Issues in Identification and Linkage of Patient Records Across an Integrated Delivery System

Max G. Arellano, MA; Gerald I. Weber, PhD

To develop successfully an integrated delivery system (IDS), it is necessary to create a data file to identify and access efficiently information on patients, guarantors, and plan members. This data file is known as the *enterprise person index* (EPI) or the *master person index* (MPI). An accurate EPI provides the "glue" for the system to ensure that the relevant patient care information is truly available from all sites; minimizes multiple collection of basic demographic and patient history information; and allows for the computation of episode cost and outcomes research.

Only recently have systems administrators fully recognized the difficulty in building an accurate EPI. Without the presence of a single unique identifying field per patient, integration of existing patient medical data from multiple providers or sites depends on the use of basic patient demographic information. If that demographic information is complete and accurate and if there is little duplication of records within individual facility MPIs, the linking of data can, in theory, be handled with minimal error. Unfortunately, the reality of combining multiple MPIs into a single EPI is a much more formidable task because of the following three major factors:

- There are no accepted standards for either collecting or storing substantive information.
- There are no standards for default values, allowing registrars to assign any value for unknown data fields rather than the desired blank or empty field.
- Data for identifying critical fields are subject to error because of changes that occur over time, registrar error, or patient misinformation.

These three factors contribute to an existing rate of file duplication within single-facility MPIs at an expected range of 3 percent to 10 percent. That is, in a file of five hundred thousand patient names, data on anywhere from fifteen thousand to fifty thousand patients may be in two or more records with different medical record numbers.

During the past eight years we have been asked to analyze more than four hundred patient index files for internal duplication and for the presence of the same patient across multiple files. We have developed procedures for standardization of data and the assignment of measures of confidence (likelihood measures) including a probability-based measure. The preprocessing and postprocessing of the data are critical to assign the likelihood measures accurately and thus identify patient record linkages.

With the introduction of new health information systems to support the enterprise effort, vendors are becoming increasingly aware that duplication within a newly created EPI frustrates the basic objectives of their customers. Vendors have therefore applied algorithms, including various substring searches on name fields, in their attempts to identify linkages of patient records. We will evaluate several vendor algorithms and compare the results with those obtained using the more sophisticated probabilistic linkage methods that we recommend.

The Importance of Standardization Across Multiple MPI Files

The increasing need to merge disparate information systems has highlighted weaknesses in traditional methods of collecting and recording data. Different systems have stored data in various formats because of a lack of commonly accepted rules for recording data. To compound this problem, most vendors do not standardize MPI files before converting to an enterprise system. A set of standardization procedures is required before converting an individual MPI to ensure that items of identification that must be compared are indeed comparable. When hospitals that have formerly maintained independent MPIs consolidate their operations, these standardization procedures must prepare data for linkage because multiple files are analyzed rather than a single file.

Recommended Standardization Procedures. *Identify fictitious data records.* Extensive procedures must be developed to identify and delete fictitious data entries such as “test patient,” “John Doe,” or “dog” from the MPI. The procedures can use a combination of key word searches and frequency counts. Records identified as fictitious will be deleted, thus ensuring that only records of members of the patient population proceed to the linkage evaluation process.

Identify default values. Some data fields require that values be supplied even when the necessary information is not available, which introduces default values into the patient database. These default values are readily identifiable in many cases (for example, 999999 in the date field). However, the presence of 010190 is not so obvious and, moreover, may be a true value.

Because default values can lead to spurious linkages, they must be identified before linkage processing can be initiated. Most of these values can be identified using a frequency test. All obvious default values should be blanked out. They can be kept in another created field, if required for reporting.

Standardize personal identifiers. It is critical to the successful application of a record linkage procedure that key personal identifiers be compared. In a typical medical record file, the key items might include name, address, birth date, gender code, race code, and birthplace code. However, there are no generally accepted, commonly followed rules for recording those items, either in format or content, as can be seen in the following examples.

- *Multiple formats for name*. Both free-field and fixed-format forms are used. Different codes and default values are used such as “Baby,” “Girl,” “BG,” or “Boy (Ethyl),” to represent a newborn. Extraneous elements such as “Mr.” and “Dr.” can be placed within the files in an inconsistent pattern.
- *Social security numbers and phone numbers*. These fields are often formatted differently, sometimes with hyphens and sometimes without hyphens, thereby leading to different length fields.

Name Standardization Illustrated. The preprocessing procedure generally used to standardize names—the name parsing scheme—will illustrate the careful approach required.

Name parsing is the process of reading name fields; distinguishing between valid and invalid name forms; identifying the last, first, and middle names; separating out invalid values and extraneous name components; and then recording the name in a standardized form. Procedures vary depending on whether the name fields are specified in a fixed-field or free-field format.

Processing invalid values. Invalid values that appear in a primary name field generally indicate that a name had not been assigned when the record was generated. Examples of invalid values are “Unknown,” “Unavailable,” “Not Given,” “Baby Boy,” and “Baby Girl,” all words that are not actually a part of the formal name. All invalid values that appear in the first and last name fields of a file must be identified and blanked out because they will conflict with any valid names with which they are compared.

Processing extraneous name components. An extraneous name component is anything that appears in a name field other than invalid values that is not a proper part of the name. Titles such as “Dr.” and “Col.” and general suffixes such as “Jr.” and “Sr.” are examples. These should be moved from the name field and placed in another field for possible future use.

Processing last name suffix and first name prefix. Many first and last names may contain spaces between valid components, such as “Mary Lou” or “De Ford.” In most cases, these spaces are not used consistently from file to file or even from record to record. To make comparisons more accurate, the spaces should be removed from between these name components. Spaces could be added to unseparated components, but it is easier to concatenate than to separate when attempting to get comparability.

Encoding phonetic names. Using an appropriate name encoding system is crucial to the identification process. Transcription and recording errors in names are primarily handled by recoding names in a manner that the computer program sees as equivalent. The system used should be chosen based on its ability to form small groups of names that are variations of each other, so that when a name is misspelled it would have a high likelihood of being in the same group as the correct form of the name. We have found a good solution to be a phonetic encoding system based on the New York State Identification and Intelligence System (NYSIIS) algorithm developed in 1970 for the last name and an enhanced Soundex system for the phonetic encoding of the first name.

The phonetic encoder may be supplemented with other procedures such as string comparison methods. These methods can be designed to detect transcription errors that may be missed by phonetic encoding schemes. Extensive nickname tables can be developed and used to ensure that all nicknames receive the phonetic code of their formal first name. Tables can also be applied for those last names that seem to frustrate the most advanced phonetic systems.

Processing hyphenated names. When a hyphenated last name is encountered, it should be quadruplicated. One record will contain the original hyphenated name (Smith-Jones); one the reverse of the original hyphenated name (Jones-Smith); one the first part of the last name alone (Smith); and the fourth the last part of the hyphenated name alone (Jones). Each of the new records may make a match that the original record did not.

The same meticulous approach to edits and standardization is required for all fields used in the linkage process. For example, using tables of predominately male and female first names, gender codes can be assigned to records with unknown gender codes that can make allowances for instances of obvious gender coding errors. Other fields such as telephone and social security number should be stripped of extraneous elements such as hyphens.

Identification of Multiple Patient Records Across Multiple MPIs

Evaluating both single MPI file duplication, known as *internal dupes,* and multiple file or *overlap* matches is relevant to the creation of the EPI. Most EPI creation projects begin with at least one internal analysis for each MPI file. Based on the budget, staff, and time available, some subset of those duplicate files will be merged both on the computer system and in the paper files. Priority lists of the records to be merged first are generally based on the date of the most recent contact and the confidence in the linkage. Considerable evaluation by people is required to ensure that the patient data are correctly integrated.[1, 2]

Projects focusing on the creation of a multiple file EPI vary considerably from the internal duplication model.

Linkages Tend to Be Much Stronger in the Overlap Evaluations. Usually, between 70 percent and 80 percent of linkages in overlaps are exact matches on name, birth date, and gender or social security number after those fields have been edited and standardized. In contrast, internal duplication within a single MPI is usually related to discrepancies in the same primary identification fields, but with rarely more than 15 percent of linkages defined by an exact match.

Linkages Generally Cannot Be Reviewed in Overlap Evaluations. The integration of hospital outpatient and inpatient MPIs, the linkage of feeder facilities with major teaching hospitals, or the mergers of several facilities within a limited geographic service area can all lead to large numbers of linkages. Any internal linkage of more than thirty thousand matches is considered burdensome for the internal record merge process. Overlaps have produced up to several hundred thousand matches.

These numbers make it impossible to review the output from the linkage process, but some response is unavoidable. Most frequently, development of the required database pointer from the unique enterprise number to the individual MPI record numbers is automated based on assigned thresholds using the confidence measures developed in the linkage evaluation. A few organizations have decided to start fresh in their EPI and build it only as patients present. Others have converted only a limited number of years of records from the old system to the new ("backload").

Table 5.1 compares the typical differences between internal and multiple file evaluations. Using two individual MPI files from three multihospital systems and the overlap of patients across those files, we prepared the table showing the percentage of linked pairs with the same value in first and last name, birth date, social security number, and gender. The linkage uses a probability-based procedure. Matches in the name fields are based on standardized names,

Table 5.1. Comparison of Internal Master Person Index and Overlap Matches on Key Identification Fields (Percentage Matching of Total Linkages)

	First/Last Name	*Birth Date*	*Social Security Number*	*Gender*	*Exact Match*	*Total Linkages*
Project A						
Internal 1	38	73	15	55	15	53,677
Internal 2	75	91	41	97	33	35,227
Overlap	76	85	42	90	66	261,203
Project B						
Internal 1	41	67	-	88	16	8,256
Internal 2	41	60	34	94	13	19,952
Overlap	81	87	69	96	70	90,077
Project C						
Internal 1	44	78	-	95	25	12,311
Internal 2	57	78	19	94	37	1,057
Overlap	87	95	64	99	81	93,092

so they understate the true rate of error. We have included the proportion of exact matches on the combined fields as well.

Name and birth date discrepancies that are pervasive in the creation of duplicate records within a single MPI are a smaller factor in the overlap discrepancies, whereas social security number matches are much more frequent in the overlap analysis. As indicated earlier, the proportion of exact matches in the overlap example far exceeds that from the comparable internal analyses. The preceding framework allows us to compare alternative matching procedures—substring and probabilistic—for the development of the EPI.

Enterprise Load Matching Algorithms

During the enterprise load, or the creation and installation of the EPI, it must be determined whether each record evaluated either represents a person already in the file or requires the creation of a new index number. Usually, algorithms for identifying multiple records for a patient are applied as an enterprise load is in process. That process is analogous to a multiple MPI file overlap analysis.

If an incoming record represents a patient already in the enterprise file, a pointer is assigned from the existing enterprise number. When there is no match, a new member number and record are created in the enterprise database. The trade-off in this process is clear. A conservative approach to matching may lead to duplication of enterprise numbers for an individual even when there is little duplication in the existing individual MPIs. But the decision process must protect against having a single enterprise number pointing to more than one person.

Our probability-based approach to linking patient records helps us develop baselines for evaluating the effectiveness of several deterministic (exact match) algorithms. Our approach assigns an overall confidence measure based on the totality of the information available for linkage purposes, which is the sum of calculated levels of significance for each variable used.

The resulting algorithm recognizes that two characteristics of data files must be incorporated into the patient identification decision criteria: the frequency of occurrence of the values for each field used in personal identification, and the expected error rates for each of those data elements. Probabilistic patient identification provides a powerful tool because its linkage decision criteria are developed to emulate human judgment and skill in evaluating discrepancies—that is, the expertise of an experienced clerk. Probabilistic patient identification procedures use all the available identifying information to match multiple records accurately for the same person in spite of name variations and recording errors. The success of the search can be measured using a likelihood ratio, the probability of a successful search divided by the probability of an unsuccessful search.[3, 4]

A health information system vendor algorithm, in comparison, usually uses some type of exact match, a deterministic procedure, based on a combination of name, birth date, gender, and social security number. Substring comparisons (that is, those using only part of each name segment) are introduced to allow a degree of "fuzziness" in the process. We will evaluate three algorithmic procedures derived from the operational guidelines of several major health information system vendors. Based on matches for the data fields shown in Exhibit 5.1, each algorithm searches for instances in which two records represent the same person.

The variation in the three vendor algorithms highlights the complexity and the arbitrary underpinnings of exact match criteria. Social security number matches are required in two of the three algorithms. In one case, matches on blanks are included whereas another requires that the two social security numbers not be blank. When the social security numbers need not match, they cannot be discrepant (ignoring blanks). In two of the algorithms, complete last name must match, whereas in the third, the first five characters of the last name must match. Two of the algorithms require that the first six characters of the first name be used, whereas the third uses the first three characters.

Two critical questions are answered in our analysis of the algorithms: How does the total number of linked pairs found through the application of each deterministic algorithm compare with that found by application of the probabilistic linkage process? What are the discrepancies within each primary identification field associated with each algorithm's failure to identify the "very good" and "good" linkages?

Analysis 1. Table 5.2 summarizes the number of linkages identified using the three deterministic algorithms and compares them with the linked record pairs identified using the probabilistic process. The probability-based linkages are divided into two groups based on our experience with the quality of linkages—"very good" and "good." A third category, "high likelihood," a linkage that would require human review, is not shown.

Exhibit 5.1. Comparison of Three Vendor Algorithms

Vendor Algorithm A

- Social security number (not including blanks)
- First three characters of first name and first five characters of last name
- Gender

Vendor Algorithm B

- Complete last name and first three characters of first name
- Birth date
- Gender
- Social security number in either or both records can be blank

Vendor Algorithm C

- Social security number (including blanks)
- Complete last name
- First six characters of first name
- Gender
- Birth date

Table 5.2. Comparison of Personal Record Linkages Using Deterministic Algorithms or Probabilistic Procedure (Number of Matches Found by Overlap Project by Vendor Algorithm)

	Project 1	*Project 2*	*Project 3*
Deterministic (Exact Match) Linkages			
Vendor Algorithm A	95,836	53,918	55,019
Vendor Algorithm B	89,842	18,859	27,138
Vendor Algorithm C	108,181	46,575	59,128
Vendor Algorithms A and B	185,678	72,777	82,157
Probability-Based Linkages			
Very good	147,479	69,581	75,508
Good	78,103	17,603	15,748
Total	225,582	87,184	91,256

The combined ability of vendor algorithms A and B to identify the probability-based linkages ranges from 82 percent in Project 1 to 90 percent in Project 3, whereas the ability of algorithm C to identify the high likelihood linkages ranges from 48 percent in Project 1 to 64 percent in Project 3. A major factor in the variation among projects is the extent to which there are unknown social security numbers and gender codes. Those missing data were particularly problematic for all vendor algorithms in the Project 1 records.

The results of Project 1 highlight the potential for a questionable start to the EPI. Based on the known 225,582 very good and good linkages seen in the probabilistic model, the EPI evaluated with the deterministic model would start with 39,904 duplicates on day one if algorithms A and B together had been used, and with 117,401 duplicates if algorithm C had been used. A more detailed analysis reveals that a combination of algorithms A and B would have identified just 78 percent of the very good group and 68 percent of the good group, whereas algorithm C would identify 68 percent of the very good group and only 13 percent of the good group. As a side note, both algorithms A and C would have missed the entire set of linkages identified using the probabilistic process that require human review, and algorithm B would have identified only 3 percent of that group.

Analysis 2. We next look at the discrepancies among the individual data identification fields accounting for the variation in the algorithm's ability to link multiple person records. The combined algorithms A and B and algorithm C were analyzed. The statistics in Table 5.3 are based only on the very good and good categories of linkages identified with the probabilistic process, and show the percentage of linkages missed by each of the deterministic algorithms where the referenced field had two different values.

As the table shows, within Project 1 the birth dates differed in 25 percent of the cases where neither algorithm A nor B matched pairs of records that had been linked through the probabilistic identification process. The standardized

Table 5.3. Discrepancies Between Data Field Values in Possible Linkages (Percentage of Probabilistic Linkages Not Defined by Deterministic Algorithms)

	Algorithms A and B			*Algorithm C*		
	Project 1	*Project 2*	*Project 3*	*Project 1*	*Project 2*	*Project 3*
SSN missing				97%	41%	29%
SSN discrepancy	3%	11%	9%	6%	16%	12%
Birth date discrepancy	25%	33%	23%	12%	38%	16%
Standardized name discrepancy						
First name	39%	12%	23%	12%	12%	16%
Last name	25%	20%	38%	12%	10%	14%
First and last name	5%	11%	6%	2%	4%	2%
Gender discrepancy	25%	20%	14%	14%	7%	4%

Note: Percentages are based on the standardized name (without applying a phonetic).

first name differed in 39 percent of the same set of pairs linked through the probabilistic process.

Organizations have conducted "synchronization" projects (Table 5.3) whereby they make changes in the computerized MPI files before applying a vendor algorithm to ensure that existing overlap data are identified to the fullest extent at load time. Using the results from the more sophisticated probabilistic analysis, they make changes in the data identification fields for one or both of the linked person records. This can greatly improve the rates of identification for the vendor algorithms.

Summary and Conclusions

Historically, the health information systems community has viewed linking personal records as a mundane task. The oversimplified view that routine database manipulation can accurately identify multiple records for a single individual is erroneous, an assumption based on a misperception of the quality of the underlying data. Such data have been adversely affected by the evolution of individual facility patient indexes from multiple systems and the results of backload procedures, and the lack of focus on the need for data integrity by users of the automated systems. Much of the random, invalid data we identify on a daily basis is directly associated with the need for system users to place data in the patient record while they face the situation of having no obvious data field in which to place them. Combined with an underlying lack of standards for the collection of personal identification information, this results in pure chaos when reviewing an MPI file containing a million records at the start of a linkage evaluation project.

We have documented the considerable effort that must therefore be made in standardizing the MPI files using stringent analytical procedures and applying common edit routines before commencing record linkage. This preprocessing effort must then be supplemented with sophisticated matching procedures that can handle the dual challenge of minimizing false negatives (the failure to identify true linkages) and false positives (the incorrect linking of records that do not represent the same person).

The identification of pairs of linked records does not, however, complete an EPI loading. Because it is fairly common for a multiple facility linkage evaluation to identify more than two medical record numbers for the same patient, and the primary goal of an EPI is to assign a unique identifier for the patient which will link that patient's multiple files, it becomes necessary to develop a means of readily associating three or more records for the same patient.

One approach we have used with great success is to assign a common, sequential identification number to all linked medical record numbers for the same patient regardless of facility. The assignment of linkage identification numbers is computer-intensive and is generally accomplished with a highly iterative process. Both system memory and hard disk resources are fully tested as the number of good linkages in an overlap evaluation reaches the half-million mark or greater.

Because the primary linkage analysis goal is to develop linkages on pairs of records, with confidence levels based on the comparison of information for those two records, thresholds must be set to decide which linkages should be accepted as true without any human evaluation. If the threshold is set too low, the defined linkage groups may incorrectly join the medical record numbers for different persons. But if the threshold is set too high, there will be undesired duplication of persons in the enterprise system. As in the identification of the underlying linkage pairs, the development of a confidence measure greatly facilitates the assignment of the unique identification numbers needed in the EPI implementation.

References

1. Lenson, C. M. "Assuring an Accurate Enterprise Patient Index: The Difference Between Internal Duplicate Files and Overlap Patient Files." *J Am Health Info Manage Assoc,* 1997, *68,* 43–45.
2. Lenson, C. M., and Herr, L. M. "Preparing the Master Patient Index for an Integrated Delivery System." *J Am Health Info Manage Assoc,* 1995, *66,* 56–60.
3. Weber, G. I. "Achieving a Patient Unit Record Within Electronic Record Systems." *Proceedings: Toward an Electronic Patient Record '95.* Newton, Mass.: Medical Records Institute, 1995, *2,* 126–134.
4. Arellano, M. G., and Simborg, D. W. "A Probabilistic Approach to the Patient Identification Problem." *Proceedings of the Fifth Annual Symposium on Computer Applications in Medical Care.* Los Angeles: Institute of Electrical and Electronics Engineers, 1981, pp. 852–856.

About the Authors

Max G. Arellano, MA, is the chief scientist of Advanced Linkage Technologies of America, Inc. and has been working on the linkage problem as applied to patient identification since 1980.

Gerald I. Weber, PhD, is president of Advanced Linkage Technologies of America, Inc.