Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields

by Beth Haenke Just, MBA, RHIA, FAHIMA; David Marc, MBS, CHDA; Megan Munns, RHIA; and Ryan Sandefer, MA, CPHIT

Abstract

Patient identification matching problems are a major contributor to data integrity issues within electronic health records. These issues impede the improvement of healthcare quality through health information exchange and care coordination, and contribute to deaths resulting from medical errors. Despite best practices in the area of patient access and medical record management to avoid duplicating patient records, duplicate records continue to be a significant problem in healthcare.

This study examined the underlying causes of duplicate records using a multisite data set of 398,939 patient records with confirmed duplicates and analyzed multiple reasons for data discrepancies between those record matches. The field that had the greatest proportion of mismatches (nondefault values) was the middle name, accounting for 58.30 percent of mismatches. The Social Security number was the second most frequent mismatch, occurring in 53.54 percent of the duplicate pairs. The majority of the mismatches in the name fields were the result of misspellings (53.14 percent in first name and 33.62 percent in last name) or swapped last name/first name, first name/middle name, or last name/middle name pairs.

The use of more sophisticated technologies is critical to improving patient matching. However, no amount of advanced technology or increased data capture will completely eliminate human errors. Thus, the establishment of policies and procedures (such as standard naming conventions or search routines) for front-end and back-end staff to follow is foundational for the overall data integrity process. Training staff on standard policies and procedures will result in fewer duplicates created on the front end and more accurate duplicate record matching and merging on the back end. Furthermore, monitoring, analyzing trends, and identifying errors that occur are proactive ways to identify data integrity issues.

Keywords: patient matching, patient identity, data integrity, master patient index (MPI), enterprise master patient index (EMPI), health information exchange (HIE), data discrepancy, data quality

Introduction

Deaths due to medical errors are the third leading cause of death in the United States, killing about 400,000 patients every year and more than 1,000 people each day, according to one estimate.[1] A different study estimated that 195,000 deaths occur each year because of medical errors, with 10 of 17 being the result of identity errors or "wrong patient errors."[2] These substantial problems are exacerbated as the healthcare system continues to consolidate and grow. The number of patient records in healthcare providers' databases is growing rapidly. Industry experts estimate that the average healthcare

organization's electronic health record (EHR) system has an 8 percent to 12 percent rate of duplicate records. A RAND Corporation report on duplicate records indicates that the average duplicate record rate for US healthcare organizations is 8 percent and is higher in large systems (15 to 16 percent).[3] If we translate these percentages into real numbers, a database with approximately 1 million records, or unique patients, may include up to 120,000 duplicate records. In terms of price, these duplicate records cost healthcare organizations about $96 per duplicate record, according to a research study conducted at Children's Medical Center Dallas.[4] The study also showed that in 4 percent of cases involving confirmed duplicate records, clinical care was negatively affected. Delays in initiating treatment in the emergency room were a common issue. Further, care quality issues included duplicated tests due to lack of access to previous test results and delays in surgery due to lack of patient history and physical reports. On average, repeated tests or treatment delays added $1,100 to the cost of the patient's care. Moreover, nearly 11 percent of duplicates are associated with bad debt.

EHR technologies have significantly improved patient safety and care coordination in recent years; however, some unforeseen difficulties and unintended consequences have arisen alongside these advances, including challenges associated with accurate patient matching. Despite these difficulties, the United States is experiencing an explosion in EHR adoption and use. According to 2013 estimates, the rate of certified EHR adoption among US acute-care hospitals was 94 percent, a 22 percent jump since 2011. Moreover, among acute-care hospitals, the rate of adoption of a basic EHR system increased from 9.4 percent in 2008 to 59.4 percent in 2013—a five-fold increase.[5] This rampant adoption, spurred by the Health Information Technology for Economic and Clinical Health (HITECH) Act, part of the American Recovery and Reinvestment Act of 2009, has increased master patient index (MPI) data conversion activities, which often exacerbate issues of duplicate patient records.

According to preliminary findings from the Care Connectivity Consortium, a health system using standard technology is able to identify only 10 percent of all matches.[6] To address this issue, among other identity issues, in September 2013 the Office of the National Coordinator for Health Information Technology (ONC) launched a patient matching initiative to help "improve patient matching across disparate systems."[7] ONC's final report, published in February 2014, outlined patient matching challenges and recommended changes and opportunities for the healthcare system to improve patient matching and, as a result, duplicate patient resolution. Recommendations were based on a national sampling of healthcare organizations, health information exchanges (HIEs), and EHR vendors. Their recommendations included establishing standardized patient identifying attributes, following existing data exchange standards, providing duplicate reporting capabilities, expanding best practices for patient matching, and developing and disseminating educational and training material regarding verification of accurate patient identity attributes.[8]

Patient matching provides the ability to match a unique individual with a unique set of data in a healthcare database or data set. In hospitals, this database is called the MPI or enterprise master patient index (EMPI). In the greater healthcare arena, the MPIs for large integrated delivery networks, health information organizations, accountable care organizations, HIEs, and others are more complex and have additional layers of identifiers. Within an organization, different facilities may have different medical record numbers (MRNs) for the same patient, each of which was assigned at a specific facility at which the patient was seen, such as an inpatient facility or a clinic. To add another layer, enterprise identification numbers (EIDs) link the various MRNs for one patient to a single umbrella record. Despite best practices in the area of patient access and medical record management as a method to avoid duplicating patient records, including scrubbing the MPI, using probabilistic and mathematical algorithmic tools to identify and select the correct patient record, adopting technology and educating stakeholders, and monitoring performance, duplicate records continue to be a major problem for healthcare. Reasons that duplicate records continue to plague healthcare systems include varying methods of matching patient records; departmental system silos; lack of data standardization; lack of policies, procedures, and data ownership; frequently changing demographic data; multiple required data points needed for record matching; and default and null values in key identifying fields.

Despite the extent of patient matching challenges, very little research has been conducted to date on the prevalence of duplicate medical records and the underlying causes of those duplicates. Just Associates Inc. worked in collaboration with the College of St. Scholastica to conduct research on confirmed duplicate records. The purpose of the study was to examine the underlying causes of duplicate records. The study's objectives were to provide categorization of data discrepancies on key patient-identifying data fields and to analyze the data for additional patterns that may lead to conclusions about what caused the duplicates to be created.

Methodology

The study was performed in 2014. A total of 398,939 confirmed duplicate pairs were analyzed from 10 different states, spanning more than 100 data cleanup projects. The data came from a variety of settings, including academic medical centers, community-based organizations, integrated delivery networks, a health information organization, and a children's medical center. The average size of the databases analyzed in this study was 1,813,201 unique patient records, with the median being closer to 1 million. A duplicate pair was defined as a single patient with two distinct records, and the study was limited to patients that had only two distinct records. Patients that had more than two records (multiples) were excluded from this study because of the complexity of examining records with more than one duplicate. The duplicate records were identified by a variety of MPI record-matching algorithms, with the majority of possible duplicates being identified by intermediate algorithms because the majority of cleanup projects that occurred during the time of the study used intermediate algorithms.[9] (See Appendix A.)

However, only confirmed duplicates were included in the study. Each algorithm identified duplicate records by referencing and comparing key demographic data fields. These fields included name, date of birth (DOB), gender, and Social Security number (SSN). Many intermediate and advanced algorithms also used telephone number and address in their comparison. The algorithms used in this analysis ranged from basic to advanced, with the majority being intermediate. All potential duplicate records identified by these various algorithms were evaluated and confirmed by patient identity experts, with the duplicate validity decision checked to ensure quality. The underlying causes of duplication were determined by examining mismatches between data elements commonly used in record-matching algorithms that existed in duplicate pairs. The following six data elements were examined for mismatches between duplicate pairs: first name (FN), middle name (MN), last name (LN), gender, SSN, and DOB.

The causes of mismatches between specific data elements in duplicate pairings were further analyzed. The frequency of the following underlying causes of mismatches in the six data elements described above were evaluated:

- Mismatches between any of the six fields that were a result of a blank entry in one record
- Mismatches in the MN field that were a result of the middle initial entered in one record and the complete middle name entered in the other record
- Mismatches in the SSN field that were a result of default values entered in one record
- Mismatches due to typographical errors in the FN, LN, or SSN field

In addition to identifying mismatches between duplicate pairings, additional analyses were conducted to identify swapping between the name fields and DOB subcomponents. Therefore, an analysis was conducted to identify duplicate pairs where values between the following fields were swapped: first and middle name, first and last name, middle and last name, month and day, month and year, and day and year. Five matching data fields were required for the duplicate pair (two records) to be validated as belonging to the same patient. The primary data fields included in the decision making process were LN, FN, MN, DOB, gender, SSN, telephone number, and address. Additional confirmatory fields such as alias name or guarantor may have also been used.

Statistical Methods

The R statistical software package was used to analyze this data set. The frequency of discrepancies between data elements of duplicate pairs was determined by developing IF-THEN statements and counting the frequency of each output. For example, to determine if there were discrepancies in the FN of two records of a duplicate pair, the following statement was developed: "*IF* the first name in record 1 matches first name in record 2 THEN label as TRUE *ELSE* label as FALSE." The IF-THEN statements resulted in the Boolean output of TRUE or FALSE. The frequencies of TRUE and FALSE outputs for each data field were calculated to determine the overall occurrence of discrepancies. Also, the proportions of FALSE occurrences were calculated to determine the percentage of duplicate pairs where there were discrepancies in a specific data field.

FOR-LOOP statements were used to identify character discrepancies in the FN, LN, and SSN. The FN, LN, and SSN of duplicate pairs were aligned when the data elements had the same number of characters. Figure 1 depicts the method that was adopted to determine the number of character discrepancies between data elements of duplicate pairs. In Figure 1, "Sandefer" was the LN in one record while "Sandefre" was the LN entered in the other record of a duplicate pair. The two records were aligned and positional mismatches were identified. The number of character discrepancies was determined by identifying the number of positional letter differences between the two records. In this example, there are two letter discrepancies. Green arrows indicate position matches between the characters, and red arrows indicate discrepancies. Again, the R statistical software package was utilized to make this comparison.

To determine the frequency of duplicate pairs that were the result of typographical errors, the frequencies of positional character discrepancies were calculated for the FN, LN, and SSN of duplicate pairs where the number of characters in that data field was the same in both records. Typographical errors were equated by aligning the characters of the FN, LN, and SSN in the duplicate pairs and counting the number of character discrepancies between these fields. Typographical errors were defined as instances in which a duplicate pair had one or two characters that differed between the FN, LN, or SSN. A one-character discrepancy was likely due to a single typographical error. A two-character discrepancy was likely due to transposition, such as "Michael" entered in one record and "Micheal" entered in the duplicate record.

Further analytical breakdown was performed on pairs of records in which a single data field was discrepant across the pair and on pairs of records in which only two data fields had discrepancies. These more detailed analyses identified more specifically the types of discrepancies that occurred, such as misspellings, data entry errors (transpositions or typographical errors), or complete differences in the data values (such as "Johnson" as the LN on one record and "Smith" on its duplicate record).

As in any research study, it is important to note the limitations of the method. Some margin of error may have occurred during the analysis, both in the way the statistical code was run and in the validity decisions made by patient identity experts. In addition, comparisons were made at the end of the analysis to the results of an unpublished 2007 Just Associates study in which the statistics were calculated in a similar but not identical manner.[10] Unfortunately, very little research has been conducted to date on this topic and similar topics, which made it difficult to perform meta-analysis and compare these results to those of previous studies. In addition, the majority of duplicate records were identified via an intermediate algorithm. If an advanced algorithm had been applied, more data discrepancies would have been identified and discrepancy rates would have been higher, which could have resulted in significantly different results of the study. (Refer to Appendix A for more information on algorithms.) Lastly, in the analysis, three or more character or digit differences in a name or SSN field led to the mismatches being defined as different records when in fact they could have potentially been records of the same person.

Results

A variety of counts and percentages were calculated to determine the most prevalent types of data discrepancies. Table 1 depicts the most prevalent single and combination data discrepancies identified in the study.

The MN was the data element that had the greatest proportion of discrepancies between duplicate pairs. It was discrepant in 58.1 percent (231,566) of the duplicate pairs. The SSN was discrepant in 53.4 percent (213,003) of the duplicate pairs. FN discrepancies occurred in 22.6 percent (90,035) of the duplicate pairs, and LN discrepancies occurred in 23.7 percent (94,661) of the duplicate pairs. DOB discrepancies occurred in 6.2 percent (24,781) of the duplicate pairs. Discrepancies in gender occurred the least frequently, accounting for 6.0 percent (34,014) of duplicate pairs (see Table 2).

When limiting the analysis to just the duplicate pairs that had discrepancies in the MN, 62.0 percent (143,504) of the MN discrepancies were due to a blank entry in one record of a duplicate pair. Additionally, 20.3 percent (47,097) of the MN discrepancies were a result of the middle initial entered in one record and the full MN entered in the other record. The FN was swapped with the MN in 1.6 percent (3,595) of the duplicate pairs that had discrepancies in the MN (see Table 2).

A blank entry accounted for 58.2 percent (123,894) of the SSN discrepancies. Discrepancies in the SSN as a result of a default value (e.g., 111-11-1111, 999-99-9999, 123-45-6789) accounted for 50.2 percent (106,965) of all SSN discrepancies. Duplicate pairs with SSN discrepancies often had a blank entry for one record and a default value for the other record (see Table 2).

Of the duplicate pairs with discrepancies in the FN, 31.3 percent (28,174) were caused by a blank entry. The LN was swapped with the FN in 7.8 percent (6,978) of duplicate pairs with FN discrepancies, while the MN was swapped with the FN in 4.0 percent (3,595) of the duplicate pairs (see Table 2).

A blank entry accounted for only 5.1 percent (4,804) of the LN discrepancies. The FN was swapped with the LN in 7.4 percent (6,978) of the duplicate pairs with LN discrepancies, while the MN was swapped with the LN in 2.0 percent (1,933) of the duplicate pairs (see Table 2).

A blank entry accounted for 16.5 percent (3,972) of the gender discrepancies. Other discrepancies in gender were a result of a discrepant gender being entered (see Table 2).

Of the duplicate pairs with discrepancies in the DOB, only 0.4 percent (103) were due to a blank entry. The most frequent DOB discrepancy was with the day component, which differed in 48.3 percent (11,977) of the pairs with DOB discrepancies. The month differed in 22.1 percent (5,486) of these pairs, and the year differed in 40.0 percent (9,909) of the pairs. Swapped month and day accounted for 2.6 percent (635) of the duplicate pairs with discrepancies in the DOB. No duplicate pairs had a swapped month and year or a swapped year and day (see Table 2).

As shown in Table 3, typographical errors that included misspellings and transpositions (as determined by the count of duplicate pairs with one-character and two-character discrepancies) in the FN accounted for 53.1 percent of duplicate pairs with discrepancies in the FN. Typographical errors in the LN accounted for 33.6 percent of duplicate pairs with discrepancies in the LN.

FN and LN discrepancies were further analyzed to categorize the discrepancies as typographical errors or misspellings rather than a different name. Table 3 portrays the frequency of the number of letter discrepancies between the FN (n = 18,503) and LN (n = 77,912) fields of duplicate pairs when the fields had the same number of letters. The percentage is calculated as the count of letter discrepancies divided by the total number of discrepancies for that field. Three or more discrepancies in a name field was defined as being a “different” name.

Finally, typographical errors in the SSN accounted for 7.8 percent of duplicate pairs with discrepancies in the SSN (Table 4). More than 92 percent of duplicate pairs with SSN discrepancies occurred with three or more digits being discrepant, indicating that the discrepancy is not likely to be the result of a typographical error. Table 4 shows the frequency of the number of digit discrepancies between aligned SSNs of duplicate pairs (n = 87,473). The percentage is calculated as the count of digit discrepancies divided by the total number of discrepancies for that field. Three or more discrepancies in a SSN field was defined as being a “different” SSN.

Figure 2 shows the frequency of duplicate pairs in which one of the key demographic data points was discrepant. The highest percentage of single-field discrepancy was in the MN, accounting for 58.30 percent of the duplicate pairs. The second most prevalent discrepancy was SSN, with 53.54 percent. The

LN was the only discrepancy in 23.84 percent of the pairs, while the FN was discrepant in 22.72 percent. The DOB was discrepant in 6.30 percent of the pairs, and gender was discrepant in 6.32 percent.

Finally, Table 5 shows the frequency of duplicate pairs with co-occurring discrepancies where two or more of the data fields were discrepant. For example, a duplicate pair of records for a patient may have a discrepancy involving the MN and FN fields but also have a discrepancy in the SSN field. This pair of records would be counted in the table in the FN/MN cell, the FN/SSN cell, and the MN/SSN cell. Again, the highest co-occurring discrepancies are due to issues with the SSN, FN, MN, and LN fields.

Discussion

As noted above, there are numerous reasons for duplicate records. This study focused on data capture and standardization with the goal of ascertaining the underlying cause of various types of data discrepancies. Further explanation of data-related reasons for duplicate records is provided below:

- **Lack of data standardization.** Demographic field names vary across the healthcare system databases, and the fields typically do not have a one-to-one match. For example, one system might have three different name fields: first name (FN), middle name (MN), and last name (LN). Another system may have only one field for the name, allowing the user to enter the patient name in any order: Megan L Munns; Munns, Megan L; Megan Munns; Munns, Megan; and so on. Matching these fields is not an easy task, and name suffixes compound this problem.
- **Frequently changing demographic data.** In today's society, people commonly change their name, addresses, and phone numbers and occasionally change their gender. Data collection also can be a problem. For instance, the MN and SSN frequently are not collected. Human error, policy gaps, and lack of standardization all contribute to data inaccuracy. In individual facilities and in the healthcare industry as a whole, there is a lack of consistency in the data collected.
- **Required multiple demographic data points.** To effectively match one patient with another, several data points are necessary. If records do not have enough matching data points and patient matching is based on limited data, there is a risk of overlaid records. An overlaid record occurs when two different patients are associated with the same record.
- **Default and null values in key identifying fields.** Having default values or no information in fields such as SSN and MN can reduce an organization's ability to accurately match patient records.

Overall, this study demonstrates that duplicate patient record discrepancies are often due to a blank entry or a default entry in one of the key identifying fields, with the majority being in MN and SSN fields. Because of the complex nature of record matching and the decreased capture of an accurate, valid SSN, the MN is becoming ever more important to appropriately identify duplicate records. Additionally, name mismatches arise repeatedly as a result of misspellings or transpositions, accounting for 53.14 percent of all FN mismatches and 33.62 percent of all LN mismatches. Yet this analysis demonstrates that the MN and SSN were discrepant in 21.88 percent of all discrepant records, which was the highest of any multifield discrepancy (see Table 1). Furthermore, the record-matching algorithms in use today catch duplicate records that have exact or similar data in key identifying fields, but the percentage of records that match exactly is less than 5 percent of the discrepant records. Health information management departments appear to be doing a better job of cleaning up the duplicates that match exactly in these key identifying fields and/or registration areas are searching the MPI for patients before creating a new record. Of further importance is the number of gender discrepancies, which is now slightly higher than DOB discrepancies. Possible explanations for the number of gender discrepancies could be the increased number of transgender patients and/or reference labs' creation of shell records for lab patients without capturing data in this field. However, most of the discrepancies identified consisted of elementary errors

identified by basic and intermediate algorithms. If advanced algorithms had been used to determine potential matching records, these discrepancy patterns might have been significantly different.

Comparing this study's results with those of the unpublished 2007 Just Associates study reveals some drastic differences (see Table 6 and Table 7). In the current study, only about 5 percent of the duplicates had no discrepancies in the six key identifying fields, whereas in 2007, nearly 18 percent had no data discrepancies in these fields. Additionally, in the current study about 59 percent of the duplicate records had two or more discrepancies in the key demographic data, compared with 48 percent in 2007. This finding suggests that front-end registration processes are creating fewer exact-match duplicates, the exact-match duplicates created are being actively fixed and therefore do not appear in this data set, or both. Moreover, discrepancies in DOB fields have decreased substantially over the years, from 14.9 percent in 2007 to 6.2 percent in the current study. This improvement in data capture could be due in part to the patient safety rules pertaining to accurate patient identification. Decreased capture of the SSN, which often is the only unique identifier for a person, and decreased accuracy in name collection calls for the industry to pay closer attention to data capture accuracy and the need for advanced algorithms that tolerate multiple data discrepancies. Because the current study used a different statistical program to compare the types of data discrepancies, not all of the changes identified above can be extrapolated to indicate a change in data capture practices in the industry. However, the findings described in the first three sentences of this paragraph are valid comparisons because they were computed using exact data match programs in both studies.

So what does this information mean, and what can be done to address the issues presented by duplicate patient records? Initial steps that can be taken to lessen the burden of duplicate records include partnering with colleagues in patient access to establish standard policies and procedures, such as patient searching protocols, standard name entry conventions, and questions that registrars can ask the patient in order to determine if the patient has ever been to the facility or practice before. Capturing the full MN would help substantially to verify the patient's identity and is particularly useful to distinguish between twins and patients with common names. Additionally, creating a searchable field to store the last four digits of the patient's SSN (separate from the full SSN field) would greatly improve duplicate record prevention. Historically, patients have been hesitant to share their full SSN, and rightfully so with the rise in identity thefts. Because of this, other industries such as banking and financial institutions have begun to capture the last four digits of the SSN, which is useful in identity verification and gives customers ease of mind knowing that they are not sharing their full SSN. This technique is commonly utilized in the finance industry, and patients would be less hesitant to share the last four digits than they would be to share the full SSN. This technique is not common in healthcare; however, some EHR vendors have added a separate data field for the last four digits of the SSN, and at least one large HIE is using this field in its patient matching algorithm.

In addition, staff should be trained and frequently assessed on their knowledge regarding identity data capture and the consequences of inaccurate or incomplete data capture on duplicate record creation. It is also imperative to track who is creating duplicate records and where those employees work. Tracking this information can identify a need for additional training or changes to policy and procedures. Some EHRs also offer the ability to track demographic changes to records. Tracking these reports and quality-checking demographic changes for appropriateness may also trigger further training for staff, both administrative and clinical. In addition, organizations should take steps to evaluate the underlying causes of duplicate patient records, including an assessment of common data discrepancies, evaluation of people and process augmentation, technology solutions, and ongoing MPI maintenance.

The ONC's Patient Matching Initiative report clearly states that people and process improvements are needed to improve patient identity data integrity.[11] Technology is needed as well because errors will always occur, and as databases get larger and larger, human manpower alone cannot solve this issue. Effective tools to fix errors that occur are needed. Utilizing more advanced record-matching algorithms, along with other technology solutions such as smart cards and biometrics, is key. Advanced algorithms tolerate both multiple data entry errors and changes in a patient's demographic data. The higher-quality algorithms are able to limit the number of false positives that show up in a standard duplicate queue and catch more real duplicates than are caught by basic or intermediate algorithms, minimizing false

negatives.[12] Algorithms should be continuously improved to review additional key fields for patient matching. Currently, most algorithms utilize only standard demographics, such as name, DOB, gender, SSN, and in some cases address and phone number. The use of additional fields can help to identify more real duplicates. Some of these additional fields include previous names (to help detect overlays, as well as to identify an existing record for that patient), insurance data, policy numbers, guarantor data, and next of kin. Furthermore, front-end search algorithms could be improved by allowing searches using other demographic fields (e.g., telephone number or the last four digits of the SSN) and by installing biometric technologies or card readers.

Conclusion

To improve patient matching, increasing the use of more sophisticated technologies is critical. For example, using biometrics, smart card readers, and advanced algorithms (or fine-tuned algorithm matching criteria) can all help increase front-end and back-end matching. Record search algorithms in scheduling and registration systems specifically are in need of major enhancements. For example, adding last four digits of a SSN as a searchable field and adding complementary fields in the matching criteria, such as telephone number, next of kin, guarantor, or insured, would find more potential duplicates and could accommodate more data discrepancies. Apart from technological advances that can be implemented, additional improvements can be made to the current state of demographic data capture. As demonstrated in Table 6 and Table 7, MPI data discrepancies overall appear to be increasing, as more duplicate records now have multiple data discrepancies than in the 2007 study. As seen in the results of this study, the SSN and full MN had the lowest percentages of being a reliable demographic data element. Increasing data capture in these fields can also significantly increase matching possibilities.

No amount of advanced technologies or increased data capture will completely eliminate human errors. Creating policies and procedures for front-end and back-end staff to follow is foundational for the overall data integrity process. Training staff on standard policies and procedures will result in fewer duplicates created on the front end and more accurate duplicate record matching and merging on the back end. Furthermore, proactive ways to identify data integrity issues include monitoring, analyzing trends, and identifying errors that occur. The sooner issues are detected, the sooner health information management and registration leaders can address the concern.

Implications of ineffective patient matching are substantial because the MPI is the backbone of the EHR. Severe patient-care issues can occur and resources are wasted when systems are inundated with duplicate records. Patient safety is a major concern for many organizations, yet it is necessary to increase awareness of the safety, legal, financial, and compliance concerns created by duplicate and overlaid medical records. Additionally, the migration to value-based reimbursement may be the driver for healthcare organizations to ensure that their MPI data are clean. If an organization's data are not sufficiently clean, risk-based contracting will create significant financial consequences that are likely to capture the attention of C-suite executives. Lastly, a primary goal of the EHR Incentive Program, with its criteria for meaningful use of EHRs, is to "improve quality, safety, efficiency, and reduce health disparities."[13] Several of the meaningful use criteria have the number of "unique patients" as a denominator. Duplicate records increase the number of "unique patients," thereby making it more challenging for providers to reach the minimum core criteria.

This research study was limited by the exclusion of multiples (defined as the existence of more than two records for a patient), the fact that the majority of cases were identified by an intermediate algorithm, and the lack of other published research studies to use as a basis for comparison. Further research is needed to demonstrate the improvement in patient matching when advanced technologies are in place, better data capture of other demographic identifiers such as full MN or last four digits of the SSN is achieved, and multiple duplicate record sets are included in the analysis. Solid data demonstrating the effectiveness of these possible solutions are needed to further educate healthcare leaders and health information management professionals on patient matching challenges and opportunities for improvement.

Beth Haenke Just, MBA, RHIA, FAHIMA, is the founder and CEO of Just Associates in Centennial, CO.

David Marc, MBS, CHDA, is an assistant professor and director of the health informatics graduate program at the College of St. Scholastica in Duluth, MN.

Megan Munns, RHIA, is an associate identity manager at Just Associates in Centennial, CO.

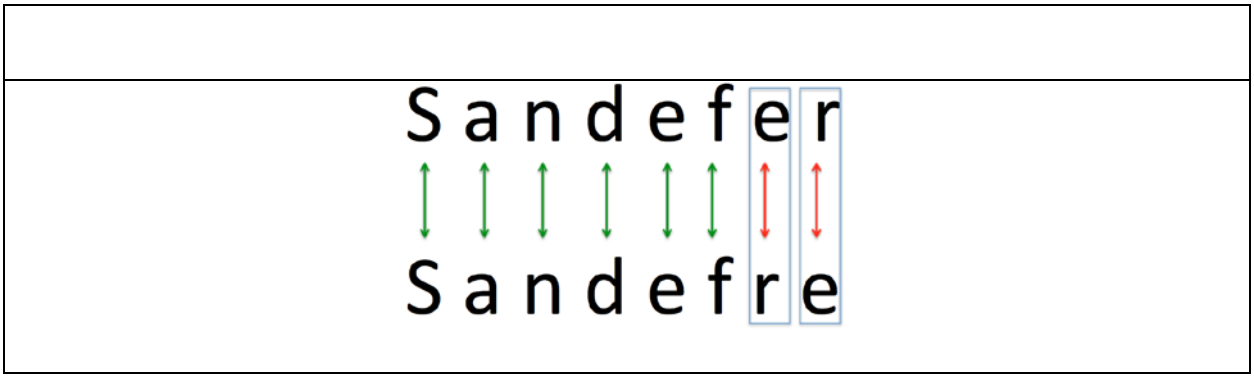
Ryan Sandefer, MA, CPHIT, is an assistant chair and professor in the Department of Health Informatics and Information Management at the College of St. Scholastica in Duluth, MN.

Notes

1. McCann, Erin. "Deaths by Medical Mistakes Hit Records." *Healthcare IT News*, July 18, 2014. Available at http://www.healthcareitnews.com/news/deaths-by-medical-mistakes-hit-records.
2. Smart Card Alliance. *Effective Healthcare Identity Management: A Necessary First Step for Improving U.S. Healthcare Information Systems*. 2014. Available at http://www.smartcardalliance.org/resources/pdf/Healthcare_Identity_Brief.pdf.
3. Hillestad, Richard, James H. Bigelow, Basit Chaudhry, Paul Dreyer, Michael D. Greenberg, Robin C. Meili, M. Susan Ridgely, Jeff Rothenberg, and Roger Taylor. *Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System*. Santa Monica, CA: RAND Corp., 2008. Available at http://www.rand.org/pubs/monographs/MG753.html.
4. Just Associates. "Studies in Success: Children's Medical Center Dallas." Available at http://www.justassociates.com/Articles-MPI-Childrens-Medical-Center-Dallas.html.
5. Charles, Dustin, Meghan Gabriel, and Michael F. Furukawa. *Adoption of Electronic Health Record Systems among U.S. Non-federal Acute Care Hospitals: 2008–2013* (ONC Data Brief No. 16). Office of the National Coordinator for Health Information Technology, May 2014, p. 1. Available at http://www.healthit.gov/sites/default/files/oncdatabrief16.pdf.
6. Heflin, Eric, Sid Thornton, and Reg Smith. "An Approach to Understanding and Resolving Inter-organizational Patient Matching." Presented at the Care Connectivity Consortium, February 24, 2014, slide 7.
7. Stevens, Lee. "ONC Launches Patient Matching Initiative." *Health IT Buzz*, September 11, 2013. Available at http://www.healthit.gov/buzz-blog/health-innovation/onc-launches-patient-matching-initiative/.
8. Morris, Genevieve, Greg Farnum, Scott Afzal, Carol Robinson, Jan Greene, and Chris Coughlin. *Patient Identification and Matching Final Report*. Prepared for the Office of the National Coordinator for Health Information Technology. February 7, 2014. Available at http://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf.
9. AHIMA. "Managing the Integrity of Patient Identity in Health Information Exchange (Updated)." *Journal of AHIMA* 85, no.5 (May 2014): 60–65. Available at http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_050658.hcsp?dDocName=bok1_050658.
10. Heflin, Eric, Sid Thornton, and Reg Smith. "An Approach to Understanding and Resolving Inter-organizational Patient Matching."
11. Morris, Genevieve, Greg Farnum, Scott Afzal, Carol Robinson, Jan Greene, and Chris Coughlin. *Patient Identification and Matching Final Report*. Prepared for the Office of the National Coordinator for Health Information Technology.
12. Just Associates. "Studies in Success: Epic with IDOptimize." Available at http://www.justassociates.com/Articles-Epic-with-IDOptimize.html.
13. HealthIT.gov. "Meaningful Use Definition & Objectives." Available at http://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives.

Figure 1

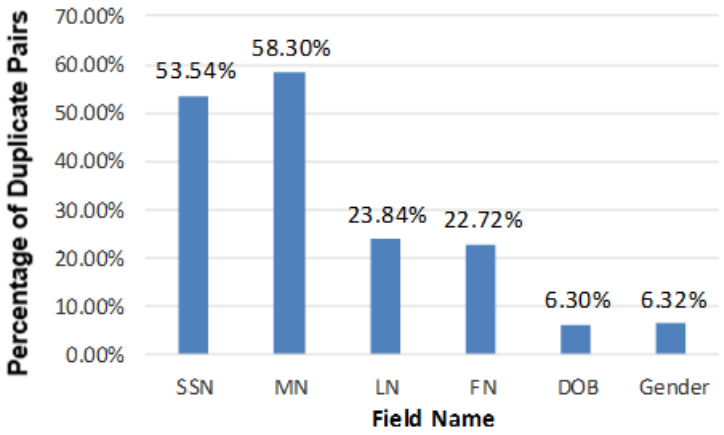
Example of Character Matching Method Used to Compare Name Fields of Duplicate Records



Note: Green arrows indicate position matches between the characters. Red arrows indicate discrepancies.

Figure 2

Percentage of Duplicate Pairs with a Mismatch in the SSN, MN, LN, FN, DOB, or Gender Field (*n* = 398,939)



Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Table 1

Percentage of Discrepant Pairs by Discrepancy Category

Discrepancy Type	**Duplicate Pair Counts by Discrepancy Category**	**Percentage of Total Duplicates (*n* = 398,939)**
MN	29,396	7.37%
SSN	75,115	18.83%
MN + SSN	87,304	21.88%
LN	21,362	5.35%
LN + SSN	7,158	1.79%
LN + MN + SSN	7,102	1.78%
LN + MN	25,510	6.39%
SSN + DOB	3,097	0.78%
FN + MN	30,753	7.71%
MN + SSN + DOB	2,764	0.69%
FN	10,148	2.54%
DOB	5,925	1.49%
FN + MN + SSN	6,573	1.65%
FN + SSN	6,636	1.66%
MN + DOB	6,025	1.51%
LN + FN + MN	14,175	3.55%
LN + FN	4,787	1.20%
Gender	854	0.21%
Other	34,343	8.61%
Total without discrepancies	**19,912**	**4.99%**
Total with discrepancies	**379,027**	**95.01%**

Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Table 2

Proportion of Mismatches between Data Discrepancy Subtypes

Data Element	Percentage of Total Duplications (*n* = 398,939)	Percentage of Duplicate Pairs with Discrepancy by Data Element
First Name	22.6%	FN–MN swap: 4.0% FN–LN swap: 7.8% FN blank in one record: 31.3%
Middle Name	58.1%	FN–MN swap: 1.6% LN–MN swap: 0.8% MN blank in one record: 62.0% Middle initial in one record and full MN in the other record: 20.3%
Last Name	23.7%	FN–LN swap: 7.4% LN–MN swap: 2.0% LN blank in one record: 5.1%
Gender	6.0%	Gender blank in one record: 16.5%
SSN	53.4%	Default SSN: 50.2% SSN blank in one record: 58.2%
DOB	6.2%	DOB blank in one record: 0.4% Month does not match: 22.1% Day does not match: 48.3% Year does not match: 40.0% Month and day swapped: 2.6% Month and year swapped: 0% Day and year swapped: 0%

Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Table 3

Frequency of Letter Discrepancies in First Name and Last Name

	Number of Discrepancies							
Field	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**>8**
First Name (*n* = 18,503)								
Count	6,753	3,078	1,394	2,026	2,186	1,948	725	393
Percentage	36.50	16.64	7.53	10.95	11.81	10.53	3.92	2.12
Last Name (*n* = 77,912)								
Count	17,771	8,419	9,071	8,464	6,225	10,468	12,815	4,679
Percentage	22.81	10.81	11.64	10.86	7.99	13.44	16.45	6.01

Table 4

Frequency of Digit Discrepancies in Social Security Number (*n* = 87,473)

	Number of Discrepancies								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Count	4,712	2,087	3,837	794	1,842	7,058	17,845	24,874	24,424
Percentage	5.4	2.4	4.9	0.9	2.1	8.1	20.4	28.4	27.9

Table 5

Percentages of Co-occurring Mismatched Fields

	FN	**MN**	**LN**	**Gender**	**SSN**	**DOB**
FN	-	**15.3**	**6.8**	2.1	5.4	0.8
MN	**15.3**	-	**13.6**	4.2	**28.7**	3.1
LN	**6.8**	**13.6**	-	1.4	5.4	0.9
Gender	2.1	4.2	1.4	-	2.6	0.4
SSN	5.4	**28.7**	5.4	2.6	-	1.8
DOB	0.8	3.1	0.9	0.4	1.8	-

Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Table 6

Comparisons of the Causes of Duplicate Patient Records in an Unpublished 2007 Study by Just Associates and the Current Study

Type of Discrepancy	Percentage of Total Pairs (*n* = 244,356) in the 2007 Study	Percentage of Total Pairs (*n* = 398,939) in the Current (2014) Study	Change from 2007 to 2014
DOB discrepancies	14.9	6.2	Decreased 58%
FN–LN swap	<0.1	7.8	Increased 7,700%
FN–MN swap	0.7	4.0	Increased 471%
Gender discrepancies	3.0	6.0	Increased 100%
SSN discrepancies	44.4	53.4	Increased 20%
Two or more data discrepancies	48.2	59.2	Increased 23%
No data discrepancies across six key identifying fields	17.6	5.0	Decreased 72%

Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Table 7

Additional Trends Suggested by Results That Were Different from Those of the 2007 Unpublished Just Associates Study

Field Category	Key Trends	Explanations
Last Name	• There were fewer typographical errors and misspellings. • There were twice as many pairs with completely different LNs.	• Possibly due to more compound names
First Name	• Different names are more frequent than they used to be. • FN and LN swaps are 10 times more frequent than the previous study showed.	• Prevalence of nicknames • Registrar not being cognizant of data field being entered
Middle Name	• MN differences are much higher. • Analysis of blanks in MN decreased.	• Middle name captured more often, resulting in higher rates of discrepancy
Social Security Number	• 58% have a blank or default value in one or both records, compared to 36.5% in 2007. • There is a significant volume of typographical or transposition errors.	• Patients still hesitant about sharing SSN, but data capture increased over the years • Many systems require SSN, forcing registrars to change a digit in patient's SSN (because the patient's real SSN is already entered in the original patient record)
Date of Birth	• DOB capture has increased overall over the years. • Birthdate day mismatches and year mismatches are both less frequent than they used to be.	• DOB is collected more accurately than it used to be • Likely helped by Joint Commission initiatives regarding positive patient identification
Gender	• Gender discrepancies have increased over the years.	• Possible explanations: sex-change operations, data entry errors, and reference lab specimen registrations (if the lab test is not gender specific, registrar keeps the default entry)

Abbreviations: DOB = date of birth, FN = first name, MN = middle name, LN = last name, SSN = Social Security number.

Appendix A

Record Matching Algorithms

Record matching algorithms are used to perform front-end record searches, to link records across disparate systems, and/or to identify possible duplicate records on the back end. Algorithms are defined as follows.

- **Basic algorithm:** Simplest technique for matching records; utilizes deterministic matching.
  - The data elements must match exactly (or have an exact partial match) in order to return a particular record match.
  - Comparisons are usually made only on name, birth date, Social Security number (SSN), and sometimes gender.
- **Intermediate algorithm:** Uses more advanced programmatic techniques than basic algorithms to compare records.
  - Phonetic encoding systems and sometimes equivalent name tables are utilized to counter misspelled names and nicknames. Arbitrary/subjective field match weights are assigned to key patient-identifying attributes such as last name, first name, date of birth, and SSN, resulting in a record match weight score.
  - May utilize programs to address transpositions, digit rotations, and typographical errors.
- **Advanced algorithm:** Utilizes the most sophisticated tools for matching records and relies on mathematical and statistical theory.
  - Core intelligence includes probabilistic theory and mathematical/statistical models, which are applied to determine likelihood of a match on specified data elements.
  - It includes machine learning and neural networks, which use forms of artificial intelligence that simulate human problem solving.