Achieving a Patient Unit Record Within Electronic Record Systems

Gerald I. Weber, Ph.D.
President
Advanced Linkage Technologies of America, Inc.

Originally published in *Proceedings: Toward an Electronic Patient Record '95*, pp. 126-134, March 14-19, 1995

BIOGRAPHY

Gerald I. Weber, Ph.D., is a principal and President of Advanced Linkage Technologies of America, Inc. (ALTA) in Berkeley, California. Dr. Weber joined ALTA in 1987 after more than twenty years experience as a health care economist and analyst focusing on issues related to medical manpower, physician and hospital payment, and managed care rate setting. Under his direction ALTA has focused its effort on applying its proprietary person record matching algorithms to the patient identification challenge within the health care sector.

ABSTRACT

Problem Definition

Assuring that there is one unit medical record number for each patient presents a difficult challenge to electronic record systems whether the focus is on a single patient index or multiple patient indexes.

1. In the former situation multiple (split) records for patients most often result from discrepancies among identifying fields such as name, birthdate and Social Security Number (SSN) within the registrar query and the base file during an on-line search. Typical split-records within a single file are portrayed by the following two examples.

MR#	Patient Name		Birthdate	SEX	RACE	SSN
127473	QUIRINO	MARTHA	10-19-954	F	1	886215060
533813	QUIRINO	MARTA	10-19-954	F		886214060
276440	GRANADOS	FRANK J	04-19-961	M	6	887738652
458773	GRANADOS	JOSE	04-19-962	M	1	887738652

2. When multiple patient indexes are linked there are a much smaller proportion of matched records with discrepancies among significant fields. Manual review becomes less feasible, however. Multiple automated systems are involved, the patient paper records will be stored in multiple locations, and generally there is a much larger total number of linkages.

Extent of Discrepancies

The best fields to use for patient identification are those that are generally available, change infrequently and can be provided with accuracy by the person furnishing the information. The nine fields listed below are those usually considered for identification purposes either in batch evaluations to identify split-records or for registrar on-line searches. In the latter case, the time required to input information is also critical to the decision on usage. For each field we have provided the percentage discrepancy among the fields <u>in linked record pairs likely to represent the same patient</u> (according to ALTA's confidence measures) within 20 single file evaluations with a total of over 6.5 million records.

Percentage Discrepancy in Single File Split-record Linked Records (Individual Percentages Weighted by Total Records in the File)

Field	Percent Discrepant
Last name	39
First name	22
Middle initial	16
Birth month	3
Birth day	7
Birth year	10
Social Security Number	36
Mother's Maiden Name	30
Telephone	39

These discrepancy rates are not likely to represent the error rates in the total MPI file as they were created from the set of split-records. Rather, they provide a guidepost of the errors that are leading to the creation of multiple records for a patient. The evidence is quite clear that discrepancies in the fields used in patient record searches are strongly associated with the creation of the multiple records. Since the last three fields in the list are present for both records in a small portion of identified pairs, their associated percentages must be interpreted with considerable caution.

Operation

The need for a sophisticated response to patient identification arises from the general inability of automated databases to define a primary key capable of uniquely identifying records for specific individuals. Because of the unreliability and/or unavailability of information, correct identification is problematic. Some of the most common obstacles to accurate patient identification are:

1. The absence of standards for the recording of names.
2. Misspelling within names.
3. Change in last name and hyphenated last names.
4. AKAs and nickname use for first name.
5. Variance in presented birthdates.
6. Transcription error in the Social Security Number and the use of one SSN by multiple family members.

Those problems are inherent within the patient registration process. Critical operational factors that further amplify the creation of split-records are listed below.

1. Decentralized registration areas with different policies and procedures.
2. Registration clerks who work on a computer system with inadequate training and use poor searching techniques. Inadequate audit trails have made the identification of problematic registrars difficult to identify.
3. Cultural factors may lead to inconsistent information such as the switching of given and middle names.
4. Pressure in the admitting department to decrease patient registration time.
5. Emergency registration of patients with information provided from secondary sources. The registrars in the Emergency Room are less trained and under intense time pressure.
6. Inadequate software for patient identification. Soundex searches tend to be weak; systems cannot accommodate AKA (also known as) searches; and there are not forced searches whereby registration staff is required to do a thorough check for a prior medical record number.

Minimizing False Positives and False Negatives - The Ultimate Challenge

Biometric identification procedures or the perfect patient identification card are not likely to be implemented in the near future. Thus the problematic environment described above will continue to require automated systems to locate patient records where there are discrepancies within the commonly used fields (e.g. fuzzy matches).

Where there are discrepancies between the identifying field information within two files that are to be linked, or between the identifying field information provided at time of registration with that in the Master Patient Index (MPI), potential for error must be dealt with.

Two types of error must be considered. The first category focuses on the possible exclusion of potential matches from consideration by the registrar or by the reviewer of linkages after evaluations of the entire MPI to identify existent split-records. This "false negative" problem is introduced by the inability of the search criteria used within the automated system to identify candidate records that match the query record information. If formal name is used in the query and a nickname is represented in the base file or a last name change has taken place, potential matches will be missed.

The second category focuses on the inclusion of inaccurate linkages for the consideration of the registrar or for evaluation by reviewers. This "false positive" problem is perhaps best illustrated when multiple pages of patient records are brought to the screen for a registrar to review. It is introduced by the failure of the automated system to discriminate finely enough when searching the MPI file for candidate records to evaluate.

Typically, automated systems will be biased either towards the handling of false negatives or the handling of false positives as the solutions used to alleviate one problem generally exacerbate the other. Most HIS vendors have responded to a perceived false negative problem by implementing "phonetic searches" which inflate the false positive problem.

Steps in Implementation of the Optimal Solution

Five procedures are critical to the concurrent optimal solution of the false negative and false positive problems.

1. Standardization

   This process ensures that items of identification which are to be compared are comparable. It assures that any field in the two records that is to be compared has the same field length; that extraneous elements in name field (Jr., Ph.D.) are deleted; that unknown name equivalents (Baby Boy, Twin1, unk.) are deleted; that invalid values in any field are not allowed (month = 14, SSN = 999999999); and that nicknames and hyphenated last names are not problematic.

   Look-up tables can provide efficient support for many of the standardization activities. They can be developed for the extraneous elements and unknown name equivalents, and can be used to equate formal and informal first names. Additional records can be created in the search process to respond to hyphenated last names (three records are created), to aliases and to possible errors in the coding of the sex field.

2. Phonetic Encoding

Phonetic encoding of names is critical to the complete identification of potential linkages. In most vendor systems a variant of the standard Soundex phonetic encoder is used. Evaluations of the Soundex systems, including that done by the New York State Identification and Intelligence System (NYSIIS) in 1970, have found problems with the reliability and selectivity of the algorithms. Examples of standard Soundex phonetic coding failures include:

1. Terminal S (Cobb, Cobbs)
2. Leading C or K (Cain, Kain)
3. Embedded G or DG (Rogers, Rodgers)
4. Embedded T or GHT (Leitman, Leightman)

In response, the NYSIIS staff developed their own phonetic encoder. Variants of that system are used by Statistics Canada, the U.S. Bureau of the Census, and the U.S. Department of Agriculture. As with most phonetic encoding systems, the NYSIIS performs particularly well for selected ethnic groups - Spanish and Southern Europeans. It reportedly has had difficulty with Japanese and French derivative names.

ALTA staff has introduced their own enhancements to NYSIIS. However, we are now convinced that the most effective approach to most remaining encoding problems is to introduce a last name look-up table that will force the equating of the phonetic for known problematic names. Adventist Health System - Loma Linda implemented this approach during installation of ALTA's SmartPID™ weighting system and is very pleased with the results.

3. The Search

Once the data files have been standardized and the phonetic codes developed, the first part of linkage processing is the identification of candidate records for automated and/or human evaluation. This search strategy determines what data elements in the two files are to be scanned in evaluating the existence of a match. A compromise is necessary between (a) forming all possible pairwise comparisons for the evaluation and (b) evaluating only the exact match of a unique identifier. Whereas scanning too many records increases the cost of the search including human review and response time, scanning too few records increases the likelihood of missing a matching record.

Within the healthcare environment in the United States the data element closest to representing a dependable, unique identifier is the Social Security Number. There are, however, three problems with its use.

1. Most of the MPI files ALTA has evaluated have had between 60 and 80 percent presence of SSN and, in aggregate, SSN has been available in only 40 percent of the records (excluding pediatric files where it is rarely

available). Moreover, presenting patients frequently will not be able or willing to provide their SSN.

2. There is frequent use of the same SSN by family members. This can be partially explained to be the result of confusion with requests for information on the guarantor.

3. Errors in transcribing SSN are prevalent. It is unfortunate that a check digit algorithm was not implemented as part of the SSN at the time of its introduction.

Despite these problems, use of the SSN (when presented) is very efficient. Indexing supports the almost instantaneous identification of candidate linkages and there are a relatively small number of false positives even with family use.

Use of Soundex searches (including last name) are most prevalent within current HIS products. Complete dependence on these searches presents two major roadblocks to accurate, efficient identification of candidate linkages.

1. The group of candidates identified within a phonetic block can be very large - frequently they present the human evaluator large numbers of false positive linkages.

2. Change in last name is the most prevalent discrepancy associated with split-records and those situations will not be identified with the phonetic search.

ALTA has found that the use of a combination of the last name phonetic and first name phonetic is effective and efficient **after the use of an alternative match key** which we typically specify as the first two characters of the first name phonetic, birthdate and sex code. This alternative key presents relatively small candidate groups and operates through an efficient indexed search.

As a generic guideline, our staff suggests that multiple search routines that identify relatively compact candidate groups be used in an iterative approach. Those routines that closely resemble a unique identifier should be applied first. Whatever search routines are applied must adequately respond to the magnitude of last name change prevalent in the patient population. Historically that change was primarily associated with female adults; today it is also associated with babies and young children.

4. <u>The Match</u>

Most typically, responsibility for the selection of the true match(es) from the candidate list is left to the human reviewer. This leads to the registrar searching through multiple screens trying to identify a correct medical record for the patient who has presented, or simply creating a new record in frustration. In the attempt to identify true split-records within thousands of pages of output provided from naive evaluations of the MPI, blurry-eyed clerks can be seen with yellow markers reviewing patient records sorted by last name, then birthdate, and then first name. This is the result of excessive false positive matches.

In order to support the human decision maker, the automated system should analyze the records identified by the search and display a limited number of records with a high probability of being a correct match. It should delete from consideration all obviously bad matches using, in addition to the search criteria, the fields that had not been used to identify candidate records. Using all the identification data, the system should then list the remaining linkages in declining order of the assigned level of confidence.

The most advanced algorithm for assigning such a confidence level was published by two Canadian statisticians in 1969.[1] That (Fellegi-Sunter) model is currently used by Statistics Canada, the U.S. Bureau of the Census, the U.S. Department of Agriculture and the SUR project of the National Cancer Institute.

The match strategy introduced by Fellegi-Sunter and enhanced by ALTA's Chief Scientist, Max Arellano, uses two factors for each patient identification field that is present.

1. The relative frequency of occurrence for the value in a matched field. (A match on Smith receives a lower weight than a match on Scitovsky; a match on birth month receives a lower weight than a match on birth day.)

2. The extent of discrepancy within the fields. (A miss on male last name receives a greater negative weight than a miss on female last name; a miss on telephone number receives a very small negative weight.)

5. <u>Review and Selection</u>

The most intelligent and able decision maker with respect to the identification of true linkages of patient records is the human. When provided a small number of alternatives and adequate data, they are able to effectively and efficiently process the data and make judgments. The role of the automated procedures described above is to facilitate the human process through the following.

1. Limiting the number of linkages that must be reviewed.
2. Limiting the records considered within each decision.
3. Assigning a confidence measure to each linkage to support the human decision.

Summary - Benefits and Costs

ALTA's experience with evaluations of MPIs from more than 40 organizations suggests that, on average, about 5 percent of medical records in a file will contain information for the same patient as another medical record. Those split-records are clinically problematic and associated with direct cost to the delivery system.

1. The quality of clinical care is compromised. Critical medical information such as drug allergies may not be available. One Medical Records Director has reported that her hospital has identified 25 incident reports monthly associated with split medical records.
2. The cost of additional search for information known by the attending physician to be in the patient records.
3. The potential for incorrect and/or incomplete data for billing, executive information and outcome analyses.
4. Potential failure to satisfy the JCAHO requirement to offer the capability to identify and retrieve a complete patient record.
5. Interference with the accurate implementation of the Computerized Patient Record, Corporate Patient Indexes and Community Health Information Networks where integrity and completeness of patient data is a primary objective.

Clean-Up of Existent Split-Records

Automated systems with a sophisticated approach to handling fuzzy matches have been shown to provide an efficient tool to identify existing split-records within the MPI while supplying a very small proportion of false linkages for review. The primary cost is related to the human effort to review and merge records within both the automated environment and paper files. ALTA's client hospitals have implemented procedures which allow for cost-effective response to identified split-records.

1. The dedicated merge effort is limited to those patients most likely to present for care. They typically are identified as patients who have experienced care within some recent period (two years is typical).
2. Review of identified linkages is concentrated on those matched records identified as most uncertain to be the same patient.
3. Staff or contracted clerks are dedicated to the merge task. Clients have indicated that productivity of clerks doubles within a month of dedicated activity. Our best estimate is that 3 to 6 split-records can be merged per hour with variance associated with the number of automated (departmental) systems requiring correction and the specific vendors involved, and the medical record storage environment (to what extent offsite).
4. Microcomputer support and customized reports are developed to facilitate the process.
5. An indicator flag is placed in the MPI for those split-records that are not merged. If the patient presents, a report is sent to Medical Records requesting that action be taken to evaluate and merge the records if appropriate.

Prevention of Split-Records

The majority of the split-records are related to the following factors.

1. Last name change.
2. First name nicknames.
3. Typos in the name fields.
4. Inaccurate birthdate information.
5. Transcription error in the Social Security Number.

Appropriate on-line search procedures can overcome those factors which are a natural part of the registration environment. Training of registration staff is certainly a critical element in reducing the introduction of duplicate records. However, the staff can also greatly benefit from enhanced automated search capability. The development of that procedure must take account of the intense conditions under which registrars operate. The cost of search procedures that are currently available appears to be exceeded by the present expense associated with ongoing review and correction of identified split-records. Moreover the evolution of healthcare delivery and information systems even further augments the benefits from improved search procedures and the presence of a patient unit record.

Conclusion

As an individual hospital links its diverse information systems the medical record number must provide the glue to tie data together. The single facility computerized record can only be a reality if the entire patient experience can be accurately accessed. Linking of files across facilities to create Corporate Patient Indexes and Community Health Information Networks will be greatly facilitated by MPIs that are relatively cleaned of split-records. As those files grow in size, reaching many millions of records, the requirement for efficient complete on-line search will be compulsory.

History has provided Health Information Management Departments a considerable burden in managing computerized data files that are problematic and replete with multiple records for patients. Advances in hardware and software technologies now offer the potential to greatly alleviate the existing problem, and to protect against future deterioration of that improved environment. The clinical and financial benefits would clearly seem to outweigh the investment required. A successful effort will require that sufficient budget be made available, that careful planning be undertaken; and that the effort be a cooperative, joint responsibility of Registration and Health Information Management staff with the full cooperation and support of the MIS Department.

FOOTNOTE

[1] Fellegi, I. and Sunter, A., "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1183-1210, December 1969.