

Data Matching Algorithm An Example from Oklahoma

The data matching algorithm used by the Oklahoma Department of Mental Health and Substance Abuse Services (ODMHSAS) was developed using an article by Matthew A. Jaro published in the *Statistics in Medicine Journal*, Vol. 14, 491-198 (1995) titled Probabilistic Linkage of Large Public Health Data Files. The matching algorithm involves blocking, matching weights, and a threshold of likelihood. Blocking is used as a first step to reduce the number of records compared within two files. Blocking is an exact match on selected variables between records within each file that are highly reliable, e.g., sex. Weights are assigned to additional variables that may not match exactly, due to spelling variability or number transposition. Finally, the two records are considered to match or not depending on the sum of the weights. The sum is compared to a predetermined threshold of likelihood. If the sum is above the threshold, the records are considered to be for the same individual and, if the sum is below the threshold, the records are considered to be for different individuals.

ODMHSAS uses a Microsoft SQL 7.0 program to perform the data matching. In the program, one record in the first file is compared to all records in the second file. The blocking is done on sex, date-of-birth components (month+day, day+year, month+year) and social security digits. If sex and date-of-birth components match exactly, the record in the second file is kept and checked for matching on SSN. If both of the SSN fields are not blank, the first three numbers of the SSN match or the second three numbers of the SSN match or the last three numbers of the SSN match the record in the second file is kept for further matching. If one of the SSN fields is blank, the record in the second file is kept for further matching if both of the last name fields are not blank and the first three letters of the last name match.

Following the blocking, weights are assigned to each record as follows:

SOCIAL SECURITY NUMBER	
9 digits in the correct order	22.95
8 digits in the correct order	16.89
7 digits in the correct order	8.44
Blank (either field)	0.00
Less than 7 digits	- 2.38
LAST NAME	
All letters match	9.58
First 3 letters match	5.18
Blank (either field)	0.00
No match on first 3	-3.62
FIRST NAME	
All letters match	6.69
First 3 letters match	3.37
Blank (either field)	0.00
No match on first 3	-3.27

MIDDLE INITIAL	
Match	3.65
DATE-OF-BIRTH	
Match	6.22
INITIALS	
If Sex=M and 7 digits of SSN match and first and last name are missing and no match on first and last initial	-8.00
If Sex=F and 7 digits of SSN match and first name missing and no match on first initial	-8.00

Weights are assigned to each variable based on the probability of the records matching on that variable. There are two probabilities associated with each variable: the probability of a match on the variable, given the records matched are for the same individual (m); and the probability of a match on the variable, given that the records matched are not for the same individual (u). Using the two probabilities, the weight is calculated as the logarithm to the base two of the ratio of m and u .

Following the assignment of weights, the weights for a pair of records are summed and compared to the threshold of likelihood of 17.73. To establish the threshold of likelihood, visual inspection of the matched records is required to determine the best threshold that will result in the fewest false positives and false negatives.

The weights and threshold described above are for the variables listed in this example. Often not all those same variables are included in an available dataset. Some state agencies may not have, or are not willing to provide, the full SSN. Therefore, the SSN may be excluded completely, or the last four digits will be provided instead of the full SSN. Each time a new set of variables is used for matching, the weights may need to be recalculated and the threshold will need to be reset. For further information on probabilistic matching, blocking, and computations of weights, consult the Jaro paper referenced above.