	9.4	9.6	10.1	10.8	11.1	11.3	12.1
Missing	Mismatch	Mismatch	Exact Match	Exact Match	80	8.1	7.1	7.4	7.5	7.8	8.2	8.5	8.6	8.6	8.6
Exact Match	Exact County Match	Exact Match	Exact Match	Mismatch	79	15.8	14.6	14.7	15.1	15.3	15.7	16.3	16.8	17.0	17.0
Missing	Exact State Match	Partial Match	Exact Match	Partial Match	76	9.6	8.4	8.6	8.7	9.1	9.5	10.0	10.6	10.8	11.3
Missing	Exact Zip5 Match	Mismatch	Exact Match	Exact Match	76	10.4	8.9	9.0	9.2	9.6	10.3	11.0	11.9	12.2	12.6
Missing	Mismatch	Exact Match	Exact Match	Exact Match	75	11.3	9.1	9.6	9.8	10.5	11.3	12.4	12.9	12.9	13.0
Missing	Exact County Match	Mismatch	Exact Match	Exact Match	74	9.5	8.3	8.5	8.6	9.0	9.5	9.8	10.5	10.9	11.6
Missing	Exact State Match	Mismatch	Exact Match	Exact Match	74	9.1	8.0	8.2	8.3	8.6	9.1	9.4	9.8	10.1	10.5
Exact Match	Exact Zip5 Match	Partial Match	Exact Match	Partial Match	72	21.3	18.4	19.1	19.7	20.3	21.7	22.4	22.9	23.0	23.4
Exact Match	Exact State Match	Partial Match	Exact Match	Exact Match	65	19.1	16.2	16.8	17.0	18.2	19.2	20.1	20.7	20.9	21.2
Exact Match	Exact Zip5 Match	Mismatch	Exact Match	Exact Match	64	19.2	16.7	17.0	17.4	18.2	19.4	20.2	21.1	21.2	21.7
Exact Match	Exact Zip5 Match	Partial Match	Exact Match	Mismatch	62	15.0	13.5	13.6	13.9	14.2	15.1	15.6	16.4	16.7	16.8

Variable Categories					Count*	Score Mean	Score Percentiles								
Admin ID	Residence	Age	Gender	Name			p1	p5	p10	p25	p50	p75	p90	p95	p99
Partial Match	Mismatch	Partial Match	Exact Match	Mismatch	53	9.5	8.4	8.6	8.7	9.2	9.7	9.8	9.9	9.9	9.9
Exact Match	Exact County Match	Exact Match	Exact Match	Partial Match	52	22.3	19.5	19.9	20.3	21.2	22.4	23.5	23.9	24.1	24.2
Missing	Exact County Match	Exact Match	Exact Match	Mismatch	44	8.1	7.3	7.5	7.6	7.8	8.1	8.3	8.6	8.8	9.4
Missing	Exact State Match	Mismatch	Exact Match	Partial Match	41	8.9	7.9	8.1	8.4	8.7	8.9	9.2	9.5	9.9	10.1
Partial Match	Exact Zip5 Match	Mismatch	Exact Match	Mismatch	35	10.7	9.0	9.3	9.4	10.4	10.9	11.1	11.7	11.8	12.3
Exact Match	Mismatch	Partial Match	Exact Match	Exact Match	34	17.6	15.9	16.1	16.1	16.7	17.5	18.7	19.3	19.4	19.7
Missing	Mismatch	Mismatch	Exact Match	Partial Match	34	8.2	7.5	7.7	7.7	7.9	8.2	8.5	8.5	8.6	8.6
Missing	Exact Zip5 Match	Exact Match	Mismatch	Mismatch	32	8.6	8.0	8.0	8.1	8.2	8.6	8.7	9.2	9.2	9.9
Partial Match	Exact County Match	Mismatch	Exact Match	Mismatch	32	10.1	8.1	9.0	9.0	9.6	10.3	10.8	10.9	10.9	11.1
Missing	Exact Zip5 Match	Partial Match	Exact Match	Partial Match	31	12.6	10.8	10.9	11.0	11.5	12.2	13.5	14.7	15.4	15.7
Missing	Exact County Match	Mismatch	Exact Match	Partial Match	30	9.4	8.6	8.6	8.6	8.9	9.3	9.7	10.4	10.7	10.9
Mismatch	Mismatch	Partial Match	Exact Match	Exact Match	30	9.4	8.6	8.6	8.7	8.9	9.3	9.8	10.3	10.6	10.7
Missing	Exact State Match	Missing	Exact Match	Exact Match	27	8.4	7.0	7.5	7.6	7.9	8.3	8.9	9.2	9.9	10.3
Mismatch	Exact State Match	Partial Match	Exact Match	Exact Match	27	10.1	8.8	9.2	9.3	9.4	9.7	10.8	11.2	11.6	11.8
Exact Match	Exact Zip5 Match	Missing	Exact Match	Exact Match	24	19.8	16.4	17.9	18.2	19.2	20.2	20.6	21.0	21.0	21.4
Mismatch	Exact Zip5 Match	Partial Match	Exact Match	Exact Match	24	13.0	10.5	10.9	11.2	12.2	13.2	13.9	14.2	14.7	15.1
Exact Match	Exact State Match	Exact Match	Exact Match	Mismatch	23	14.5	13.7	14.1	14.1	14.1	14.5	14.8	14.9	15.2	15.3
Missing	Exact County Match	Partial Match	Exact Match	Partial Match	23	9.9	8.7	8.8	8.9	9.3	9.7	10.5	11.0	11.0	12.5
Missing	Exact Zip5 Match	Mismatch	Exact Match	Partial Match	23	10.2	8.5	8.7	8.9	9.5	10.1	11.1	11.9	12.1	12.2
Partial Match	Exact State Match	Exact Match	Exact Match	Mismatch	23	11.4	10.6	10.7	10.7	11.0	11.3	11.5	11.8	13.0	13.7
Exact Match	Exact State Match	Exact Match	Exact Match	Partial Match	21	21.0	18.1	18.8	19.1	20.0	21.1	22.4	22.5	22.8	23.0
Mismatch	Exact Zip5 Match	Mismatch	Exact Match	Exact Match	20	10.9	9.6	9.6	9.8	9.8	10.3	12.2	12.6	12.6	12.9
Exact Match	Mismatch	Exact Match	Exact Match	Partial Match	19	19.5	17.3	17.7	18.0	18.5	19.6	20.7	21.2	21.3	21.3
Partial Match	Mismatch	Exact Match	Exact Match	Mismatch	19	9.9	9.5	9.6	9.6	9.6	9.8	9.8	9.9	11.2	11.3
Missing	Mismatch	Missing	Exact Match	Exact Match	18	7.9	6.9	6.9	7.2	7.4	8.0	8.6	8.6	8.6	8.6
Missing	Exact County Match	Missing	Exact Match	Exact Match	17	9.5	8.1	8.1	8.2	8.5	9.7	10.3	10.6	10.9	11.0
Missing	Exact Zip5 Match	Missing	Exact Match	Mismatch	16	4.4	3.8	3.9	4.0	4.1	4.3	4.8	5.2	5.2	5.3
Partial Match	Exact State Match	Exact Match	Mismatch	Mismatch	16	11.1	10.5	10.6	10.6	10.8	11.1	11.2	11.9	11.9	12.6
Partial Match	Exact Zip5 Match	Partial Match	Exact Match	Exact Match	16	19.7	17.5	18.0	18.2	18.8	19.6	20.5	21.6	21.6	23.2
Partial Match	Exact Zip5 Match	Mismatch	Mismatch	Mismatch	16	10.3	9.0	9.1	9.2	10.0	10.6	10.8	11.2	11.2	11.5
Partial Match	Mismatch	Partial Match	Mismatch	Mismatch	16	9.5	8.4	8.5	9.3	9.4	9.4	9.6	10.1	10.1	10.8
Mismatch	Exact County Match	Partial Match	Exact Match	Exact Match	16	10.2	8.7	9.2	9.3	9.5	10.3	10.6	11.4	11.4	11.4
Exact Match	Mismatch	Exact Match	Exact Match	Mismatch	15	13.0	12.7	12.7	12.7	13.0	13.0	13.1	13.1	13.1	13.1
Missing	Exact County Match	Exact Match	Exact Match	Partial Match	14	14.2	12.5	13.5	13.5	13.8	14.1	14.7	15.6	15.6	15.7
Missing	Missing	Partial Match	Exact Match	Exact Match	13	8.9	7.8	8.1	8.1	8.2	8.9	9.3	9.9	9.9	10.1
Mismatch	Exact County Match	Mismatch	Exact Match	Exact Match	13	10.4	8.4	8.9	8.9	9.5	10.1	11.2	11.9	11.9	13.1
Missing	Exact Zip5 Match	Missing	Exact Match	Partial Match	12	11.2	9.2	9.7	9.7	10.1	11.3	12.4	12.5	12.5	12.6

* Cells suppressed with counts <12.

Appendix D: Characterization of PUBLIM Approach in Fellegi and Sunter Framework

This appendix translates this report's PUBLIM match scoring method in the Fellegi and Sunter (F&S) setting to provide a context for comparisons with alternative linkage methods found in the literature.⁸

Representation of the F&S Framework

Drawing on the notation and setting of Appendix B, consider the problem of matching N_A individuals in dataset A to N_B individuals in dataset B . Denote observations in A as $i \in A$, $i = 1, \dots, N_A$, and observations in B as $j \in B$, $j = 1, \dots, N_B$. Designate the comparison vector for individuals j and i as $\gamma_{ij} = \gamma_{ij}(\dots, \gamma_k(i, j), \dots)$.

A matching rule "scores" outcomes of γ_{ij} to determine the ranking of the linkage combinations (i, j) . Define $\varphi_i = \text{Max}_{j \in N_B} \varphi(\gamma_{ij} | j \in N_B)$ as the score assigned to individual i , with possible values $\varphi_i = r$, $r = 1, \dots, R$.⁹ The F&S framework interprets γ_{ij} as a random vector, which induces a distribution for φ_i specified here as $\text{Pr}(\varphi_i = r)$.

A linkage decision rule determines whether to assign a linkage for individual i to individual j based on the ranking of j 's match score. The F&S framework provides the optimal ranking of these scores. To specify this ranking, designate the discrete variable $y_i = 1$ when i is linked ($= 0$ when not linked), and define the probabilities

$$m_r = P(\varphi_i = r | y_i = 1) \quad \text{and} \quad u_r = P(\varphi_i = r | y_i = 0).$$

F&S prove that the optimal ranking orders the ratios $\frac{m_r}{u_r}$ from highest to lowest. Applying this

theorem, individuals j receiving scores $\varphi_i = r$ yielding the highest values of these ratios are assigned $y_i = 1$, while those with the lowest may or may not be linked depending on that threshold of the ratio determined to be sufficiently weak to justify assignment.

Application of F&S Optimality Rule

For the linkage approach described in this report, $\gamma_{ij} = \gamma(\dots, p_k(i, j), \dots) = \gamma_i(\dots, p_{ki}, \dots)$ where the p_{ki} 's equal the probabilities individual i in dataset A has the same characteristics as the matched individual j in dataset B for the different variable categories. The elements p_{ki} possess a continuous distribution over the unit interval. The scores equal $\varphi_i = \varphi(\gamma_i) = S_i$, having the distribution

$S_i \sim f(S_i)$ where f is a continuous density with a domain over positive numbers. Given the calculated linkage score S_i for individual i , $y_i = 1$ when $S_i \geq \theta$ and $y_i = 0$ otherwise, with θ a prescribed threshold.

The m_r and u_r probabilities take the form:

⁸ See the brief entitled "Comparison of LINKAGE and Fastlink Methods for Assigning Finder File Linkages" to review a comparison of the PUBLIM approach described here to the Fastlink approach developed in Enamorado, T. Fifield, B., and Imai, K., "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," American Political Science Review (2019).

⁹ The value r is a single value, but its value need not have a 1-to-1 correspondence to a unique j . Also, the values of r can be continuous on a set.

$$m_{S_i} = \Pr(S_i = r \mid y_i = 1) = \frac{f_1(S_i)}{P(S_i \geq \theta)} \quad \text{where } f_1(S_i) = f(S_i) \text{ when } S_i \geq \theta \text{ and } = 0 \text{ otherwise ,}$$

$$u_{S_i} = P(S_i = r \mid y_i = 0) = \frac{f_0(S_i)}{P(S_i < \theta)} \quad \text{where } f_0(S_i) = f(S_i) \text{ when } S_i < \theta \text{ and } = 0 \text{ otherwise .}$$

The ratios of these probabilities equal:

$$\frac{m_{S_i}}{u_{S_i}} = \frac{P(S_i < \theta)}{P(S_i \geq \theta)} \frac{f_1(S_i)}{f_0(S_i)} = \begin{cases} \infty & \text{when } S_i \geq \theta \\ 0 & \text{when } S_i < \theta. \end{cases}$$

This expression corresponds to the m_{ij}/u_{ij} estimates in equation (4) of Section 2.2. The posterior probability of linkage the (i, j) pair being a member of the inferred match set equals one if their score S_i exceeds the threshold θ , and equals zero if their score falls short of this threshold.

Accordingly, the linkage decision rule assigns linkages ($y_i = 1$) to those individuals i with scores S_i exceeding the threshold θ , and does not link ($y_i = 0$) individuals in dataset A with scores below this threshold. Appendix B describes a method for setting θ to bound the probability of assigning false matches.