Fine-Grained Record Integration and Linkage Tool

Pawel Jurczyk,[1,2,3]* James J. Lu,[3] Li Xiong,[3] Janet D. Cragan,[1] and Adolfo Correa[1]

[1]National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia
[2]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee
[3]Emory University, Mathematics and Computer Science, Atlanta, Georgia

Received 9 May 2008; Revised 8 August 2008; Accepted 26 August 2008

BACKGROUND: As part of the surveillance program to monitor the occurrence of birth defects in the metropolitan Atlanta area, we developed a record linkage software tool that provides latitude in the choice of linkage parameters, allows for efficient and accurate linkages, and enables objective assessments of the quality of the linked data. **METHODS:** We developed and implemented a Java-based fine-grained probabilistic record integration and linkage tool (FRIL) that incorporates a rich collection of record distance metrics, search methods, and analysis tools. Along its workflow, FRIL provides a rich set of user-tunable parameters augmented with graphic visualization tools to assist users in understanding the effects of parameter choices. We used this software tool to link data from vital records (n = 1.25 million) with birth defects surveillance records (n = 12,700) from the metropolitan Atlanta Congenital Defects Program (MACDP) for the birth years 1967–2006. **RESULTS:** Compared with the data linkage performed by conventional algorithms, the data linkage of birth certificates with birth defect records in MACDP using FRIL was more efficient. The linkage based on FRIL was also accurate, showing 99% precision and 95% recall. Based on positive user feedback, new features continue to be developed, and the tool is being adopted in several other data linkage projects in MACDP. **CONCLUSIONS:** A software tool that allows significant user interaction and control, such as FRIL, can provide accurate data linkages for birth defect surveillance programs and allows an objective assessment of the quality of linked data. *Birth Defects Research (Part A) 82:822–829, 2008.* © 2008 Wiley-Liss, Inc.

INTRODUCTION

Birth defects surveillance programs conduct a number of activities, such as monitoring the prevalence of birth defects, maintaining case registries for epidemiologic studies, evaluating morbidity and mortality associated with birth defects, and providing data for education and health policy decisions related to prevention (Correa et al., 2007). An essential component of many of these activities is the linkage of databases, a process of matching two or more databases to validate information collected in one database with that from another dataset or to acquire additional information on possible cases of birth defects. Traditional interactive tools for record linkage provide users with limited control, mostly allowing options to specify similarity measures between records and decision models (Campbell, 2005; LinkageWiz, 2008; Thoburn et al., 2007). Although the linkage tool may allow a number of options for the search algorithm, the combination of available choices typically does not provide enough parameter granularity to produce results that are easily discernible, rendering the objective assessment of the accuracy of the data linkage difficult.

The problem of record linkage can be described as follows. Given sets of records A and B, find a partition of $A \times B$ consisting of sets M (matched), U (unmatched), and P (possibly matched) that satisfy $M = \{(a, b) \mid a = b\}$ and $U = \{(a, b) \mid a \neq b\}$, where a and b represent the records in sets A and B, respectively. A widely adopted record linkage approach is the probabilistic approach (Fellegi and Sunter, 1969). First, a vector of similarity scores (or agreement values) is computed for each pair. Then the pair is classified as either a match, non-match, or possible match on the basis of an aggregate of the similarity scores. Among the methods used for such classification, there are rule-based methods that allow human experts to specify matching rules, unsupervised learning methods such as expectation-maximization (EM) that

Presented in part at the American Medical Informatics Association 2008 Annual Symposium, November 8–12, 2008, Washington, DC.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

*Correspondence to: Pawel Jurczyk, Mathematics & Computer Science, Mail Stop: 1131-002-1AC, Emory University, Atlanta, GA 30322. E-mail: pjurczy@emory.edu

Published online 4 November 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/bdra.20521

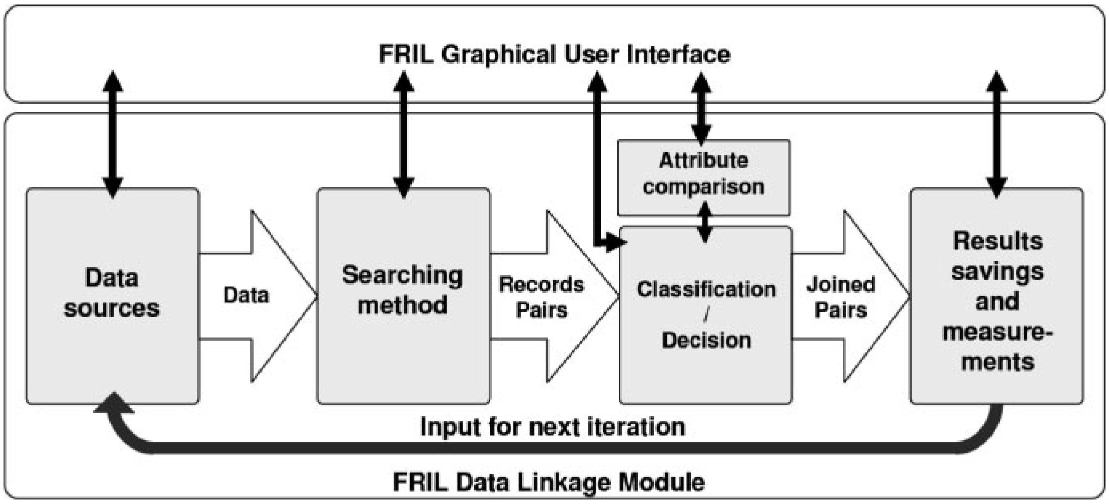


Figure 1. The FRIL architecture.

learn the weights or thresholds without relying on labeled data (i.e., pairs of records designated as matching and non-matching), and supervised learning methods that use labeled data to train a model, such as decision tree, naive Bayesian or Support Vector Machine (SVM) (Elmagarmid et al., 2007; Halevy et al., 2006; Winkler, 2006). For computing similarities between attributes, various distance functions are used and studied (Cohen et al., 2003; Navarro, 2001; Porter and Winkler, 1997; Ristad and Yianilos, 1996). Despite recent developments in data linkage methods, the literature on the application and evaluation of these methods for data linkages in birth defects surveillance projects (and in public health surveillance projects in general) is limited.

We developed an open source fine-grained record integration and linkage tool (FRIL) to link a birth defects surveillance database from the Metropolitan Atlanta Congenital Defects Program (MACDP) with a birth certificate database as a test case. In this article, we describe the spectrum of user-tunable parameters available in FRIL, evaluate the linkage of MACDP records with vital records, and discuss the potential use of FRIL in record linkages that involve birth defect surveillance data.

METHODS

The workflow of FRIL is shown in Figure 1. The user specifies the initial input files. Each run requires the user to specify the search method, the distance function in the attribute comparison module, and the decision model. Output consists of sets M, U, P, and various summary statistics. Sets U and P may be fed back into FRIL with a different set of parameters. We describe each module in turn.

Data Sources

An initial issue that the user must address when selecting attributes for comparison is to resolve possible discrepancies between schemas of *A* and *B*. Attributes are often labeled differently in the two data sets to be linked (e.g., ''Baby Name'' vs. ''B Name''), and in some cases two attributes of one source map onto a single attribute of the other (e.g., mapping between fields ''FirstName'' and ''LastName'' in the first and field ''Name'' in the second data file), or data needs to be parsed prior to matching. FRIL allows users to identify attributes from one source to the other. We call this user input *attribute selection and mapping*. The resulting set of attributes to be used in the linking is denoted Φ.

Searching Methods

Search methods refer to algorithms for determining which pairs of records to compare between the data sources A and B and the attributes on which comparisons are made. FRIL implements the *nested loop join*, *sorted neighborhood method* (SNM), and *blocking search method* (BSM). Nested loop join performs an all-to-all comparison between *A* and *B*, thus requiring $|A| \times |B|$ comparisons. It is useful for small input data files.

Technical details. The SNM first sorts records of A and B over the relevant attributes, and follows by comparing records within fixed windows ω_A and ω_B of records as these windows are advanced along the data files. Sorting moves records that have similar values (relative to the selected attributes) close together, presumably to within ω_A and ω_B of each other. This avoids the need to compare each record of one file against the entire data set of the second file. We call user inputs to ω_A and ω_B *window sizing*.

The BSM first groups records into *blocks* and then performs a nested loop join within each block. Such a grouping places records that have identical or similar values (with respect to a chosen blocking function) in the same block. This improves efficiency by avoiding the need to compare records that are classified into separate blocks. The subtle difference between SNM and BSM is that in SNM, users specify, a priori, values that fix the window sizes ω_A and ω_B, but in BSM, sizes of the blocks may vary. We call user inputs of the blocking function and its associated attributes the *blocking parameter*.

Table 1
FRIL Parameter Space

Description	Possible values
SNM window size selection	Two integer numbers greater than 0
Attribute selection and mapping	Any subset combination of the data source attributes, including possible merging and splitting of attributes
Attribute weighting	Real numbers between 0 and 1
Sort ordering (only for SNM)	All permutations of the attributes
Blocking parameter (only for BSM)	All permutations of attributes chosen in attribute selection and mapping and distance functions
Distance function selection	For each pair of attributes, distance function from a set of available functions
Attribute scoring	Two real numbers in the range [0, 1] with respect to the distance function; 0 indicates identical values
Record scoring	Two real numbers in the range [0, 1] indicating acceptance and rejection thresholds of records with respect to the attribute scoring

BSM, blocking search method; FRIL, fine-grained record integration and linkage tool; SNM, sorted neighborhood method.

When Φ contains more than a single attribute and searching is based on SNM, the choice of the dominant sorting attribute plays a critical role. In Atlanta, we found that when linking birth defects surveillance datasets with other databases, while the baby name attribute carries the greatest weight, using mother name as the dominant sorting attribute allows additional matches to be found. The reason is that data discrepancy occurs more frequently on baby names, and if baby name is used as the dominant sorting attribute, similar records may end up outside the comparison windows. We call the user input of the dominant sort attribute *sort ordering*.

Attribute Comparison

In this module, users choose distance functions that will be used when comparing fields between two data sources. Different pairs of fields may require different distance functions. For instance, when names are analyzed, a distance function that considers possible misspellings is appropriate. On the other hand, when fields representing body height are analyzed, numeric distance functions should be adopted.

Technical details. FRIL provides edit distance, Soundex, Q-gram, and equality distance functions (Cohen et al., 2003; Navarro, 2001; Porter and Winkler, 1997; Ristad and Yianilos, 1996). All the functions have the same type: $a \times a \rightarrow [0,1]$, where a is an attribute in Φ. The smaller the function value, the higher the probability that the two inputs are an exact match. FRIL allows users to choose different distance functions for each attribute in Φ. We refer to this input as a *distance function selection*.

For each distance function f, the user indicates the threshold for acceptance and rejection via a simple form of fuzzy logic. Specifically, if f_a is chosen as the distance function for attribute a, the user specifies the maximum $\max_{fa} \in [0,1]$ and the minimum $\min_{fa} \in [0,1]$ values for outright rejection and acceptance, respectively. For values between max_{fa} and min_{fa}, we use the membership function m_{fa}:

$$\begin{aligned} m_{fa}(s_1,s_2) &= 1 \text{ if } f_a(s_1,s_2) \leq \min_{f_a} \\ &= 0 \text{ if } f_a(s_1,s_2) > \max_{f_a} \\ &= \frac{\max_{f_a} - f_a(s_1,s_2)}{\max_{f_a} - \min_{f_a}} \text{ otherwise} \end{aligned}$$

We call this set of user inputs *attribute scoring*. If $min_{fa} = 0$ and $max_{fa} = 1$, then the above function is the same as a continuous similarity function used in typical probabilistic linkage methods.

Classification/Decision Module

This module is responsible for identifying pairs of records that are considered matches or possible matches. As different pairs of compared fields have different importance, users may assign weights that appropriately reflect the importance of each pair. For instance, between name and body height, matching on name is more important and hence should carry a higher weight than matching on body height.

Technical details. A weight is a real number $\alpha_a \in [0,1]$ assigned to each attribute a in Φ. We refer to this user input as *attribute weighting*. The final matching score for a pair of records r_1 and r_2 is the normalized weighted sum over all attributes:

$$\text{score}(r_1,r_2) = \frac{\sum_{a\in\Phi} \alpha_a m_{fa}(\pi_a(r_1), \pi_a(r_2))}{\sum_{a\in\Phi} \alpha_a}$$

π is the projection operator of the relational algebra. The user may specify two weights, min_t and max_t to indicate the overall scores for match rejection and acceptance. Linked records with scores above max_t are considered matching, below min_t are unmatching, and in between are probable matches. A goodness-of-fit score is reported based on the following membership function:

$$\begin{aligned} M(r_1,r_2) &= 1 \text{ if score}(r_1,r_2) \geq \max_t \\ &= 0 \text{ if score}(r_1,r_2) < \min_t \\ &= \frac{\text{score}(r_1,r_2) - \min_t}{\max_t - \min_t} \text{ otherwise} \end{aligned}$$

We refer to this user input as the *record scoring*. As an example, let $\Phi = \{a\}$, f_a be the edit distance, $min_{fa} = 0.5$ and $max_{fa} = 1$, $\alpha_a = 1$, $min_t = 0$ and $max_t = 1$. The following shows the scores and match results for three input record pairs (edit distance returns number of edits as a fraction of the length of the longer string).

r_1	r_2	#edits	f_a	M
''AARON''	''ARON''	1	0.2	1.0
''AARON''	''ADAM''	4	0.8	0.4
''AARON''	''HUGH''	5	1	0.0

Observe that the Boolean join (or exact match) condition is a special case of the above discussion and may be obtained by choosing equality as the distance function, choosing $\min_{fa} = \max_{fa} = 0$, and $\min_t = \max_t = 1$. Table 1 includes a summary of the full space of parameters in FRIL.

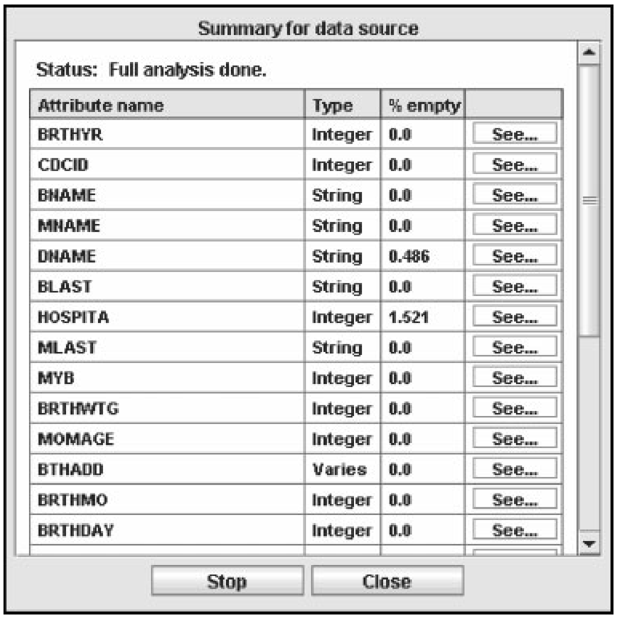
Support for User's Decisions

Choosing the best combination of values over all the parameters in FRIL can be a challenging task. Two solutions exist. The first is to provide, as much as possible, tools that help users understand the effects of parameter choices quickly. To that end, for several key user-specified parameters, FRIL contains real-time visual displays of the results over representative data samples. The second approach is to provide automatic parameter suggestions through machine-learning techniques and allow users to modify suggested values. In FRIL, this approach has been incorporated for choosing the attribute weighting. We present below some of the tools available in different FRIL modules that assist users in the attributes configuration process.

Data Sources

To assist users with attribute selection and mapping, FRIL provides a data source summary. The summary contains information on each attribute in the data sources, including name, type, and the percentage of null values found in the data sources. A screenshot of one such summary is shown in Figure 2. Sample values for each attribute can be displayed by clicking the ''See…'' button.

The second step in attribute selection and mapping is the reconciliation of schemas provided by two data sources. FRIL enables users to graphically map attributes between schemas. An example screenshot is shown in Figure 3. Attribute converters can be used to modify the attributes into a common format for linkage. Available converters include split converters that separate a single attribute into two attributes, merge converters that merge multiple attributes into one attribute, and trim converters that allow various string operations. In splitting, data are separated based on regular expressions.



Summary for data source

Status: Full analysis done.

Attribute name	Type	% empty	
BRTHYR	Integer	0.0	See...
CDCID	Integer	0.0	See...
BNAME	String	0.0	See...
MNAME	String	0.0	See...
DNAME	String	0.486	See...
BLAST	String	0.0	See...
HOSPITA	Integer	1.521	See...
MLAST	String	0.0	See...
MYB	Integer	0.0	See...
BRTHWTG	Integer	0.0	See...
MOMAGE	Integer	0.0	See...
BTHADD	Varies	0.0	See...
BRTHMO	Integer	0.0	See...
BRTHDAY	Integer	0.0	See...

Figure 2. Summary of data source.

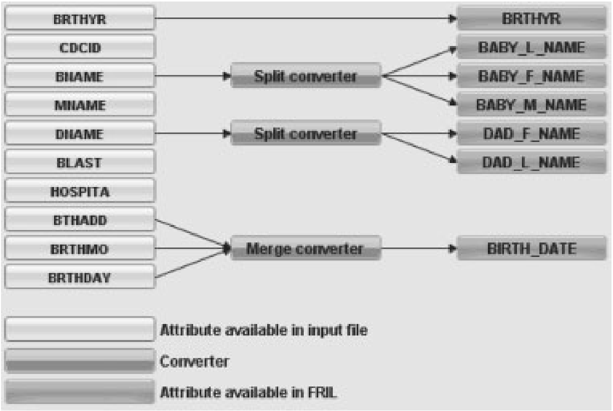


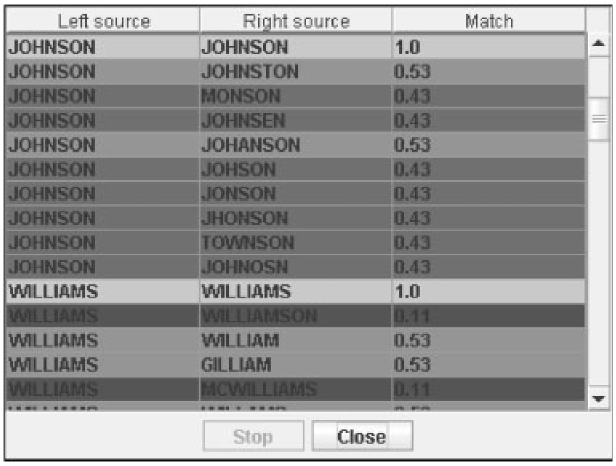
Figure 3. Data source configuration.

If schema reconciliation is configured correctly, users do not need to normalize the data within each data source; FRIL will perform the reconciliation automatically.

Attribute Comparison

After attribute selection and mapping is complete, users choose, for each attribute, a distance function and its associated attribute scoring values. To guide the user, dynamic distance function and attribute scoring analysis are provided. Whenever a user makes a change to these parameters, a brief summary of results is displayed. Note that the analysis is performed independently for each pair of matched attributes.

Figure 4 is a screenshot of the user interface for supporting distance function selection and attribute scoring selections for last names. Values from the data sources



Left source	Right source	Match
JOHNSON	JOHNSON	1.0
JOHNSON	JOHNSTON	0.53
JOHNSON	MONSON	0.43
JOHNSON	JOHNSEN	0.43
JOHNSON	JOHANSON	0.53
JOHNSON	JOHSON	0.43
JOHNSON	JONSON	0.43
JOHNSON	JHONSON	0.43
JOHNSON	TOWNSON	0.43
JOHNSON	JOHNOSN	0.43
WILLIAMS	WILLIAMS	1.0
WILLIAMS	WILLIAMSON	0.11
WILLIAMS	WILLIAM	0.53
WILLIAMS	GILLIAM	0.53
WILLIAMS	MCWILLIAMS	0.11

Figure 4. Support for distance parameters selection.

are shown with a numeric score, f_a, which indicates the level of match on the basis of the configuration of parameters. As the user adjusts the parameter values, the display is updated accordingly in real time. This gives users the ability to tune all the parameters until a desired result is obtained, before linking is performed.

Classification/Decision

To help users set the attribute weighting parameter, FRIL implements the EM method. The algorithm learns the weights without relying on labeled data. The user is responsible for specifying a method for sampling data sources and a search method for identifying tested pairs of records. The goal of sampling is to speed computation by reducing the number of records from the original data sources. The purpose of the search method in the EM algorithm is similar to that of the search methods for record linkage; it aims to identify pairs of records that will be used by the algorithm for computing weights. The options include all-to-all and blocking search methods. After the algorithm finishes learning, it provides a set of weights that can be used for attribute weighting.

Experiment

An objective of our experimental evaluation is to present a process for efficiently obtaining the best possible linkage between two input data sources. The MACDP program is an active population-based surveillance system for birth defects that was established in 1967 by the Centers for Disease Control and Prevention, Emory University, and the Georgia Mental Health Institute. The program collected information on more than 12,700 cases of birth defects among offspring of residents of the five central counties of Atlanta from 1997 through 2006. As part of the surveillance program, a birth certificate database of 1.25 million records of children born in Georgia for the same years was obtained from the Georgia Department of Human Resources. The goal of linking the two data sets is to match each record from the MACDP database with its corresponding record in the birth certificate database. However, the two sources contain numerous metadata (i.e., schema level) and object-data heterogeneities. Metadata heterogeneities (e.g., different number of digits used for encoding year of birth) are resolved in FRIL through the attribute selection and mapping parameter.

Object-data heterogeneity examples include incorrectly recorded information and missing data values. These are more difficult to handle and require the full range of FRIL features to resolve. The focus of the Results section is on this type of heterogeneity.

Metrics

The two data sets of interest in this project had been linked previously by using a deterministic, rule-based approach and a combination of running the program and manual inspections. We used the results from the linkage as our gold standard, *G*, against which the results from our data linkage approach are compared. For evaluation of our data linkage results, we used two standard metrics: precision and recall.

Table 2
Characteristics of Columns in Datasets

	Percentage of non-null values	
Column name	MACDP data	Birth certificate data
Birth date (baby)	*100*	*100*
Name (baby)	*100*	*100*
Birth date (mother)	100	100
Name (mother)	*100*	*100*
Birth date (father)	82	83
Name (father)	82	83
Hospital #	*100*	*99.5*
City	100	97.5
Zip code	*100*	*99.8*
Sex (child)	100	100

MACDP, Metropolitan Atlanta Congenital Defects Program.

$$\text{precision} = \frac{\#\text{ of true positives}}{\#\text{ of true positives} + \#\text{ of false positives}}$$

$$\text{recall} = \frac{\#\text{ of true positives}}{\#\text{ of true positives} + \#\text{ of false negatives}}$$

A true positive is a pair of correctly matched records, while a false positive is one that is incorrectly matched. For measuring improvements across experiments, precision and recall are better metrics than sensitivity and specificity. In particular, specificity can be unduly influenced by the large number of true negatives in large data sets.

RESULTS

Data Characteristics

Among the 187 MACDP data attributes, Table 2 provides statistics about the data in some the more intuitively important attributes in our data sources. In general, columns that have many *null* values are not good candidates to include in the join condition. Sources of *null* values vary and may indicate unknown, inapplicable, or unrecorded parameters. Therefore, comparing attributes with *nulls* provides less information than comparing attributes with *non-nulls*. Initial choices of attributes are highlighted in italic font in the table. Other attributes in the data sources were not used in our experiments.

Linkage Experiments (SNM)

With the attributes for the linking process fixed, we now describe experiments aimed at finding parameters of the join condition that produce the best linkage result. The six remaining parameters are distance function selection, attribute scoring, attribute weighting, sort ordering, record scoring, and window sizing.

Experiment 1. Our initial values for the first four parameters are shown in Table 3. The top-down ordering of the attributes in the table corresponds to the sort ordering parameter. Initial values for record scoring are $\min_t = \max_t = 0.61$. In all experiments, we used $\omega_A = \omega_B = 8$. For attributes with likely misspellings, the edit distance function is deployed with acceptance and rejection thresholds specified (as fraction of the length of the longer string) in the table. Results produced by the join con-

Table 3
Initial Join Condition

Column name	Metric	Weight
Name (baby)	Edit dist. (min_{fa} = 0.2, max_{fa} = 0.25)	0.4
Birth date (baby)	Equality	0.25
Name (mother)	Edit dist. (min_{fa} = 0.2, max_{fa} = 0.25)	0.2
Zip code	Equality	0.1
Hospital #	Equality	0.05

Table 4
Join Condition Used in Experiment 3

Column name	Metric	Weight
Name (baby)	Edit dist. (min_{fa} = 0.15, max_{fa} = 0.25)	0.3
Birth date (baby)	Equality	0.35
Name (mother)	Edit dist. (min_{fa} = 0.2, max_{fa} = 0.25)	0.2
Zip code	Equality	0.1
Hospital #	Equality	0.05

dition as specified were good: precision at 95% and recall at 86%.

Experiment 2. We reviewed the false positives generated above and observed that non-matching date of birth was an important cause. We refined the join condition by increasing the weight assigned to the date of birth attribute to 0.35 and reduced the weight of the baby name attribute to 0.3. This resulted in a 98% precision and a slightly decreased recall value of 85%.

Experiment 3. An examination of the remaining false positives showed a strong correlation to non-matching baby name, and in particular to the overly relaxed acceptance threshold for the edit distance function for baby name. We restricted the threshold by 25% (min_{fa} = 0.15,max_{fa} = 0.25), and it resulted in an improved precision value of 99% with no change in the recall. The join condition used in this experiment is presented in Table 4.

Experiment 4. To address the relatively low recall rate, we sifted through records that appeared in the gold standard *G* but that were not matched in Experiment 3. For most of these records, we observed that no attempted links were made by FRIL. The reason lies in the sort ordering we used. Using baby name as the dominant sorting attribute, two records with dissimilar values will occur far apart in the sorted files, beyond the window size for comparison. However, significant mismatches in baby name often occur as the result of data entry conventions; e.g., for babies that have not been given a first name, the letter B is used to denote ''Baby'' (e.g., ''Smith B''). In some cases, similar records appeared more than 1,200 records apart in the sorted file when sorted on baby name (Fig. 5a). While mother name is a semantically less significant attribute (i.e., carries less weight), it is a better dominant sorting attribute for many cases because of fewer variations in how its values are recorded. Figure 5b illustrates how the problem of Figure 5a is solved through a different sort ordering.

Rather than increasing the window sizes, which would hamper computational efficiency, we handled the problem with another feature of FRIL: the join summary. It allowed us to create an output of those MACDP records not joined in the initial run of the experiment (>1,500 records), and use them as input in a second run of the experiment under a different set of parameters. By changing the dominant sorting attribute to mother name, the second run linked nearly 900 of the unmatched records from the first run. Thus the combined effect of the two runs yielded 99% precision and 95% recall. With a small window size of 8, each run of FRIL takes approximately 20 min. The overall time for completing the four experiments took less than 2 days. The remaining unmatched records have non-matching names, dates of birth, etc. Those records were joined manually in *G* with the assistance of human expertise. We also found four linkages that did not appear in *G*. This suggests another utility of FRIL: it can be used as a verification tool for existing linkage results. Figure 6 shows a summary of the four experiments.

Linkage Experiments (BSM)

Experiment 5. We used the join condition presented in Table 4. Recall that for the blocking search method, a function for blocking with associated attributes needs to be specified. Our initial test used Soundex for the blocking function and baby name as the associated attribute. As a result, only records that had the same Soundex code for baby name ended up in the same block. These

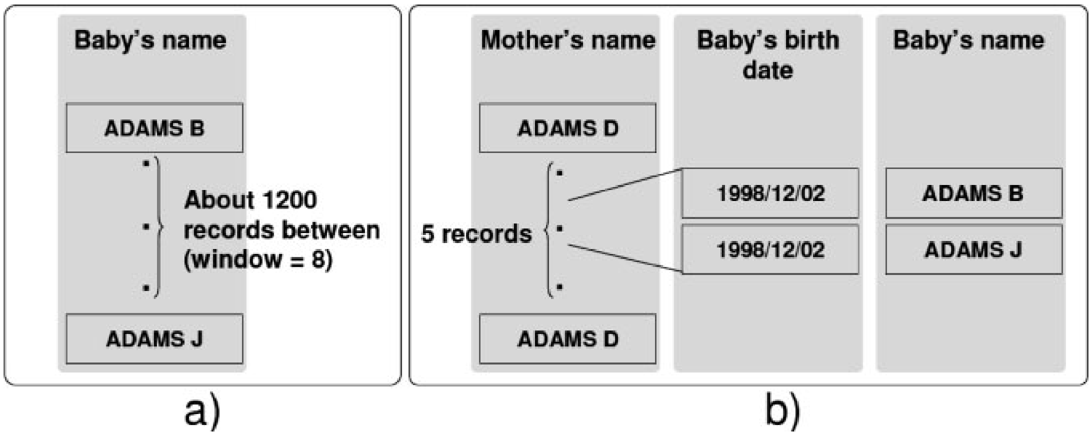


Figure 5. Impact of sort ordering on compared records in SNM search method.

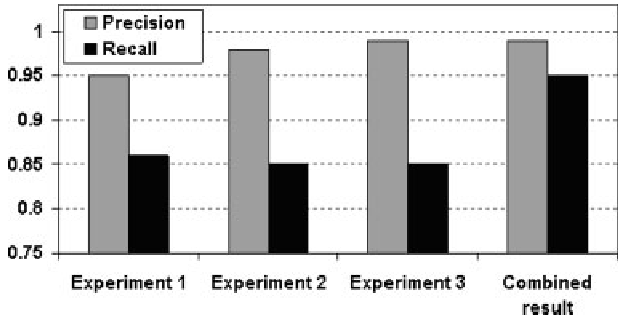


Figure 6. Summary of results for SNM search method. Combined result contains linkages from two runs.

Table 6
Precision and Recall Comparisons between Weights Obtained Manually and Weights Computed by the EM Method

	Original weights (%)		Weights computed by EM (%)	
	SNM search method[a]	Blocking search method	SNM search method	Blocking search method
Precision	99	98	99	99
Recall	95	88	86	86

[a]Denotes results from two iterations of the algorithm.

EM, expectation-maximization; SNM, sorted neighborhood method.

choices yielded 84% recall and 99% precision, which are very similar to the results of Experiment 3 using the SNM.

Experiment 6. Following the insights learned through Experiment 4, we changed the attribute used for blocking to mother name. This produced recall and precision of 88% and 98%, respectively. An inspection of the data showed the same effect as Experiment 4: similar records were more likely to be grouped together when blocking on mother's name.

Comparing the SNM and BSM, the latter is computationally more efficient and demands less configuration effort from the user. During our experiments, each run of the SNM took approximately 20 min, compared to an average of 12 min for the BSM. BSM requires specification of blocking function and its attribute, while the SNM requires user inputs on window sizes and a sort ordering. In general, while the BSM is easier to use, results of the SNM are more directly connected to its input parameters and hence allow for more systematic manipulation. For example, the user can intuitively understand that larger window sizes will produce more accurate results but will take longer to complete. A different selection of the blocking function in the BSM, on the other hand, may produce results that are drastically different. Therefore, it is much more difficult to predict the accuracy of linking results through refinements of the parameter values.

Evaluation of the EM Method for Attribute Weighting

To test the effectiveness of the EM method for choosing the initial weights of attributes used in the join condition, we sampled 5% of the data from the inputs to train the algorithm. As a search method used by the algorithm, we used blocking on baby name. Table 5 presents a comparison between weights manually obtained in the previous experiments and those calculated by using the EM algorithm.

Table 5
Weights Obtained by Using EM Method

Column name	Original weight	Computed weight
Name (baby)	0.3	0.21
Birth date (baby)	0.35	0.29
Name (mother)	0.2	0.22
Zip code	0.1	0.18
Hospital #	0.05	0.1

EM, expectation-maximization.

We compared the effectiveness of the computed weights against the results obtained previously in Table 6. For the SNM, we used the top-down ordering of the attributes in Table 5 as the sort ordering. For BSM, the baby name is used as a blocking attribute and Soundex as a blocking function.

Results show that weights obtained through the EM algorithm provide very high precision for both SNM and BSM. On the other hand, recall values provided by computed weights are less than manually selected weights. The experiment suggests that a useful process may be to apply the EM to obtain high-quality initial weights, and in subsequent iterations, manually adjust the weights to improve the recall value. Alternatively, initial linked set may be used as input data for alternative linkage processes based on supervised learning algorithms.

DISCUSSION

FRIL facilitated efficient and accurate record linkage over two large data sources with a great deal of flexibility in the choice of fine-grained parameters available for tuning the linkage tasks. FRIL allowed us to link MACDP and birth certificate data efficiently and accurately (99% precision and 95% recall). By exploiting all the features of FRIL, we presented an iterative process that enabled us to find good join conditions.

FRIL is an open source application. The tool, documentation, and other useful information can be found at http://www.mathcs.emory.edu/Research/Area/datainfo/FRIL/index.html.

FRIL uses a probabilistic linkage approach and provides users with options for tuning the accuracy and performance of data linkages. The combination of parameters provides researchers with objective measures for comparing results of linked data. Among the user-controlled parameters in FRIL are certain algorithmic decision points that are usually hidden in common linkage tools such as Link King (Campbell, 2005), Link Plus (Thoburn et al., 2007), or LinkageWiz (LinkageWiz, 2008). FRIL also differs from other tools that provide little or no options for schema reconciliation. This limitation makes the linkage process complicated and requires ad hoc data preprocessing. Additionally, powerful analysis, debug and decision support modules that are accessible at every level of linkage configuration in FRIL enable users to define linkage more intuitively, faster, and avoid common errors. A tabular comparison of FRIL to existing tools can be found on the FRIL website.

FRIL embodies the standard process of record linkage tools as described in, for example, TAILOR (Elfeky et al., 2002). From the data sources, the user chooses a search method, a set of distance functions for measuring record similarity, and a decision model for accepting or rejecting a match. Iterative refinement of linkage is possible: unmatched records from one run of FRIL are available as input to a follow-up run with a different set of parameters. Graphic tools for reconciling schema discrepancy and for analyzing, validating, and summarizing results have been incorporated. In addition, computerized learning tools are being developed to suggest automatic parameters.

The benefits of FRIL extend beyond the results of linking. By revealing key algorithmic decision points for user inputs, the tool forces researchers to consider computational issues that impact accuracy and performance of the linkage process. As a result, researchers are able to judge the quality of the linked data objectively and quantitatively. For already linked data, FRIL may also serve as a validation tool, since results obtained by running FRIL can be compared with linkage results obtained previously.

Work on extending the usability and efficiency of FRIL is ongoing. These include methods for supervised learning methods and alternative search methods based on clustering. Borrowing query optimization techniques from databases, window size and sort ordering may also be suggested through learning techniques.

In addition to conducting data linkages with vital records, MACDP conducts linkages with other datasets such as the National Death Index and the database of the Metropolitan Atlanta Developmental Disabilities Surveillance Program (Yazdy et al., 2008). In addition, MACDP has recently conducted linkage of MACDP records with an ambient air pollution database to examine the possible association of air pollution with heart defects. We anticipate conducting data linkages with new databases in the near future to improve case ascertainment for a number of conditions that are likely to be better reported in records on outpatient visits or other databases (e.g., some genetic disorders, fetal alcohol syndrome), to examine other possible outcomes, such as cancer, associated with birth defects and to examine the possible association of maternal lupus with birth defects. For such efforts, it will be important to have a record linkage tool that is efficient and accurate. We are optimistic that FRIL will facilitate many future data linkage projects based on birth defects surveillance data not only for MACDP but also for other birth defects surveillance programs.

ACKNOWLEDGMENTS

We thank Don Gambrell for providing the results of the initial linkage of the MACDP database with vital records, Matt Strickland for testing FRIL and providing helpful feedback, and MACDP staff for their suggestions.

REFERENCES

Campbell KM. 2005. Rule your data with The Link King (a SAS/AF application for record linkage and unduplication). SUGI 30 Proceedings. April 10–13, Philadelphia, Pennsylvania. Available at http://www2.sas.com/proceedings/sugi30/toc.html.

Cohen W, Ravikumar P, Fienberg S. 2003. A comparison of string metrics for matching names and records. Proceedings of the KDD-2003 Workshop on Data Mining Standards, Services and Platforms. p 13–18.

Correa A, Cragan JD, Kuick ME, et al. 2007. Metropolitan Atlanta Congenital Defects Program 40th anniversary edition surveillance report. Birth Defects Res A Clin Mol Teratol 79:65–186.

Elfeky MG, Elmagarmid AK, Verykios VS. 2002. TAILOR: a record linkage toolbox. Proceedings of the 18th International Conference on Data Engineering. p 17–28. IEEE Computer Society, Washington, DC, USA.

Elmagarmid AK, Ipeirotis PG, Verykios VS. 2007. Duplicate record detection: a survey. IEEE Trans Knowledge Data Eng 19:1–16.

Fellegi IP, Sunter AB. 1969. A theory for record linkage. J Am Stat Assoc 328:1183–1210.

Halevy A, Rajaraman A, Ordille J. 2006. Data integration: the teenage years. Proceedings of the 32nd International Conference on Very Large Data Bases. p 9–16. VLDB Endowment.

LinkageWiz. 2008. Record linkage software, Version 5.0. LinkageWiz Inc. Available: http://www.linkagewiz.com/.

Navarro G. 2001. A guided tour to approximate string matching. ACM Comput Surv 33:31–88.

Porter E, Winkler W. 1997. Approximate string comparison and its effect on an advanced record linkage system. U.S. Bureau of the Census, Research Report. Available at http://www.census.gov/srd/papers/pdf/rr97-2.pdf.

Ristad ES, Yianilos PN. 1996. Learning string edit distance. Technical Report CS-TR-532–96, Princeton University, Department of Computer Science. Available at: http://www.cs.princeton.edu/research/techreps/TR-532-96.

Thoburn KK, Gu D, Rawson T. 2007. Fundamentals of linking public health datasets. Link Plus: probabilistic record linkage software. Probabilistic Linkage Webinar 2, March 30. Available at: http://nahdo.org/cs/media/p/182.aspx.

Winkler W. 2006. Overview of record linkage and current research directions. U.S. Bureau of the Census, Technical Report. Available at: http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

Yazdy MM, Autry AR, Honein MA, et al. 2008. Use of special education services by children with orofacial clefts. Birth Defects Res A Clin Mol Teratol 82:147–154.