IDENTIFYING INDIVIDUALS IN A MULTIFACETED SERVICE ORGANIZATION

By

Cecelia Jane Madison

A CAPSTONE

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Biomedical Informatics

May 2012

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

_____

This is to certify that the Master's Capstone Project of

Cecelia Jane Madison

"*Identifying individuals in a multifaceted service organization*"

has been approved

_____

Judith R. Logan, MD, MS
Capstone Advisor

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I owe an unfathomable debt of gratitude to Judy Logan, who far surpassed her duties as an advisor in supporting my efforts during this project.

I am also immensely grateful to Jamie Meyers, not just for allowing me access to her systems and data, but also for promptly and thoroughly answering every question I asked.

Thank you most especially to Vincent and Susan Madison for the phenomenal intellectual, emotional, financial, and parenting support during the past five years.

ABSTRACT

Entity identification is the process of finding semantically related records in disparate databases. In the absence of a global unique identifier, determining which of the different records pertain to the same entity can be difficult. Disparate databases within an organization represent a significant barrier to the use of that organization's data.

Central City Concern is a multifaceted service organization which assists the homeless population of Portland, Oregon. Multiple different services are provided by and at different facilities. Over time, each facility independently developed individual mechanisms and procedures for collecting and storing client data. As a result, no cohesive method exists either to aggregate the organization's data or to identify multiple records for an individual across facilities.

An algorithm was developed that uses deterministic matching techniques to solve the problem of entity identification in the organization's different databases. This algorithm will be used to construct a master index that will link each of the facilities' internal identifiers for an individual client. The algorithm was used to classify a typical dataset against the organization's electronic health record data, and manual review demonstrated that the algorithm correctly categorized more than 99% of the records.

INTRODUCTION

Entity identification is the process of finding records pertaining to the same person in multiple distinct datasets. When the datasets do not share a global unique identifier, other fields must be used in the attempt to link the records. Several approaches can be used. Deterministic matching methods require that the variables selected fields match exactly between the two datasets. Probabilistic matching methods assess the degree of similarity or difference between the variables, and in some cases, the relationships between different variables can be used to aid the assessment.

Disparate databases can result from the slow growth of a multifaceted organization, as each different branch may develop its own database for internal use. One such organization is Central City Concern (CCC), a not-for-profit organization in Portland, Oregon that addresses homelessness with many different facilities. Available services include mental and physical health clinics that strive to especially attract people who avoid traditional health care settings. In addition, other branches of the organization offer drug and alcohol rehabilitation, as well as both housing and employment assistance. The goal of CCC is to assist its clients in "achieving self-sufficiency".[1]

The many branches of the evolving organization have, over the years, developed their own methods of collecting and storing data. Each facility, for instance, uses a different internal identifier for the clients it serves; furthermore, data collection and storage procedures differ among the facilities. As a result, a mechanism to easily track clients' use of multiple services is lacking. Datasets from different branches of the organization must be laboriously hand-linked — with uncertain accuracy — and assessments of the

efficacy of the programs rely on client testimony. The organization has grown large enough to need a data warehouse, for both production and research purposes.

Together with Oregon Health & Science University, CCC was awarded a grant from the National Institutes of Health to study the data that CCC has amassed. Many ideas could be explored with this data. Two of the most important are an analysis of the correlations between the different types of problems suffered by clients and a quantitative assessment of the efficacy of CCC's programs. Before any analysis can be done, however, the data in the different databases must be linked through entity identification.

This paper describes the development and evaluation of an algorithm that will be used to link records across CCC's databases. The algorithm's focus was to resolve the problem of entity identification within the different datasets. Using the deterministic matching method, the algorithm should be able to link records while minimizing not only error rates but also the amount of manual review required.

BACKGROUND

Entity Identification

The fundamental task of record linking is ascertaining that records from different branches of an organization refer to the same person. The procedure used to manage this task will impact the completeness and accuracy of the result. The process of determining whether two distinct data records are semantically related, or refer to the same real world object, is known in the literature as entity identification. In the absence of a shared unique identifier, a number of methods can be used to link records. These methods occur in two categories: deterministic and probabilistic matching.

The deterministic matching technique requires an exact match of variables in all fields selected for matching the records.[2,3,4] A variation of this technique is *n – 1* deterministic matching, which expects an exact match in all but one of the fields (for example, 3 out of 4 chosen fields must match).[2,5] These methods tend to have high specificity, but low sensitivity. This means that record pairs that are linked are very likely to refer to the same client (the entity in CCC's data), but that many records will be missed by being classified incorrectly as non-matches.

Other important metrics to consider in record linking are the rates of false positives and false negatives. A false positive occurs when two records are linked but actually pertain to distinct entities, or different people, and a false negative occurs when two records that do pertain to the same entity are not linked.[6] Deterministic linking's tendency toward high specificity correlates with a low rate of false positive occurrence, but the choice of matching variables heavily influences the occurrence and prevention of matching errors.

Drawbacks to the deterministic matching method include its vulnerability to data entry error and its generally low sensitivity. The $n - 1$ variant of this method can substantially raise the procedure's sensitivity, or, in other words, substantially lower the false negative rate. This improvement in sensitivity results from allowing one of the set of matching variables per record to not match. While doing so can compensate for data entry errors, success is highly dependent upon the overall uniqueness of the set of matching variables chosen. The set must be robust and unique enough to confer confidence in the link even in the event of one variable not matching. A generally acceptable set of matching variables includes last name, first name or initial, date of birth, and gender.[4,6]

In contrast to deterministic matching, probabilistic matching does not require an exact match of variables, but rather assigns weights to the matching variables based on their degree of similarity. A positive weight, or reward, is given for fields that match closely, and a negative weight, or penalty, is given to fields that do not match; a record is considered a link or non-link based on a comparison of the sum of the matching variable's weights to a threshold.[2,4,7] Bayesian probability theories are sometimes used, and one well known procedure for determining the weights to assign is the Fellegi-Sunter model, in which likelihood ratios are used to assess the similarity between fields.[7] One important criticism of these models is the assumption that the variables in a record are independent of each other. To address this concern, some researchers have developed models that incorporate the dependencies among variables into the weighting assignments.[7,8]

Probabilistic matching techniques tend to have better sensitivity than deterministic matching techniques, largely because they are not as affected by administrative errors in

the data that can lead to false negatives. It has been contended that this advantage of increased sensitivity may carry a concomitant burden of requiring more intensive human intervention in the record linking process, however, because the use of probabilistic matching methods may result in a greater proportion of records without certainty of a match.[9] When limiting the amount of records in need of manual review is a goal of the entity identification process, probabilistic matching techniques become less attractive, despite the potential to decrease instances of false negative occurrences.

A number of further concerns exist irrespective of the matching method used. So-called "homonyms" and "synonyms" tend to lead to errors in linking records.[10] The former describes the case of two distinct entities that happen to have very similar attributes, while the latter describes the case of a single entity for which attributes may have different values. An example of a homonym is the case of two different people who have, by coincidence, similar or identical names—such as Cecelia and Cecilia. Examples of a synonym include last name changes upon marriage, divorce, or adoption, and records for the same person being entered under both the given first name and a nickname—such as Cecelia and Celia.

METHODS

Choices for Central City Concern

CCC intends to eventually build a data warehouse. As a first step, a master index will be created that will link individual clients' internal identifiers from the different branches of the organization. In this way, client information can be aggregated fairly easily as needed, despite being stored in the organization's disparate databases. Numerous challenges hinder the process of constructing this index. Beyond the technical difficulties of working with the distinct and different databases, a significant challenge is presented by characteristics of CCC's client population. Not only does this population lack many instances of the identifying demographic data used to link records, but the assumption must also be made that some data will not be accurately reported.

In light of the research detailed above, deterministic matching techniques were chosen for performing the entity identification necessary in constructing the client index for CCC's data. The comparatively simpler methodology of the deterministic matching technique was preferred, and it was decided to attempt to improve sensitivity with data processing.

The primary goals for the matching algorithm's performance were to limit both the false positive rate and the number of records requiring manual review. Minimizing the false positive rate is important because the occurrence of incorrect matches would detract from the quality of any research using this data. It is possible that imposing a limit on the false positive rate will cause some records not to be matched that could be, but, as with other projects facing this dilemma, it was decided that greater damage could be done to future research by wrong matches than by missed matches. Limiting the amount of manual

review needed is important to alleviate the burden on CCC's limited staff resources of maintaining the client index and the planned data warehouse.

A maximum of one percent of matches was determined to be acceptable for the false positive rate, and a maximum of ten percent of records was chosen as the limit for potential matches needing manual review. These limits were chosen because they were believed to be effective for both research and production needs. A false positive rate of less than one percent can be supposed to keep the numbers of erroneous matches negligibly low without introducing the concern of significant numbers of missed matches. Manual review of ten percent of records was felt to be an acceptable cost for maintaining an otherwise automated entity identification process.

When the algorithm is implemented as an application that will be run against datasets, it must separate each record in the incoming dataset into one of three categories: record-pairs considered certain matches, record-pairs considered possible matches and requiring manual review, and single records considered certain non-matches. The results of the algorithm will be used to create and update the index in several ways; when matching records are found, the index will capture the internal identifiers used by the different branches of the organization. In addition, any new records from the incoming dataset that are not already found in the index will be appended to the index in order to be available for possible future matching.

For this project, two datasets were used (Tables 1 and 2). One was derived from the electronic health record used by CCC's health clinics and contained 106,932 records. It is believed to have the most complete and accurate data of any of their

Table 1. Fields and formatting notes for the CCC EHR demographics table

Field	Notes
Patient Profile ID	Numeric data
Last Name	Variable length character data Inconsistent use of hyphenation and spaces Occasional inclusion of suffixes
First Name	Variable length character data Inconsistent use of spaces
Middle Name	Variable length character data
Date of Birth	Date format Occasional inclusion of default time entry
Social Security Number	Character data No dashes or spaces

Table 2. Fields and formatting notes for the CCC housing assistance agency dataset

Field	Notes
Client UID	Numeric data
Last Name	Variable length character data Inconsistent use of hyphenation and spaces Occasional inclusion of suffixes
First Name	Variable length character data Inconsistent use of spaces
Date of Birth	Date format
Social Security Number	Character data with dashes: xxx-xx-xxxx

systems. The other dataset used comprised a typical periodic report from CCC's housing assistance agency and contained 1019 records.

Algorithm Development

Previous matching procedure. Currently, CCC's Director of Information Technology maintains responsibility for linking records from the organization's different branches whenever it is necessary to do so. She generally makes two attempts at entity identification with each dataset, and almost exclusively, she uses the electronic health record derived table as the dataset against which she performs the record linkage.

Through experience, she has come to believe that the combination of last name and social security number are the most successful matching variables, and she uses them to do the majority of the record linking. After using this first method, she uses the combination of last name and date of birth as matching variables in a second attempt to link more records, in particular those which have missing or incorrect data in the field of social security number.

Due to the independent development of the different databases, no consistent formatting of the data elements exists. Before any attempt at entity identification can be made, the formatting discrepancies must be resolved. The minimum mandatory reconciliation of data element formats that must be done in this instance is of client social security numbers. In the client index, the social security number will be stored as a nine character text string, and incoming datasets were revised to use this same format.

Certain matches. The first goal of the algorithm is to link all the records that are "certain" matches. Available matching fields are first name, last name, social security number, and date of birth. In this project, five sets of records were considered matched and were successively removed from the dataset before making further attempts at linkage.

- The first set was comprised of those records that matched on all four fields.
- The next three sets were those that matched on social security number and any two of the other three fields.

- The last set was formed from those that match exactly on first name, last name, and date of birth, but which have only a single digit discrepancy in social security number.

If any field was null in either dataset, but the other three variables matched, the records were considered a match, even in the case where one of the social security number fields was null.

Possible matches. After the “certain” matches were identified and removed from the dataset, the algorithm must distinguish between those records which do not have matches and those which have "possible" matches and require manual review. Records that matched on social security number alone – with the field being not null in both datasets – and records that matched on the combination of last name and date of birth were flagged for manual review. All remaining records were considered non-matching.

Certain non-matches. The non-matches are defined in negative terms. These are records from the dataset which do not match any records in the index or second dataset on social security number, the combination of last name and date of birth, or any triple of first name, last name, date of birth, and social security number.

This algorithm was developed working with Microsoft Excel and Access, the software that CCC currently uses for storing and managing their data. The data processing to format the social security number field was done using Excel. The record linking was performed in Access, using the SQL view. The necessary steps to implement the algorithm can be viewed in Appendix 1.

Algorithm Implementation

The performance of the developed algorithm was tested with a real and typical dataset of 1019 records from CCC's housing assistance agency matching against the demographic data table derived from their EHR, as described above. Because individual clients do not necessarily use more than one of CCC's services, there was no expectation that either of these datasets contained records from all CCC clients.

The social security number in the housing agency dataset was reformatted to be stored as character data with no spaces or dashes. Record linkage was begun by matching the datasets on all four of the chosen matching variables: social security number, date of birth, last name, and first name. The record pairs resulting from this query were categorized as "certain" matches and removed from the housing agency dataset. Next, the four iterations of the planned n – 1 deterministic matching were performed between the EHR demographic dataset and the remainder of the housing agency dataset. At the end of each iteration, the record pairs were categorized as "certain" matches and removed from the housing agency dataset. Examples of the SQL queries are shown in Appendix 2.

The next proscribed step of the algorithm is to determine which record pairs should be categorized as "possible" matches. Two queries were used to identify these candidate record pairs; one matched the datasets on the combination of last name and date of birth, and the other matched on social security number only. After removing these records from the housing dataset, the remainder of records therein were categorized as non-matches.

Every "certain" or "possible" matched record pair was reviewed manually to determine whether or not the algorithm performed correctly. The results of the matching queries were transferred to a spreadsheet, where the columns were rearranged to make corresponding fields adjacent; for example, the last name fields were moved to adjacent columns. Then, every character of every field was meticulously checked, both to verify that the query results were what was expected and also to investigate the discrepancies that appeared in non-matching fields.

RESULTS

The algorithm categorized 681 record pairs as "certain" matches. This represents 66.8% of the records in the housing dataset. Of the 681 record pairs, 515 matched on all four variables, and 166 matched on three of the four variables. Table 3 shows the numbers of records determined by the algorithm to be "certain" matches. The use of all four available fields confers a high degree of certainty that the matches are correct, and the addition of the $n - 1$ deterministic matching technique allowed more than another third of records to be matched.

The next steps in the algorithm classified 22 record pairs (2.16% of the housing dataset) as "possible" matches (Table 4) leaving 316 records in the housing dataset as non-matches.

Manual Review

All record pairs were manually reviewed, both to verify the query results as well as to evaluate the non-matching fields. Possible errors include record pairs that were matched

Table 3. Results of application of algorithm for certain matches

Fields used in *n* and *n* – 1 deterministic matching	Number of records matched	% of housing records matched
First name, last name, date of birth, SSN	515	50.5
First name, last name, SSN	20	1.96
First name, date of birth, SSN	46	4.51
Last name, date of birth, SSN	51	5.00
First name, last name, date of birth	49	4.81

Table 4. Results of application of algorithm for possible matches

	Number of records matched	% housing records matched
Possible matches	22	2.16
Pairs matched on combination of last name and date of birth	8	0.785
Pairs matched on social security number only	14	1.37

but should not have (false positives). Record pairs that did not match but should have (false negatives) were not sought out. Results are shown in Table 5.

This process shows that the algorithm worked correctly for 99.2% of the housing records. This assumes a ver low false negative rate (see the discussion below on this topic). In addition, the false positive rate was within acceptable limits at 0.79%. Of the 22 "possible" matches, 17 were true matches and 3 were non-matches and 2 were indeterminate. A non-exhaustive manual review was performed for the 316 non-matching housing records, and no matches were found. This suggests an optimistic yet reasonable very low false negative rate.

Table 5. Results of manual review of matching process. Non-matching records are from the housing dataset only

	Manual match process				
Algorithmic match process		Match	Possible match	Non-match	Total
	Match	673	8	0	681
	Possible match	17	2	3	22
	Non-match	0	0	316	316

Of the record pairs that matched, but with less than four of the fields being identical, the majority of the fields that did not match had human-understandable errors such as typographical errors, transposed digits, or formal versus "pet" names. Eight record pairs, however, were identified that had nontrivial differences between the non-matching fields. In other words, the record pairs had been matched based on three fields, but the values in the fourth field were substantially different. For example, these fields had differences in first name such as Cecelia in one dataset and Jane in the other. Four of these differences occurred in first names. One possible explanation for this discrepancy is that it is not uncommon for people to be known by their middle names but use their first names on some documents. Two of the nontrivial differences occurred in last names, and two occurred in dates of birth. Discrepancies in last names could be accounted for by marriage, divorce, or use of an alias. If, indeed, all eight of these records were erroneously matched, the false positive rate would be 0.79%.

It would be reasonable to link these 8 record pairs in spite of the discrepancies, however, based on two rationales. The first is that this maximum false positive rate was below the set upper limit of 1%, and therefore acceptable, and the other is the sense that these records were not likely to have been erroneously linked – in other words, for the reasons given above, the discrepancies were not enough to indicate that the records actually belonged to different clients.

Twenty two record pairs were classified as "possible" matches by the algorithm. Upon review, it was determined that 17 of these 22 record pairs were matches and 3 were definitely not matches. The other two records remained uncertain. The algorithm defines one set of "possible" matches as those record pairs that match on the combination of last

name and date of birth, but not on first name or social security number. From the sample dataset of 1019 records, eight record pairs fell into this category of "possible" matches. Seven of them were true matches, and one was a non-match. The other category of "possible" matches is those record pairs that only match on social security number. Fourteen such record pairs were found in the sample data set; ten of them were determined to be true matches, and two of them were determined to be non-matches. Two of the record pairs could not be definitively categorized as either matches or non-matches.

The remaining 316 records from the housing dataset were designated non-matching by the algorithm. Attempts to discover false negatives among them on non-exhaustive manual review did not reveal any definite false negative records. It is possible, however, that some duplication of clients still exists, but based on the limited fields provided, none could be detected. Missing values, errors, or discrepancies in more than two of the matching variables in either or both datasets would cause some records to falsely be categorized as non-matches.

This test of the algorithm demonstrated that it correctly categorized more than 99% of the records. The false positive rate of 0.79% falls below the imposed limit of 1%, and the 22 "possible" matches requiring manual review represented 2.16% of the sample dataset, which is well below the suggested 10% limit.

DISCUSSION

Central City Concern, an organization serving the homeless population of Portland, Oregon, evolved to comprise many different facilities in the quest to offer a comprehensive spectrum of assistance programs to their clients. As these many branches of the organization developed, however, no comprehensive plan guided the formation of their data storage procedures. The incompatibility of the resulting disparate databases presents a significant barrier to the productive use of CCC's vast accumulation of data for government reporting purposes, evaluation of services, and research. Supported by a grant from the National Institutes of Health, CCC is working with Oregon Health & Science University to resolve the lack of cohesiveness in the disparate databases.

This paper describes the development and evaluation of an algorithm that, when implemented in a software application, will help build and maintain a master client index. If implemented, it will identify records pertaining to an individual entity – a single client – that are stored in different databases. The algorithm will first be used to build an index consisting of all CCC clients; later, it will be used to update the index with new data. The algorithm is designed to sort the records of two distinct datasets into three categories: certain matches, possible matches, and certain non-matches. Goals defined at the outset were to achieve a maximum false positive rate of 1% and to limit the manual review needed of possible matches to 10% or less of the records.

When the algorithm was used to perform entity identification between two real CCC datasets, the results showed a very low false positive rate and no detectable false negative rate. It is unrealistic to believe that none of the un-matched records actually should have

been linked, but none were identified during evaluation. The false positive rate was calculated as if eight record pairs identified during manual review of the "certain" matches did not represent the same entity, which is the worst case. It is possible that one or more of these record pairs are matches. Both the false positive rate (0.79%) and the low proportion of record pairs requiring manual review (2.16%) were encouraging indicators that this algorithm and the resulting master client index will meet the organization's needs.

Although the evaluation results are encouraging, the impact of this project's success does have significant limitations, and foremost among them is that the algorithm was only tested with two datasets. Error rates could be far different when entity identification is attempted on other datasets, and a high priority for future work should be to evaluate the algorithm's performance on other datasets. The same matching variables will be available in the data from all of CCC's branches, but the correctness of the data could vary.

In addition, the records which were deemed non-matches may in fact contain some duplicate records. To be duplicates, however, the client's name, date of birth and social security number would all three have to have been different in the records.

Another important limitation is that no attempt was made to automate the algorithm; only its design was considered. In order to be most useful for purposes of both production and research, the algorithm will need to be implemented in a software application that takes a dataset as user input. The application then needs to be able to independently link that dataset to the index – or to the EHR demographic dataset, as a substitute for the index –

and categorize every record in the new dataset as a match, possible match, or non-match. Ideally, the application will be able to update the index automatically with the match and non-match records, leaving only the task of disposition of the possible matches to be executed by a human. While such an application is theoretically feasible, it must be written and tested in order to determine its usability for CCC.

The algorithm performed well during this project, but manual review of the matched records revealed several possible improvements. Most prominent among these is data processing of client last names. The simplest work that could be done with these would be to identify names with a space or hyphen and remove the space or hyphen during the linkage procedure. Many instances were observed of records not matching on last name because one dataset used a hyphen while the other used a space. If records such as these also had missing fields or typographical errors in other fields, they would remain erroneously non-matched. Inconsistent use of suffixes, such as "JR," also contributed to records not matching despite pertaining to the same client. An additional step of data processing to identify and remove suffixes should improve linkage results.

The final significant limitation to this project lies in the false negative rate. Although this rate was reported to be presumptively zero, it cannot be definitively said to be so, for a number of reasons. First of all, an exhaustive attempt was not made to identify possible matches between the EHR demographic dataset and the set of non-matches derived from the housing agency dataset. A link between these datasets on last name only was reviewed without identifying any matches, but the possibility does remain that more records should have been matched but were not. It may never be possible to definitively

define a false negative rate in these datasets. Due to the population's use of aliases, demographic data pertaining to the same person could be simply un-matchable.

Future work to build upon this project will focus on the construction of a data warehouse. It may be possible to do integrate these data sources logically, with the master client index as the keystone. The index would serve to record the existence and location of records in CCC's different databases. Although access to the original data would not be provided, the index could be used to identify a set of records to be amassed from their original sources. An architecture providing direct access to the original records would require interfaces between the different systems. Alternatively, a data warehouse could also use a physical architecture, in which all of the data from all of the sources would be integrated within a single system. In either case, many questions concerning responsibilities for maintaining up to date data would need to be resolved.

As future work is being planned, further evaluation of this and other entity identification algorithms may prove worthwhile. In spite of this author's considered opinion that deterministic matching techniques are sufficiently robust for record linkage in this instance, exploration with other datasets may show otherwise.

CONCLUSION

This paper describes the development and evaluation of an algorithm that performs entity identification for a multifaceted organization with many disparate databases. Deterministic matching techniques were used to link datasets with a false negative rate near 0 and a maximum false positive rate of 0.79%. This project demonstrated that a relatively simple algorithm can be used successfully to detect records originating in heterogeneous databases but pertaining to the same client.

REFERENCES

1. Central City Concern website: www.ccconcern.org; last accessed 16 May 2012.
2. Méray N, Reitsma JB, Ravelli ACJ, Bonsel GJ. 2007. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidem. 60:883-891.
3. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, Mamlin B, INPC Management Committee. 2005. The Indiana Network for Patient Care: a working local health information infrastructure. Health Affairs. 24(5):1214-1220.
4. Li B, Quan H, Fong A, Lu M. 2006. Assessing record linkage between health care and Vital Statistics databases using deterministic methods. BMC Health Services Research 6:48.
5. Ganesh M, Srivastava J, Richardson T. 1996. Mining entity-identification rules for database integration. KDD-96 Proceedings.
6. HIMSS Patient Identity Integrity Workgroup. 2009. Patient Identity Integrity. HIMSS.
7. Alemi F, Loaiza F, Vang J. 2005. Probabilistic master lists: integration of patient records from different databases when unique patient identifier is missing. Health Care Manage Sci. 10:95-104.
8. Lim E-P, Srivastava J, Prabhakar S, Richardson J. 1993. Entity identification in database integration. IEEE.

9. Grannis SJ, Overhage JM, McDonald CJ. 2002. Analysis of identifier performance using a deterministic linkage algorithm. AMIA 2002 Annual Symposium Proceedings: 305-309.
10. Cummins D. 2007. Patient identification: hybrids and doppelgängers. Ann Clin Biochem. 44:106-110.

APPENDIX 1

Entity Identification Algorithm

1. Establish matching variables.
    a. Social security number
    b. Date of birth
    c. Last name
    d. First name
2. Format matching variables.
    a. Social security number: 9 characters, with no dashes
    b. Date of birth: MM/DD/YYYY
3. Link datasets on all 4 variables.
    a. Remove matched record pairs from datasets.
    b. Categorize record pairs as certain matches.
4. Link unmatched records from datasets on 3 variables.
    a. Four iterations, as below:
        i. Remove matched record pairs from datasets after each iteration.
        ii. Categorize record pairs as certain matches.
    b. Social security number, date of birth, last name.
    c. Social security number, last name, first name.
    d. Social security number, date of birth, first name.
    e. Date of birth, last name, first name.
        i. Identify record pairs with multi-digit discrepancies in social security number.

ii. Un-link identified record pairs; do not remove from datasets.

5. Link unmatched records from datasets on date of birth and last name.
    a. Remove matched record pairs from datasets.
    b. Categorize record pairs as possible matches.
6. Link unmatched records from datasets on social security number.
    a. Remove matched record pairs from datasets.
    b. Categorize record pairs as possible matches.
7. Categorize unlinked records from both datasets as certain non-matches.
8. Perform manual review of possible matches.
    a. Categorize record pairs as certain matches or certain non-matches.
    b. Un-link records categorized as certain non-matches.
9. Update client index.
    a. Variables stored in index are distinct internal identifiers used in datasets.
    b. Certain matches: update index with both identifiers in each record.
    c. Certain non-matches: update index with single identifier in each record.

APPENDIX 2

Queries

Abbreviations

- SSN: social security number
- DOB: date of birth
- last: last name
- first: first name

This results of this query are linked record pairs to be designated certain matches.

```
SELECT Dataset1.*, Dataset2.*

FROM Dataset1 a INNER JOIN Dataset2 b

ON a.first = b.first

AND a.last = b.last

AND a.SSN = b.SSN

AND a.DOB = b.DOB
```

The results of this query are the records from one of the datasets not linked by the previous query.

```
SELECT Dataset1.*

FROM Dataset1 a LEFT JOIN Dataset2 b

ON a.first = b.first

AND a.last = b.last

AND a.SSN = b.SSN

AND a.DOB = b.DOB

WHERE b.identifier IS NULL
```

The queries to link records on three variables and remove matched record pairs from the datasets share the structure of those above. As an alternative, the following single query can be used to identify all of the certain matches at once. Data processing in order to identify multi-digit social security number discrepancies was not performed as part of this project. Because of this constraint, records matched on the combination of last name, first name, and date of birth could not be automatically considered certain matches. They were included in this query, but still required manual review.

```
SELECT Dataset1.*, Dataset2.*
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB
UNION
SELECT Dataset1.*, Dataset2.*
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB
WHERE a.identifier NOT IN
(SELECT a.identifier
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB)
UNION
SELECT Dataset1.*, Dataset2.*
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN
WHERE a.identifier NOT IN
(SELECT a.identifier
```

```
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB)
UNION
SELECT Dataset1.*, Dataset2.*
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.SSN = b.SSN AND a.DOB = b.DOB
WHERE a.identifier NOT IN
(SELECT a.identifier
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB)
UNION
SELECT Dataset1.*, Dataset2.*
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.DOB = b.DOB
WHERE a.identifier NOT IN
(SELECT a.identifier
FROM Dataset1 a INNER JOIN Dataset2 b
ON a.first = b.first AND a.last = b.last AND a.SSN = b.SSN AND a.DOB = b.DOB)
```