Integrating the New York Citywide Immunization Registry and the Childhood Blood Lead Registry

Vikki Papadouka, Paul Schaeffer, Amy Metroka, Andrew Borthwick, Parisa Tehranifar, Jessica Leighton, Angel Aponte, Ruron Liao, Alexandra Ternier, Stephen Friedman, and Noam Arzt

In February of 2004, the New York City Department of Health and Mental Hygiene completed the integration of its childhood immunization and blood lead test registry databases, each containing over 2 million children. A modular approach was used to build a separate integrated system, called Master Child Index, to include all children in both the immunization and lead test registries. The principal challenge of this integration was to properly align records so that a child represented in one database is matched with the same child in the other database. To accomplish this task as well as to identify internal duplicate records within each database, an artificial intelligence record linkage system was created. The preliminary results show high rates of accurate merging of records both within and between the two databases. The 4,610,585 records contained in both databases before Master Child Index implementation consolidated into 2,977,290 records in the integrated system. The matching system eliminated 523,720 duplicate records within the two databases and matched and merged 1,109,575 records between the two databases. The Department of Health and Mental Hygiene plans to further develop the Master Child Index and use it as the department-wide, record-matching system.

KEY WORDS: artificial intelligence, data quality, database, duplicate records, immunizations, integration, lead, New York City

The goal of integrating health information from different databases is to improve health care by providing more complete information to medical professionals, individuals, and families, as well as to inform more comprehensive public health interventions. The greatest challenge to accomplishing this goal is to accurately match and merge health records without a unique identification number across databases to create a single, complete, and accurate health record per person.

In February of 2004, the New York City (NYC) Department of Health and Mental Hygiene (DOHMH) completed the integration of its childhood immunization and blood lead test registry databases, each containing over 2 million children. This article describes the background, implementation, and preliminary results of this effort, followed by a discussion of plans

Dr. Borthwick is the President of ChoiceMaker Technologies, Inc., the company that developed ChoiceMaker 2.

Corresponding author: Amy Metroka, MSW, Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, 2 Lafayette Street, 19th Floor, New York, NY 10007 (e-mail: ametroka@health.nyc.gov).

Vikki Papadouka, PhD, MPH, is Director of Research and Evaluation of the Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, New York.

Paul Schaeffer, MPA, is Director of the Master Child Index, Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, New York.

Amy Metroka, MSW, is Director, Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, New York.

Andrew Borthwick, PhD, is President, ChoiceMaker Technologies, Inc., New York City.

Parisa Tehranifar, DrPH, is Deputy Director of Research and Surveillance, Lead Poisoning Prevention Program, New York City Department of Health and Mental Hygiene, New York.

Jessica Leighton, PhD, is Assistant Commissioner, Bureau of Environmental Disease Prevention, New York City Department of Health and Mental Hygiene, New York.

Angel Aponte, BA, is a Computer Specialist, Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, New York.

Ruron Liao, MS, is an Associate Staff Analyst, Lead Poisoning Prevention Program, New York City Department of Health and Mental Hygiene, New York.

Alexandra Ternier, MPH, is a Public Health Epidemiologist, Citywide Immunization Registry, New York City Department of Health and Mental Hygiene, New York.

Stephen Friedman, MD, MPH, is Assistant Commissioner, Bureau of Immunization, New York City Department of Health and Mental Hygiene, New York.

Noam Arzt, PhD, is President, HLN Consulting, LLC, San Diego, California.

J Public Health Management Practice, 2004, November(Suppl), S72–S80

for integrating the department's disease surveillance programs.

In 1999, the leadership of the NYC DOHMH issued a directive to integrate the department's existing Citywide Immunization Registry (CIR) and childhood blood lead test registry, known as LeadQuest (LQ). Already, by 1999, both registries were large: 1.7 million children in the CIR, and 1.3 million in LQ. The rationale for integration was clear: CIR and LQ target the same population of preschool and school-aged NYC children. Both programs were separately devoting vast resources to matching and merging records for potentially the same children in separate data systems. The inability to relate information for the same child across separate systems resulted in missed opportunities for identifying children at risk for both vaccine-preventable disease and lead poisoning.

The benefits to integration were many. First, each registry had valuable information to offer the other. The CIR, which loads birth certificate information from vital records for all children born in NYC (approximately 2,400 records per week), offers a population base to LQ, which contains the records of only those children tested for lead. The LQ contains more complete, accurate, and standardized address information than the CIR, owing primarily to its link to Geosupport, the NYC Department of City Planning's comprehensive, standardized database of addresses. More accurate addresses would improve both CIR's outreach to parents to recall children needing immunizations and strengthen its geographical analyses of immunization coverage levels across the city to target interventions to the most at-risk children.

Second, both programs expected system integration to result in leveraging each program's resources for outreach and education to providers, health plans, parents, and communities. The CIR, having been designed for easy access by medical professionals through telephone, fax and the Web, offered LQ a means to make available lead test information to providers. The LQ had been designed primarily to support case management of blood lead tests and surveillance. For the CIR, adding lead screening histories to its Web-based immunization registry was seen as expanding the value of this application to providers and increasing its use. Easy access by providers to combined lead screening and immunization records for their patients was expected to reduce missed opportunities to immunize or screen children for lead poisoning, as well as avoid unnecessary vaccinations and lead tests. Additionally, both programs were working toward establishing data-sharing projects with external partners, including Medicaid Managed Care Organizations (MCOs) and the NYC public school system. Integration of data systems across the two programs would allow both programs to participate in these data-sharing processes, with the purpose of improving identification of and outreach to children needing immunizations and blood lead tests.

The NYC DOHMH expected that the integration of the two data systems would be a positive return on investment. Creating a centralized record matching and merging system for the two programs would dramatically reduce time and effort (ie, costs, borne by each program for this task). It was estimated that the full-time equivalent (FTE) staff devoted by the immunization and lead programs, combined, would be reduced from 12 FTEs to less than 6 FTEs.

The need for protecting the confidentiality of information contained in the integrated system, while allowing for appropriate access by internal DOHMH staff and external partners (ie, health care providers, MCOs, and authorized agencies), was carefully considered. A combination of legal and technical means was employed to limit access to and deter misuse of CIR and LQ information.

Before the integrated system was created, the NYC Health Code regulating the disclosure of confidential CIR data was amended to permit access to lead test records by health care providers and authorized agencies, subject to the same confidentiality requirements as those afforded to records in the CIR. These included requiring the provider or agency to submit sufficient identifying information to DOHMH to uniquely identify the child under his or her care. The information disclosed by DOHMH to providers and authorized agencies is limited to first name, last name, date of birth, gender, and a system-assigned unique identification number, along with the names and dates of immunizations, dates and results of lead tests, and recommendations for needed immunizations and lead tests. No locating information such as address or telephone number is revealed to providers, MCOs, or agencies. Further, providers and agencies receiving immunization and lead test records are not permitted to re-disclose this information, except for purposes of protecting the health of the child or others.

All persons having access to CIR and LQ records, both internal and external, are required to sign confidentiality statements specifying that they will use CIR and LQ information solely for public health purposes. Finally, to deter and detect inappropriate use of CIR and LQ information, the integrated system was designed to include a detailed audit log to track all users, both internal and external, who search, find, and view records. The date and time of each user's activity is recorded. The logs are monitored by DOHMH staff to detect suspicious activity.

Before proceeding with the integration, the Immunization and Lead Poisoning Prevention programs engaged in a planning process to identify a model for

Before proceeding with the integration, the Immunization and Lead Poisoning Prevention programs engaged in a planning process to identify a model for integration with the least amount of disruption to each registry's day-to-day operations and core mission, while laying the foundation for the benefits of integration to be realized.

integration with the least amount of disruption to each registry's day-to-day operations and core mission, while laying the foundation for the benefits of integration to be realized. A modular approach was agreed upon in which a separate integrated system, called Master Child Index (MCI), would be built to include all of the children in both the CIR and LQ, with each child having an assigned unique identifying number. This modular approach to integration offered the considerable advantage of allowing CIR and LQ to remain independent while gaining the ability to link information for the same children across the two data systems.

The Problem: Duplication of Child Records

Given that both CIR and LQ are continually updated by incoming records, creating a single record per child was a challenge to both CIR and LQ, even before integrating the two large registry databases. Young children, unlike adults, do not have unique identifying numbers. Although social security numbers may be assigned to children shortly after birth, these numbers are generally not collected by health care providers and, therefore, are unavailable to DOHMH for matching. Further, information for very young children may be incomplete or subject to change. For example, many children receive their first vaccination (ie, Hepatitis B) at birth, when they frequently have not yet been named. The matching system must correctly merge early immunization records submitted with the first name listed as, for example, "Boy," to later immunization records with the actual first name listed. Additionally, the diversity and unfamiliarity of names of NYC's multiethnic population (over 52% of births in NYC were to foreign-born mothers in 2002[1]) increase spelling errors in reporting.

Names, birth dates, addresses, telephone numbers, and parent information are taken into consideration for matching children's records when available. However, this information is not consistently provided on immunization and lead test submissions. As a result, while a matching system must be able to leverage this information when it is available, it must not be reliant on all of these fields being completed or being completed correctly.

Without a sufficiently powerful matching and merging application, MCI would include a large number of separate records for the same child, usually referred to as "duplicate" records. (We are using the term "duplicate" record to describe one or more records belonging to the same person, but containing different demographic information.) The success of the integration depended on the ability to accurately match and merge the large population of children in each database to create one, common population for both CIR and LQ in the MCI.

The Solution: An Automated, Artificial Intelligence Approach

Before the integration, both CIR and LQ were using custom-designed software to automatically match and merge incoming records. The LQ matching system employed a specific set of criteria, known as rules, for making matching decisions. Combined with human review by 5 full-time equivalent (FTE) staff, LQ's record duplicate rate was kept at an acceptable level of an estimated 10%.

The CIR was also using a rules-based approach to automatically match incoming records. This system alone proved ineffective, resulting in a 50% duplication rate. The CIR added an automated clean-up process using artificial intelligence (AI) matching software developed in collaboration with a consultant.[2] Immediately after its initial run, this matching system reduced the duplication rate from 50% to 15%–20%. The duplication rate would quickly climb, however, because large numbers of incoming duplicate records would continue to be added. CIR staff repeatedly ran the software to maintain the duplication rate at 15%–20%. Reducing the duplication rate further to a more acceptable level of less than 10% would have required an estimated 7 FTE staff performing human review. This level of staffing was not continuously available for this task however.

During the planning phase of the integration, the CIR was already in the process of enhancing its AI matching system. The CIR and LQ teams decided to work together to develop an improved version of this matching system for the MCI. The advantage of using AI software, instead of a rules-based software, is that it can make sense of conflicting information (eg, same first name, different spellings of last name, slightly different dates of birth [DOBs]).

In a rules-based system, each rule results in a definitive decision. A rule holding that two records with different DOBs do not belong to the same child will result in separating the records, regardless of any other similarities. In contrast, an AI system could merge these records if there are similarities that outweigh the differences in DOBs. One can, of course, create more complex rules that specify, for instance, that different DOBs are permissible so long as they are similar and the child's first and last name match and mother's maiden name matches. However, writing rules to account for all the possible combinations of fields being possibly: (1) the same, (2) different but similar, (3) different, or (4) invalid or missing would result in an exponential explosion in the number of rules.

The CIR had been using AI software to clean-up duplicate records that had already been added to the database. The plan for MCI was to use the AI software to match and merge records as they entered the databases, thereby preventing the creation of duplicate records. This was expected to reduce the duplication rate to less than 10% for CIR and 6% for LQ, and to match and merge the majority of records across the two systems.

Implementation

Matching system development

The AI matching system is designed to replicate the human decision-making process. The core decision-making process is built around a notion of "clues" (better known in the AI literature as "features"). Clues are attributes of each record, which argue for or against a "match" decision. The clues are not intended to be correct all or even most of the time, are assigned weights indicating their relative importance to a matching decision in the data and are combined into an overall probability, indicating the certainty with which two records belong to the same child. The MCI matching system contains 193 different clues that use all available fields and combination of fields in the record.[3]

To generate an appropriate set of clues, the characteristics of the data in both LQ and CIR databases were studied in depth by DOHMH staff and the MCI matching system consultant. In addition, suggestions from LQ and CIR human review staff were gathered to identify attributes of the data that are important when humans make decisions about whether two records belong to the same child.

Next, a representative sample of record pairs was selected for review by DOHMH staff. The DOHMH staff marked these record pairs with 1 of 3 decisions: (1) "match" if the reviewer believed the two records belonged to the same child, (2) "no match" if the reviewer believed the records represented different children, and (3) "potential match" if the reviewer was not able to confidently make a "match" or a "no match" decision. The system, using an AI machine learning process called "Maximum Entropy Modeling" (ME),[4,5] then assigned weights to the clues based upon the human decisions. Clues that were strongly associated with DOHMH decisions received a high weight, whereas clues that were only weakly linked received a low weight. These clue weights were combined into an overall probability such that the higher the probability, the more likely that the two records belong to the same child. Consequently, the system was tuned not only to the MCI data, but also to the DOHMH staff's judgment for determining whether pairs of records matched.

Some examples of MCI matching system clues related to the first name field only are listed in Box 1.

BOX 1 Examples of clues used by the Master Child Index matching system that incorporate the "first name" field

1. Do the first names match approximately using the well-known Soundex approximate name-match algorithm,[6] or using the less well-known NYSIIS algorithm,[7] or using the US Census Bureau's Jaro-Winkler algorithm?[8]
2. Is the first name a common first name? A match on first name "JENNIFER" will be assigned a lower weight than a match on first name "VASSILIKI" because of the relative frequency of occurrence of these names in New York City's population.
3. Do the first and last names match if you swap them in one of the two records? (Frequently the first and last names are reversed in reports submitted.)
4. Are first names the same but genders are different? If the name is gender-specific (eg, "JENNIFER"), then the gender difference will be discounted.

During the development process, DOHMH periodically evaluated the matching system's accuracy. This was done by comparing the decisions made by DOHMH staff to the decisions made by the system on a second set of record pairs. The DOHMH had to carefully weigh the tradeoff between accuracy and the number of records requiring human review. If accuracy was not satisfactory, with an acceptably low number of records requiring human review, more development was done. For example, clues were modified, added, or eliminated and assignment of weights to the clues was adjusted. Testing was repeated until the tradeoff between accuracy and human review was acceptable. DOHMH staff settled on 98.96% accuracy and 3.79% human review. This was expected to result in 1.04% false negatives (duplicates missed) and 0.0% false positives (false merges) and an acceptable volume of potential matches for human review. At that point, development was considered complete.

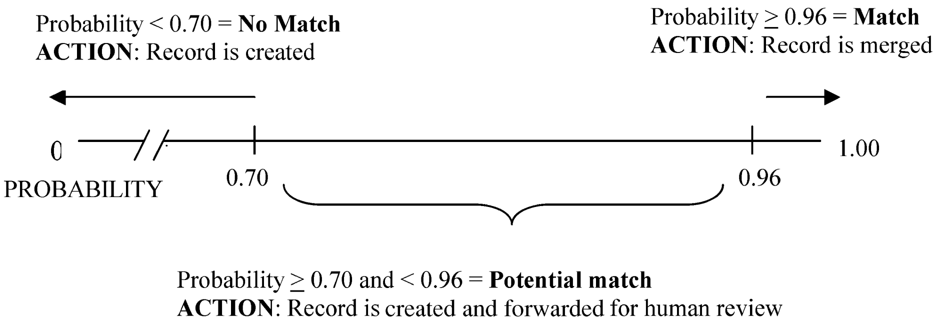
Probability < 0.70 = **No Match**
ACTION: Record is created
Probability ≥ 0.96 = **Match**
ACTION: Record is merged
0
PROBABILITY
0.70
0.96
1.00
Probability ≥ 0.70 and < 0.96 = **Potential match**
ACTION: Record is created and forwarded for human review

FIGURE 1. Probabilistic Scoring of Records and Corresponding Action in MCI.

Matching system threshold determination

The DOHMH decided on 0.96 as the "match" probability threshold, and 0.70 as the "no match" probability threshold. Thus, when an incoming record matched with an existing record with a probability of 0.96 or higher, this record would then be judged a "match" and would merge to the existing record. If the incoming record matched with an existing record with a probability below 0.70, it would then be considered a "no-match" and created in the MCI as a new record. Finally, if the incoming record matched with an existing record with a probability between 0.70 and 0.96, it would then be considered a "potential match," created as a new record, and sent to human review. This process is presented schematically in Figure 1.

While the decision to set the lower threshold at 0.70 was expected to result in a significant number of potentially duplicated records left in the system, it was believed to reduce the number of records in need of human review to a manageable size. The very high threshold of 0.96 was chosen to minimize the risk of records of different children being merged together and possibly resulting in not immunizing or not testing a child for lead poisoning.

All LQ and CIR records are evaluated by the matching system as they enter the MCI by being assigned a probability that will determine whether they will be merged, created, or created and sent to human review. There are some circumstances, however, where the probabilistic decision is overridden by a small set of rules. These rules were put in place to safeguard against false merging and force pairs to human review. For example, if a record in the database has more than 5 names, this record is sent to human review even if it matches with an incoming record above the 0.96 threshold.

Independent test on system's accuracy

The accuracy of the MCI matching system was tested using the Centers for Disease Control and Prevention's (CDC) "de-duplication toolkit," a method developed by CDC's National Immunization Program to evaluate registries' record de-duplication algorithms.[9] The toolkit, which uses a fictitious set of children's records, calculates sensitivity and specificity of the algorithm. Specificity refers to the ability of the system to identify nonduplicates and is measured as the percentage of all nonduplicate records that are correctly identified as nonduplicates. Sensitivity refers to the ability of the system to identify duplicates and is measured as the percentage of duplicate records detected out of all the known duplicates. MCI's matching system's specificity was 100% and sensitivity was 95.6%.

Results

Three types of analyses were conducted to evaluate the success of the MCI matching system in integrating the CIR and LQ databases: (1) matching and merging of "initial load" data, (2) matching and merging of current data, and (3) system accuracy.

Matching and merging of "initial load" data

The MCI database was created over a 7-week period between December 12, 2003 and January 29, 2004. This phase is referred to as the "initial load." First, the 970,567 vital records from the CIR database, most of them containing immunization information, were inserted into MCI. Next, the remaining 1.4 million CIR records that did not contain vital record information were loaded, followed by the nearly 2.2 million LQ records. Unlike vital records, which were inserted without going through the matching process, all other records were evaluated by the MCI matching system. Each nonvital CIR and LQ record was either: (1) matched and merged to an existing record, (2) not matched and created as a new record, or (3) created as a new record but also sent to the human review queue as a potential match to one or more existing records.

The number of records in each individual database was compared before and after the deployment of the

TABLE 1 ● **Number and percentage of matching results of the "initial load" data by system**

	Within system		Between system	Within and between system
	CIR	LQ	MCI	CIR, LQ, and MCI
Pre-MCI, *N*	2,426,369	2,184,216	4,086,865*	4,610,585
Post-MCI, *N*	2,065,230	2,021,635	2,977,290	2,977,290
Merged, *N*	361,139	162,581	1,109,575	1,633,295
Merged, %	14.9	7.4	27.1	35.4
Human review, *N*	74,798	56,747	95,886	227,431
Human review, %	3.1	2.6	2.3	4.9

*This number represents the sum of records in each data system after MCI's internal de-duplication, ie, 2,065,230 + 2,021,635 = 4,086,865.
CIR = Citywide Immunization Registry; LQ = Lead Quest; MCI = Master Child Index.

MCI to examine the rate of de-duplication within each system (within-system merges). Also, the total number of records that matched between the two systems (between-system or integration merges) was examined. These results are presented in Table 1.

The within-system merging was 14.9% for CIR and 7.4% for LQ. These percentages reflect the reduction in the duplication rate that existed in each system prior to MCI's implementation. The rate of merging is higher for the between-system merges. Of approximately 3 million records in MCI, 1,109,575 (37.3%) records were merged between the two systems. This percentage reflects the extent to which CIR and LQ contained records for the same population of children.

However, it is important to note that the two databases do not contain data for exactly the same population of children. LQ is an older database that began collecting blood lead tests for all NYC children in 1994, whereas the CIR was launched in 1997 and contains more complete data, including vital records, for children born in 1996 and after. It is therefore more appropriate to evaluate the "integration" by examining only those populations that are likely to have records in both systems. Table 2 presents the results of this analysis and shows that over 80% of LQ records merged with vital and/or CIR records in children born between 1996 and 2003. By contrast, the rate of merging for children born before 1996 was 40%.

The effect of removing duplicate records on each individual database was also assessed. Overall, 14.9% of records within the CIR were merged, but this percentage varied across all birth cohorts. The percentage of records merged is higher after 1996, because those cohorts contain vital records and have more duplicates. The percentage merged for those cohorts averaged 19.2. As a result of de-duplication, a number of other data quality improvements occurred. The percentage of vital records with one or more immunizations increased by 6.7%, from 67.7% to 74.4%. The percentage of children 4 months to 6 years with 2 or more immunizations increased by 5.1%, from 66.4% to 71.5%, and the percentage of children 19–35 months with 2 or more immunizations increased by 5.4%, from 65.0% to 70.4%.

TABLE 2 ● **Number and percentage of Lead Quest records merged with Citywide Immunization Registry or vital records**

Birth cohort	CIR	LQ	Integration merges	LQ records merged with CIR records, %
<1996 (no vital records)	851,460*	1,235,734*	494,595†	40.0
1996	157,818	133,368	105,280	78.9
1997	159,194	126,373	100,336	79.4
1998	154,415	124,180	99,236	79.9
1999	146,339	116,795	94,532	80.9
2000	150,899	107,048	87,802	82.0
2001	151,601	95,044	79,979	84.1
2002	148,015	74,892	63,228	84.4
2003	142,675	7,985	6,437	80.6
1996–2003	1,210,956*	785,685*	636,830†	81.1

*The total number of records in CIR and LQ in Table 2 is not exactly the same as in Table 1 because Table 1 includes children born in 2004 (CIR: 1,210,956 + 851,460 = 2,062,416 (Table 2) versus 2,065,230 (Table 1); LQ: 785,685 + 1,235,734 = 2,021,419 (Table 2) versus 2,021,635 (Table 1)).
†The total number of integration merges in Table 2 (636,830 + 494,595 = 1,131,425) is higher than the one presented in Table 1(*N* = 1,109,575). The reason for the discrepancy is that records in the Master Child Index may have more than one DOB (date of birth), and an "integrated" record may belong to more than one cohort and counted more than once.
CIR = Citywide Immunization Registry; LQ = Lead Quest.

The overall reduction in duplicate records within LQ was 7.4%, but unlike the CIR, the highest rate of de-duplication occurred in older cohorts (10.5% in the 1992 cohort). This was most likely due to the fact that LQ's pre–MCI de-duplication system was deployed after 1999 and was not used on older records. Furthermore, record merging increased the percentage of NYC children with at least 2 blood lead tests before their third birthday from 31.2% to 32.9%.

Matching and merging of current data

After the "initial load," both databases resumed their daily operations of loading current data. Unlike records that were initially loaded to MCI, which may have been previously merged to other records and become more complete, incoming records often contain fewer data elements that can be used for matching. As a result, the merging rates of new records may be lower than the "initial load" rates. The matching and merging rate of incoming (new) records was calculated by examining the percentage of records merged, created, or sent to human review, during the first few weeks after MCI's implementation.

Of the 284,191 incoming records processed through the CIR (vital records and immunization records), 225,603 (79.4%) merged with existing records, 29,471 (10.4%) were created as new records, and 29,117 (10.2%) were forwarded for human review. These percentages are roughly what would be expected given the profile of the data loaded over this period of time. Of the 284,191 incoming CIR records, 21,866 (7.7%) were new vital records and should have been created, whereas most of the remaining records should already exist in MCI. Thus, of the 29,471 new records created in MCI, only 2.7% were created from immunization records submitted for children not already in MCI. In contrast to the MCI merging rate of 79.4%, the estimated pre–MCI rate of merging for CIR was in the order of 60%. This represents an estimated 20% improvement in the match rate of incoming records, which dramatically reduces the number of duplicate records being inserted into the database**.**

A similar analysis for LQ reveals that of the 61,620 new LQ records processed, 51,983 (84.4%) merged with existing records, 5,356 (8.7%) were created as new, and 4,281 (6.9%) were sent to human review. The rate of merging in LQ is higher than in CIR because the CIR processes new vital records, and the children whose lead tests are currently processed should already exist in MCI through vital or CIR records. (Children get immunizations before they are tested for lead).

System accuracy

The specificity of the matching system was examined in production by selecting a sample of merged records and examining whether there were indications of incorrect merges. A random sample of 2,000 merges was selected and reviewed by DOHMH staff for indications of false merging. Of those, only 3 merged records did not have enough information to be definitely judged as a "merge," but could not be determined as false merges for certain. The accuracy of the MCI matching system was shown to be as high in operation, after implementation, as it was during the testing phase.

Since sampling merges in the system did not seem to point to any significant false merging activity, DOHMH data quality staff engaged in the process of finding false merges in MCI by identifying characteristics of records that would indicate false merging and by closely examining those records. One such example entailed records that contained immunizations preceding one of the dates of birth (DOB) of the child, an indication that a false merge had potentially occurred (a child can have more than one DOB if records with different DOBs were merged together). Of the 1.6 million merges performed by MCI, a total of 397 such records were identified. A sample of 30 of these records was reviewed, and staff determined that 10 were, in fact, correctly merged. The remaining potentially false merges were the result of information that was unusual or wrong in the two original records. The majority of the falsely merged records belonged to siblings. The system has mechanisms to identify siblings and gives a lower "match" probability to those records. For example, when address, last name, or parent information are the same but first name and DOB are different, a "sibling" clue is activated that carries a very high "no match" weight that overrides all the similarities, and pulls the probability to a lower score. In the cases examined, however, although the DOB was different, the first names were the same or very similar, and additional fields that should have not been the same were the same, such as Medicaid number. To safeguard against this type of incorrect matching, a rule was implemented that sends to human review records matching with probability above 0.96 and whose birth dates differ by more than a year.

The system's sensitivity in production was also evaluated by assessing the percentage of known duplicates identified and merged. For the CIR, analyses consistently pointed to 30% as the pre–MCI duplication rate for the younger cohorts for which vital records and more complete immunization reports are available. These analyses were based on sampling and the known size of NYC's birth cohort (about 125,000 births per year). The MCI automatically merged 19% of records of children born in 1996–2003. Assuming that 50% of the 3% human review will also result in merging, then, after human review, 9.5% of the duplicate records are expected to remain in the system, pointing to a sensitivity of 90.5%.

● Discussion

CIR and LQ postdeployment issues

Since MCI's deployment, CIR and LQ Quality Assurance (QA) teams have been continually assessing the system to ensure all MCI services are working correctly. These services include: (1) MCI matching system, (2) MCI administration tool—the application used by CIR and LQ staff for searching the database and reviewing potential matches and for lookup code and table maintenance, and (3) MCI core services, which encapsulate the business rules developed to govern the system.

Since deployment, several issues have been identified that need to be improved in the MCI matching system. Although over 1.6 million records were automatically merged, thousands of potentially duplicate records from the "initial load" remain in the queue to be reviewed by humans. One of the immediate challenges of LQ and CIR is to study the human review queue to identify new clues the matching system can use to increase automatic merging of records, without compromising accuracy.

Further, staff members are examining the clues from the MCI matching system to determine if the weight assigned to each data field is appropriate. For example, the current matching system contains a clue that predicts "match" based on matching immunization dates in the two records. However, the weight assigned to this clue is too low, resulting in records sent to human review that should be automatically merged.

Once these enhancements are implemented, all records currently in the human review queue will be processed through the MCI matching system again. A significant proportion of them will merge automatically, thereby reducing the volume of records in the queue. It is also expected that the matching and merging rate of new, incoming records will increase as well.

The MCI administration tool, which is the application both CIR and LQ use to review patient records, was rigorously tested during the predeployment phase of the project. Further enhancements are being planned based on feedback received from users.

Because the record matching process has become more automated, both programs have to re-evaluate the number of staff needed for human review. Prior to the integration, up to 7 FTEs were allocated to manual deduplication by CIR, and 5 FTEs by LQ. Although exact estimates have yet to be determined, FTE staff devoted by the immunization and lead programs to human review will be reduced by an estimated 50%, due to the large percentage of records automatically merged. Furthermore, DOHMH staff needs to consider that human review carries the risk of false merging. Indications are that the matching system is more accurate in merging records than humans. Further analysis must be undertaken to examine this issue.

The MCI core services, the business rules governing the system, were created jointly by both programs. Modifications to these services are now necessary based on the needs of the client programs that were not foreseen during the development phase.

Provider access to data

As a result of the integration, immunizations and lead test histories of children are now available individually, and in aggregated form, to providers, MCOs, the NYS Medicaid Program, Women, Infants and Children program (WIC), schools, parents, and legal guardians or custodians through the CIR's existing access tools. It is believed that this will reduce missed opportunities to screen for lead and avoid unnecessary lead tests, and support outreach and follow-up by providers, MCOs, and agencies to children needing lead screening. Health outcomes research could test this assumption.

Of particular relevance is the ability for providers to look up both lead screening history and immunization history for their patients over the Web using the online registry. Also, MCOs will receive both immunization and lead screening information for their members via batch file data exchange to improve QA monitoring and outreach to children needing immunizations or lead tests. Providers, parents and legal guardians or custodians, foster care, and other agency staff who call the CIR for a child's immunization record will also be able to request the child's lead screening history. Currently, the CIR receives approximately 670 requests for records per month, and staff has been trained to disseminate lead information as well.

Next steps

NYC DOHMH has decided to leverage the MCI system for department-wide use. The MCI's main function is to facilitate matching, and it was developed to be extensible to other DOHMH databases because it includes attributes common to all patient databases.

The first adult system that will employ MCI is the Communicable Disease Surveillance System (CDSS). Because CDSS contains sensitive, highly confidential data, a crucial step in this process is developing additional privacy enhancements. These enhancements may also be leveraged for future client systems that have sensitive data, including the Sexually Transmitted Disease (STD) and HIV programs. Further, in order to achieve a maximum level of de-duplication, CDSS will tune the MCI matching system with its own data. A useful feature of the MCI is the ability for participating programs to use the existing record-matching system and/or to develop a more customized system tuned to their own data characteristics.

Conclusion

New York City now has a population-based, integrated child database that contains both immunization and lead test information. This database was created by merging of two distinct and separate data systems which, before December 2003, were functioning completely independently, had different operational procedures and technologies, and generally conducted "business" autonomously.

The principal challenge of this integration was to properly align records so that a child represented in one database is matched with the same child in the other database. To accomplish this task, as well as to identify internal duplicate records within each system, DOHMH employed an artificial intelligence record linkage system. This system takes into account all available data in each record and makes a human-like decision of "match," "no match," or "potential match" requiring human review.

Before integrating, the two databases had a combined total of 4.6 million records. The new, integrated system contains 3 million records, after more than 1.6 million records within and between the two systems were merged, at a very high level of accuracy. These results demonstrate that the AI technique can be used not only to reduce duplicate records in a database but also to successfully integrate patient records between two large health databases.

Because the MCI project successfully integrated CIR and LQ, senior DOHMH management has decided that the MCI will be expanded to include adults, as well as children, to create a department-wide record-matching system.

REFERENCES

1. Bureau of Vital Statistics, New York City Department of Health and Mental Hygiene. Summary of vital statistics 2002, the City of New York. Available at: www.nyc.gov/health. Accessed April 4, 2004.
2. Borthwick A, Papadouka V, Walker D. Principles and results of the NY Citywide Immunization Registry's MEDD De-Duplication Project. Paper presented at: The 34th National Immunization Conference; Washington, DC; July 7, 2000. Available at: http://www.choicemaker.com/nic_2000_presentation.pdf. Accessed June 6, 2004.
3. Choicemaker Technologies, Inc. Key concepts in the ChoiceMaker 2 matching system. Available at: http://www.choicemaker.com/key_concepts.pdf. Accessed April 14, 2004.
4. Berger A, Della Pietra S A, Della Pietra V J. A maximum entropy approach to natural language processing. *Computational Linguistics*. 1996;22:39–71.
5. Borthwick A. A Maximum Entropy Approach to Named Entity Recognition [PhD dissertation]. New York University, Department of Computer Science 1999.
6. Knuth D. *The Art of Computer Programming.* Reading, MA: Addison-Wesley; 1998.
7. Newcombe HG. *NYSIIS Algorithm Handbook of Record Linkage.* New York, NY: Oxford University Press; 1988.
8. Porter EH, Winkler WE. Approximate string comparison and its effect on an advanced record linkage system [Bureau of Census Web site]. Available at: http://www.census.gov/srd/papers/pdf/rr97–2.pdf. Accessed April 16, 2004.
9. Centers of Disease Control and Prevention. Deduplication test cases. Available at: http://www.cdc.gov/nip/registry/dedup/dedup.htm. Accessed September 23, 2003.