A Study of Factors Affecting Record Linkage in Federal Statistical Databases

Michael D. Larsen[1] and Yuan Zhao[2]

[1]The George Washington University, Department of Statistics
[2]The George Washington University, Biostatistics Center
[1]6110 Executive Boulevard, Suite 750, Rockville, MD 20852; mlarsen@bsc.gwu.edu
[2]6110 Executive Boulevard, Suite 750, Rockville, MD 20852; yzhao@bsc.gwu.edu

Abstract: Record linkage, or exact file matching, consists of bringing together records in two or more files on the same population. Files are linked for the purposes of creating a larger database, enabling analyses that would otherwise not be possible, and counting the population. When unique, error-free identification codes are not available on both files, then record linkage can be accomplished through probabilistic methods. When implementing matching algorithms, one must choose matching variables, define for each variable what it means to agree or disagree, choose blocking factors that restrict the space of comparison pairs, and decide the level of evidence required to declare that a pair of records is probable match. The National Center for Health Statistics (NCHS) uses record linkage to match surveys, such as the National Health Interview Survey (NHIS) or the National Health and Nutrition Examination Survey (NHANES) to the National Death Index (NDI) for studies of mortality and morbidity. Based on files simulated to be similar to the NDI and a NCHS health survey, some choices that affect the performance of probabilistic record linkage are studied. The impact of several choices as well as file sizes and recording errors are compared. The work has direct relevance for improving and evaluating record linkage operations in the federal statistical system.

Key words and phrases: Exact matching; file linking; blocking; probabilistic linkage; NHIS; NHANES; National Death index.

1. Introduction

A simulated National Death Index-User data set provided by Bryan Sayer was analyzed using latent class analysis record linkage models. Factors studied included blocking factors, the impact of the inclusion of select variables in the latent class models, and the impact of errors in matching variables. Blocking factors in the NDI-User data files are very important in identifying correct matches and in reducing computational effort. The NDI-User data files can be successfully matched as long as there are sufficient matching variables available and errors in files are not too large. More matching errors are made when fewer variables are available for matching and there are more errors in the NDI-User data files.

A simulated NDI-User data set was created by Bryan Sayer. He describes the creation of the data set as follows:

> The first name data is from the 250 most popular sex specific first names. Middle names are also drawn from the first name file, but use all unique first names to draw from. Middle initial is chosen based on the probability of occurrence, and then a middle name is chosen that matches the middle initial. Not all records have middle initial or name, as about 25% of NDI records have no middle initial. Year of birth is constrained to 1900 to 1919. Year of Death is chosen to approximate age at death based on the 2000 data. Sex is split 50/50. Month and day values are chosen randomly, except that day must agree with month. All other values are chosen with probability equal to the values in the NDI file for 1979 to 2000.

Latent class analysis for record linkage has been described by Larsen and Rubin (2001), Lahiri and Larsen (2005), Larsen (2004), and references therein. See also Larsen (2010). It is a model-based statistical method for estimating the probability that two record pairs, one of each from two files, pertain to the same record. One output of the

algorithm is an effective scoring system for measuring the contribution to the likelihood of agreement overall based on individual variables. When different combinations of variables are used in the algorithm, it produces different estimated probabilities of agreement. Estimation is accomplished by maximizing the latent class log likelihood function via the EM algorithm (Dempster, Laird, and Rubin 1977).

Factors studied in this research on the simulated NDI-User data set include blocking factors, the impact of the inclusion of select variables in the latent class models, and the impact of errors in matching variables.

2. Blocking Factors

Blocking factors are used to decrease the number of comparisons that must be made. A comparison of all 10,000 simulated user submissions to all 995,043 simulated NDI records would yield a huge number of comparisons. Ideally a blocking factor eliminates record pairs that are not matches, but does not exclude any record pairs that are matches. One possible blocking criterion is SSN. Exact agreement on SSN, however, is rather strict, even for a set of actual matches. It would tend to produce blocks with zero or one potential matching pair of records. As such, it is an excellent matching variable when it is available, but it does not have much potential as a blocking variable, at least in the sense of which blocking variables are construed for this research.

The last four digits of SSN, however, can be used as a blocking variable. For each value of the last four digits of SSN there are several potential matches in the data set. In the simulated NDI-User data sets, there are on average nearly 100 cases with the same last four digits of SSN: 995,043/10,000 = 99.5. The table below summarizes the number of NDI cases by SSN4. If SSN4 is used as a blocking variable by itself, it would produce 99.5*10,000 = 9,950,000 comparison pairs. At most 5,043 of these (the number of deaths in the simulated user file) are true matches. For the purposes of this research, the number of pairs according to this blocking definition was too large to be considered.

Summary of distribution of number of cases with matching SSN4 values in the NDI data set for each SSN4 in the User data set							
Minimum	1st quartile	Median	3rd quartile	Maximum	Mean	Standard deviation	n
65	93	99	106	139	99.5	10.1	995,043

Blocking based on names is another possibility. Matches to names can be made by matching last name on the User and NDI files, birth surname on the User and NDI files, and last name on the User file and birth surname on the NDI file. The third name match scenario primarily is relevant for females who might change their name at time of marriage, and is referred to as 'maiden name' below, even though agreement would happen for individuals who had not changed their name. As can be seen in the table below, the number of record pairs defined by name-based blocking criteria is even larger than the number of record pairs defined by the SSN4 blocking criterion. For the purposes of this research, the number of pairs according to these blocking definitions was too large to be considered. It is noteworthy, however, that a significant number of potential record pairs are allowed by one blocking criterion but not the others. In considering a name-based blocking criterion for the NDI, therefore, one should consider simultaneous inclusion of cases based on multiple name criteria.

Total number of record pairs allowed by some name-based blocking criteria			
Name-Based Blocking Variables, Agree on Any in a Given Row			Number of Resulting Record Pairs
Last name			13,435,375
Birth surname			13,159,741
Maiden name			13,422,836
Last name	Birth surname		22,205,314
Last name	Maiden name		19,228,383
Birth surname	Maiden name		18,869,853
Last name	Birth surname	Maiden name	24,670,117

The next blocking criterion considered is one that requires agreement on SSN4 and on a name-based match. The table below gives the average number of record pairs per user record and the maximum number of record pairs per user record that are produced by these blocking criteria for the NDI-User data set. The results below illustrate the

extreme impact that blocking can make in an application such as matching to the NDI. Changing the blocking criteria can make more than a 100 (or a 1,000) fold decrease in the number of candidate pairs. We will refer to the blocking method in the last line of the table below as the 'narrow' blocking criterion in the rest of this article.

Average and total number of record pairs allowed by SSN4 + name-based blocking criteria			
Blocking Variables, Agree on All in a Given Row		Average number of record pairs per user record	Total number of record pairs
SSN4		99.5	995,043
SSN4	Last name	0.6356	6,356
SSN4	Birth surname	0.6320	6,320
SSN4	Maiden name	0.4162	4,162
SSN4	Any of 3 names	0.7429	7,429

The next blocking criterion considered is one that requires agreement on first letter of first name plus agreement on some last name variable. First letter of last name in principle divides the NDI records into 26 classes. It is not as extreme as SSN4, which has 10,000 classes. For the purposes of this research, however, the number of pairs according to this blocking definition was too large to be considered. As before, the results below illustrate the extreme impact that blocking can make in an application such as matching a survey to the NDI.

Maximum and total number of record pairs allowed by first letter of first name + name-based blocking criteria				
Blocking Variables, Agree on All in a Given Row			Maximum number of record pairs per user record	Total number of record pairs
	First of first name	Last name	1,770	917,227
	First of first name	Birth surname	1,711	916,897
	First of first name	Maiden name	1,711	914,328
	First of first name	Last name and/or birth surname	2,984	1,492,664
	First of first name	Last name and/or maiden name	2,984	1,308,696
	First of first name	Birth surname and/or maiden name	2,775	1,298,410
	First of first name	Any of three names	3,530	1,690,309

To further refine the blocking criteria, we then considered agreement on state of residence in addition to agreement on first letter of first name plus agreement on some last name variable. Inclusion of state of residence in principle should cut the size of blocking classes by a factor of 50. The blocking criterion defined in the last row of the table below will be referred to as the 'broad' blocking criterion in the rest of this article.

Maximum and total number of record pairs allowed by state of residence + first letter of first name + name-based blocking criteria				
Blocking Variables, Agree on All in a Given Row			Maximum number of record pairs per user record	Total number of record pairs
State reside	First of first name	Last name	164	41,740
State reside	First of first name	Birth surname	167	41,739
State reside	First of first name	Maiden name	167	39,374
State reside	First of first name	Last name and/or birth surname	280	65,116
State reside	First of first name	Last name and/or maiden name	280	57,542
State reside	First of first name	Birth surname and/or maiden name	256	57,163
State reside	First of first name	Any of three names	327	73,151

3. Matching Variables

Ten matching variables are considered in analyses in this paper. They are listed in the table below. The variable 'state.resid' (state of residence) cannot be used for matching with the 'broad' blocking criterion, because all candidate pairs are required to agree on state of residence between the NDI and the User file. SSN could potentially be used as a matching variable, but doing so might be a little odd in the NDI application. Instead, SSN would likely

be used to find exact matches or to confirm matches found with other means. SSN4 could potentially be considered a matching variable in other contexts. First name can be used as a matching variable because only first letter of first name is used as a blocking variable. Birth surname can be used as a matching variable because some record pairs qualified under the blocking criteria based on last name or maiden name alone. Further research could assess whether birth surname could effectively be excluded from matching after it plays some role in blocking.

Matching Variables			
Abbreviation	Variable	Match variable with narrow blocking	Match variable with broad blocking
Fname	First name	Yes	Yes
Sex	Sex	Yes	Yes
Race	Race	Yes	Yes
Birth.m	Birth month	Yes	Yes
Birth.d	Birth day	Yes	Yes
Birth.y	Birth year	Yes	Yes
Marital	Marital status	Yes	Yes
State.resid	State of residence	Yes	
State.birth	State of birth	Yes	Yes
Surname.birth	Surname at birth	Yes	Yes

In the studies reported below, matching variables are used in different combinations for matching the NDI-User records. First, all available matching variables are used together in a single latent class model. Second, one variable from the list above is removed and the analysis is redone. This yields nine analyses under the narrow blocking criterion and eight analyses under the broad blocking criterion. Third, four subsets of matching variables are examined. The subsets are described in the table below. In the first scenario, there is no birthday information. In the second scenario, there is no first name and only birth year. In the third scenario, there is no first name, only birth year, and no state of birth. In the fourth scenario, there is no first name, only birth year, and no surname at birth. Subsets used with the broad blocking criterion exclude the variable state of residence in addition to the matching variables indicated in the four scenarios. Future work could consider subsets of variables actually being considered for NDI-survey matches.

Matching Variables used in Four Subset Matching Scenarios; subsets used with the broad blocking criterion exclude variable state of residence				
Variable	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
First name	Yes			
Sex	Yes	Yes	Yes	Yes
Race	Yes	Yes	Yes	Yes
Birth month				
Birth day				
Birth year		Yes	Yes	Yes
Marital status	Yes	Yes	Yes	Yes
State of residence	Yes	Yes	Yes	Yes
State of birth	Yes	Yes		Yes
Surname at birth	Yes	Yes	Yes	

4. Random Errors

Random errors were introduced into the matching process by randomly changing a comparison vector element from 1 to 0 or from 0 to 1. Changing a comparison vector element from a 1 to a 0 implies that there is an error in comparing two entries that should agree on a field of comparison. Changing a comparison vector element from a 0 to a 1 implies that there is an error in comparing two entries that should disagree on a field of comparison.

Random errors were introduced independently for each of the ten comparison fields. Error rates studied were taken as 5%, 10%, or 20%. When there are these levels of random errors many records have some error. Most records, however, do not have many errors. A random agreement in most cases is insufficient to make a nonmatching record pair appear to be a matching record pair. The table below shows the probability of having 0, 1, .., 10 random errors

in an individual comparison vector when there are 5%, 10%, or 20% errors per comparison field. Future work could consider scenarios in which comparison fields have different rates of errors and scenarios in which errors across fields of comparison are correlated.

Probabilities of number of errors at 3 levels of random error	Number of random errors in ten comparisons								
	0	1	2	3	4	5	6	7	8-10
5% random error by field	0.60	0.32	0.07	0.01	0.00	0.00	0.00	0.00	0.00
10% random error by field	0.35	0.39	0.19	0.06	0.01	0.00	0.00	0.00	0.00
20% random error by field	0.11	0.27	0.30	0.20	0.09	0.03	0.01	0.00	0.00

5. Results

Results when using the narrow blocking criterion

Results are reported in this section when the narrow blocking criterion, which is defined as agreement on SSN4 and one of three last name variables, is used to define blocks. Results using the broad blocking criterion are reported in the next section. In interpreting these results, it is important to remember that the blocking criterion reduces the number of potential matches to a small number. Some individuals in the User file have only one or sometimes just two potential matches in the NDI file. Thus, the goal is for the matching variables to further clarify the matching situation for each user record. A latent class model with two latent classes was fit to the multidimensional contingency table defined by the several (binary) matching variables.

When all ten matching variables are used, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, for the deceased individuals it makes no errors in finding their matched cases.

When one variable is removed at a time and the algorithm re-run, the algorithm again produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, using any nine of the available predictor variables, for the deceased individuals it makes no errors in finding their matched cases.

When groups of variables are removed together as defined in the four scenarios of Section 3 and the algorithm re-run, the algorithm yet again produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, using the predictor variables in these four sets of predictors, for the deceased individuals it makes no errors in finding their matched cases.

Why is there such a high success rate? In interpreting these results, it is important to remember that the blocking criterion reduces the number of potential matches to a small number. Some individuals in the User file have only one or sometimes just two potential matches in the NDI file. The matching variables, then, are able to clarify the matching situation. The cases with the highest estimated probabilities of matching are the true matching cases.

A 5% chance of an error in comparing matching variables was randomly introduced. When all ten matching variables are used, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, for the deceased individuals it makes no errors in finding their matched cases. Thus, the 5% random errors did not decrease effectiveness. As was discussed, most cases have zero fields or one field with random errors, so information is not substantially degraded. When nine matching variables are used, the algorithm again produced zero errors, except in the case when day of birth is removed. In part this could be random variation due to simulation. It could be, however, that day of birth, which has 31 levels, is an important variable for further clarifying match status. In the case that day of birth was not used in matching 1 error was made in finding the 5,043 deceased individuals in the NDI file. When the four subsets of variables are used for matching and 5% errors are present, matching scenarios 1, 2, 3, and 4 made 1, 4, 4, and 5 errors, respectively. As expected, the reduced set of matching variables lead to more difficulty in matching. Still, the latent class model results are quite good.

A 10% chance of an error in comparing matching variables was randomly introduced. When all ten matching variables are used, the algorithm produces estimated probabilities that correctly match the 5,043 deceased

individuals in the User file to their records in the NDI simulated file. That is, for the deceased individuals it makes no errors in finding their matched cases. Thus, the 10% random errors did not decrease effectiveness. When nine matching variables are used, the algorithm again produced zero errors, except in the case when day of birth is removed. In the case that day of birth was not used in matching 1 error was made in finding the 5,043 deceased individuals in the NDI file. When the four subsets of variables are used for matching and 10% errors are present, matching scenarios 1, 2, 3, and 4 made 4, 5, 15, and 7 errors, respectively. As expected, the reduced set of matching variables lead to more difficulty in matching. Still, the latent class model results are quite good.

A 20% chance of an error in comparing matching variables was randomly introduced. When all ten matching variables are used, the algorithm produces estimated probabilities that correctly match all but 9 of the 5,043 deceased individuals in the User file to their records in the NDI simulated file. Thus, even the 20% random errors did not significantly decrease effectiveness. When nine matching variables are used, the algorithm produced some more errors. The number of errors when a variable is excluded from matching is given in the table below.

Matching Errors under Narrow Blocking when there is a 20% chance of error by variable and 9 of 10 variables are used for matching		
Abbreviation	Variable excluded from matching	Number of errors when variable is not used for matching
Fname	First name	13
Sex	Sex	12
Race	Race	10
Birth.m	Birth month	17
Birth.d	Birth day	15
Birth.y	Birth year	14
Marital	Marital status	12
State.resid	State of residence	17
State.birth	State of birth	19
Surname.birth	Surname at birth	10

When the four subsets of variables are used for matching and 5% errors are present, matching scenarios 1, 2, 3, and 4 made 36, 36, 59, and 38 errors, respectively. As expected, the reduced set of matching variables lead to more difficulty in matching. Still, the latent class model results are quite good.

Results when using the broad blocking criterion

Results are reported in this section when the broad blocking criterion, which is defined as agreement on state of residence, first letter of first name, and one of three last name variables, is used to define blocks.

When all ten matching variables are used, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, for the deceased individuals it makes no errors in finding their matched cases. When one variable is removed at a time and the algorithm re-run, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, using any nine of the available predictor variables, for the deceased individuals it makes no errors in finding their matched cases.

When groups of variables are removed together as defined in the four scenarios of Section 3 and the algorithm re-run, the algorithm produces estimated probabilities that correctly match most of the 5,043 deceased individuals in the User file to their records in the NDI simulated file, but some errors are made. The matching scenarios 1, 2, 3, and 4 made 8, 4, 25, and 4 errors, respectively. Scenario 3, which did not use first name, day or month of birth, or state of birth as matching variables had the most problem.

A 5% chance of an error in comparing matching variables was randomly introduced. When all nine matching variables are used with the broad blocking criterion, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. That is, for the deceased individuals it makes no errors in finding their matched cases. Thus, the 5% random errors did not decrease effectiveness. When eight of the nine matching variables are used, the algorithm again produced zero errors. When

the four subsets of variables are used for matching and 5% errors are present, matching scenarios 1, 2, 3, and 4 made 1, 0, 8, and 0 errors, respectively. The result is paradoxical, because results are better with a little error introduced. Why could this happen? This could happen because one or more variables received an exaggerated weight under one of the scenarios with the original data. The random errors could have lead to a reduction in the importance of a particular variable in terms of determining matching probabilities, which ultimately increased the match rate. As mentioned before, these are simulated results so there could be a degree of simulation variance involved. Further study of this phenomenon could be undertaken in future work.

A 10% chance of an error in comparing matching variables was randomly introduced. When all nine matching variables are used under the broad blocking criterion, the algorithm produces estimated probabilities that correctly match the 5,043 deceased individuals in the User file to their records in the NDI simulated file. When eight of the nine matching variables are used, the algorithm again produced zero errors. When the four subsets of variables are used for matching and 10% errors are present, matching scenarios 1, 2, 3, and 4 made 7, 4, 17, and 3 errors, respectively. As expected, the reduced set of matching variables lead to more difficulty in matching. Still, the latent class model results are quite good.

A 20% chance of an error in comparing matching variables was randomly introduced. When all nine matching variables are used, the algorithm produces estimated probabilities that correctly match all the deceased individuals in the User file to their records in the NDI simulated file. Thus, even the 20% random errors did not significantly decrease effectiveness. When eight of the nine matching variables are used, the algorithm produced some errors: there was one error each in matching deceased individuals when variable “birth month” or variable “birth year” were removed. Why could results be better for the broad blocking scenario versus the narrow blocking scenario? In the particular cases considered for the NDI-User data files, the narrow blocking scenario contains very few NDI cases per user record. In many aspects, the user and NDI records match quite closely. Therefore they might be difficult to distinguish. In the broad blocking scenario, there are multiple NDI cases per user record. Many of those NDI cases are quite dissimilar from the user records. Therefore, many of the additional NDI cases have little risk of being finally associated with the user cases, resulting in a false match. This rationale is speculative, but it will be interesting to study in future applications. When the four subsets of variables are used for matching and 20% errors are present, matching scenarios 1, 2, 3, and 4 made 23, 8, 48, and 10 errors, respectively. As expected, the reduced set of matching variables lead to more difficulty in matching.

6. Summary

A simulated NDI-User data set provided by Bryan Sayer was analyzed using latent class analysis record linkage models. Factors studied included blocking factors, the impact of the inclusion of select variables in the latent class models, and the impact of errors in matching variables. Blocking factors in the NDI-User data files are very important in identifying correct matches and in reducing computational effort. The NDI-User data files can be successfully matched as long as there are sufficient matching variables available and errors in files are not too large. More matching errors are made when fewer variables are available for matching and there are more errors in the NDI-User data files.

Future work could consider additional blocking criterion and alternate uses of matching variables. The alternate uses of matching variables could include degree of agreement, agreement simultaneously on two or more variable (to incorporate interactions between variables), and partial agreement definitions. Future work could also study the impact of matching errors on analyses performed after matching. The latter study interest would require a different simulated or actual data set than the one used in this article. Finally, a few research questions were identified as part of this research, as mentioned in the primary text of this paper. These topics could be investigated in future research and could be reasons that it is quite difficult to improve record linkage effectiveness in some NDI situations.

References

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 78, 221-227.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American*

Statistical Association, 100, 222-230.

Larsen, M.D. (2004). Record linkage using finite mixture models. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives.* Wiley. 309-318.

Larsen, M.D. (2006). Entry for the Sage *Encyclopedia of Measurement and Statistics* on ``Latent Class Analysis."

Larsen, M.D. (2006). Entry for the Sage *Encyclopedia of Measurement and Statistics* on ``Record Linkage."

Larsen, M.D. (2006). Ideas for secure record linkage. *Privacy in Statistical Databases 2006*. [CDROM]. CENEX-SDC Conference. Rome, Italy.

Larsen, M.D. (2010). Record Linkage Modeling in Federal Statistical Databases. *FCSM research conference 2009*. http://www.fcsm.gov/09papers/Larsen_II-C.pdf.

Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Sayer, B. (2006). Comparing BigMatch Results to Current National Death Index (NDI) Selection Methods. *Proceedings of the Survey Research Methods Section* (Alexandria, VA: American Statistical Association): 3648-55.

Appendix: File Layouts

NDI File (N=995,043)

Item	Start	End
State of Death	1	2
NYC indicator (not used)	2	5
Certificate Number	6	11
Last Name	12	31
First Name	32	46
Middle Initial	47	47
Sex	49	49
Month of Death	50	51
Day of Death	52	53
Year of Death (two digit)	54	55
Race	56	56
Age Units (for age at death)	57	57
Age (2 digits)	58	59
Month of Birth	60	61
Day of Birth	62	63
Year of Birth (two digit)	64	65
State of Birth	66	67
Marital Status	68	68
Social Security Number	69	77
State of Residence	78	79
Birth Surname	80	88
NYSIIS Last Name	100	109
NYSIIS First Name	110	119
NYSIIS Birth Surname	120	129
Year of Death (4 digits)	135	138
Year of Birth (4 digits)	139	142

User File (10,000 people, 5,043 deceased, 4,957 non-deceased 128,270 submission records)

Item	Start	End
Last Name	1	20
First Name	21	35
Middle Initial	36	36
Social Security Number	37	45
Month of Birth	46	47
Day of Birth	48	49
Year of Birth (4 digit)	50	53
Birth Surname	54	71
Age Unit	72	72
Age (2 digit)	73	74
Sex	75	75
Race	76	76
Marital Status	77	77
State of Residence	78	79
State of Birth	80	81
Year of Death (or Interview)	82	85
State of Death	86	87
Certificate Number	88	93
Type of Submission Record	98	99
Vital Status (Deceased or Alive)	100	100
Middle Name	101	115
NYSIIS Last Name	116	125
NYSIIS First Name	126	135
NYSIIS Middle Name	136	145
NYSIIS Birth Surname	146	155