



*International Business Machines Corporation  
294, Route 100  
Somers, New York 10589  
914-766-1694*

**HIT Policy Committee**  
**Privacy & Security Tiger Team**  
**Patient Linking Hearing**  
**Thursday, December 9, 2010**

**Written Testimony by Scott Schumacher, Ph. D., Initiate Chief Scientist, IBM Information Management**

Matching patient records – indeed, matching any records – uses mature, well-understood technology. Healthcare providers, payers, health exchanges, data providers, and public agencies have been using probabilistic matching solutions for over a decade. For example, the solution provided by Initiate Systems, now part of IBM, is deployed in nearly a dozen countries, matching records in multiple languages and scripts. Within our base of over 200 customers, our solution deals with wide ranges of scale, data quality, data types, and accuracy requirements.

The solution is also being used today in a range of other industries to match and link non-healthcare records in order to create an industry-specific “single version of the truth” – a single record of a single individual that includes all of that individual’s industry-relevant information.

The insurance industry matches insurance holders with policy and payment information; the financial industry matches account holders with account information; the retail industry matches buyers with buying history and buying habits.

The public sector uses this matching and linking technology in law enforcement and the Department of Homeland Security to connect a vast amount of intelligence from around the world to help identify and thwart terrorist threats.

## **The Technology**

The solution links and matches records using statistical decision theory. It uses what is known as probabilistic matching technology<sup>1</sup> (see attachment 1), which makes maximal use of the information content of each record.

Probabilistic matching technology, for example, would rate two records with names Albert Einstein as a much more likely match than two records with names John Smith, since the latter is a much more common name more likely to occur at random. Similarly, the technique takes into account data quality in discrimination. For example, if a data set has a high error rate in an attribute like gender, then a mismatch on gender does not have the same effect as it would where the error rate was lower.

### *Scalability*

Over the last decade, we have seen a significant increase in the number of records held by our various customers. Ten years ago, a 2 million record set was considered a large file. Now, that is considered a laptop-size problem. We have multiple customers performing real-time matching on data sets of over 500 million records; our largest installation contains over 7 billion records. Typical volumes that we see today are:

- Provider organizations - between 1 million and 25 million records
- Healthcare exchange organizations - between 10 million and 100 million records
- Commercial organizations - between 100 million and 1 billion records

### *Accuracy*

Data completeness, data quality, and accuracy are intimately linked. The probabilistic solution can accommodate data quality issues, but matching is an information-based problem – no algorithm or process can compensate for the lack of information.

As such, most systems are designed around the false-positive (erroneously linking records) rate (or type 1 error). In typical healthcare installations, we design for a false-positive rate of about one in 1 million.

False negative results (not linking records when they should be linked) depend on data quality, the number of records being used, and how complete those records are. Most standard false-negative rates in our healthcare customer base fall in the two to 10 percent range, depending on data quality. We should note that in terms of accuracy, one size does not fit all, even in the same enterprise. Acceptable error rates for point-of-care applications are far different from those for reporting or secondary use.

While matching solutions cannot control the amount of the information they collect, the Initiate solution uses information theory to estimate the false-negative rate a user will receive. Based on this estimate, the organization has the ability to gain a greater understanding of the quality of the data and completeness of records being accessed and used by the system.

## **Identifiers**

There has been much debate over the use of various patient identifiers<sup>2</sup> (see attachment 2).

While the type of identifier is less important, it is critical to note that the effectiveness of any single identifier is based on the assignment mechanism and the workflow around its use. For example, if a patient forgets an ID card and is given a temporary number, the temporary and permanent identifiers need to be reconciled. And, if a patient goes to one facility and receives an identifier, then goes to another facility and receives another identifier, the assignment mechanism needs to account for duplication.

Even if the assignment and reconciliation mechanisms are in place, the workflow must also have the ability to subsume information from places and processes that do not have identifiers.

We all agree on the importance of identifying patients – of ensuring that information about a patient is accurate and accessible regardless of where that patient is being treated. We also agree on the importance of eliminating risks and securing patient information – of maintaining patient privacy and protecting personally identifiable information. There is a danger, however, in using any single identifier<sup>4</sup> (see attachment 4).

In fact, relying on a single identifier negates the possibility of using additional identifiers that may help increase accuracy. For example, while some patients may be reluctant to provide social security numbers, they may be comfortable providing a drivers license number.

Best practices in matching and linking say there is no need to choose between identifiers. All patient identifiers are good. When using matching and linking software, any kind of identifier is merely another data point by which to more accurately identify a patient.

The IBM Initiate software has worked within many environments using a range of identifiers, or no identifiers. Our approach is holistic. It uses all available attributes – identifiers and demographic data – and is considered a best-practices approach.

## **Biometrics**

Biometrics is another area that has elicited hot debate within the healthcare community. It's a highly effective concept – and, at this time, biometrics is still a concept based on its limitations.

In the short term, biometrics faces three primary obstacles: reliability, acceptability, and price.

Reliability issues often revolve around cleanliness, particularly with more ordinary, less expensive devices such as hand scanners. Reluctance of individuals to participate in providing biometric information is a second obstacle.

Organizations can overcome both of these with a higher-end system that is accurate and more secure, but the cost of that system represents the third barrier. With biometric devices, the less expensive the device the less reliable the results; more reliable results require a greater cash outlay.

As with the question of identifiers, within the current linking-and-matching software environment, any biometric information is good. Biometric information of any kind is simply an additional attribute; **a holistic system that uses all available attributes is considered a best-practices approach.**

### **Provider Identification**

While the focus of this testimony is patient matching, provider matching carries an equal importance and should not be viewed as an identification silo. IBM Initiate technology is used to address the provider matching requirements in payer, provider, and Health Information Exchange (HIE) organizations. Research, new healthcare delivery models, and HIE will require that organizations be able not only to match patients to their records and match providers to the appropriate patient records, but also to understand the relationships of the patient and provider records. This may be particularly important as privacy and consent policy is developed and executed.

### **Conclusion**

Healthcare organizations have been successfully matching and linking patient records, and providing improved care, for over a decade. Using probabilistic matching technology makes optimal use of the information at hand.

The IBM Initiate software does more than simply provide matching and linking capabilities. It is a complete solution that includes processes to identify and resolve errors, inconsistencies, and missing information. And, a complete solution is required to address the many aspects of matching records<sup>3</sup> (see attachment 3). The software is a complete, proven process – based on a scientific approach –

that accurately identifies patients and creates “a single version of the truth” across disparate enterprises and around the world.

## **Recommendations**

Accuracy is the key to successful patient identification and, as we have discussed, the information content of the record is the key driver to accuracy. Many approach the information content challenge by trying to specify a “minimal set of attributes” required for linking. Unfortunately, this is too simplistic – it doesn’t address issues such as regional variation in cultures, data collection quality, or variations in accuracy requirements.

As we develop comprehensive exchanges, we need to develop processes for onboarding based on statistical or information-based profiling. The goal is to create a controlled, trusted system.

## **List of attachments:**

1. Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision,
2. Patient Identification in Three Acts, Journal of AHIMA / April 2008, Lorraine Fernandes, RHIA and Michele O’Connor, MPA, RHIA, FAHIMA
3. Data Governance and Data Stewardship: Critical Issues in the Move toward EHRs and HIE, Journal of AHIMA / May 2009, Lorraine Fernandes, RHIA and Michele O’Connor, MPA, RHIA, FAHIMA
4. Universal Health Identifier, IBM White Paper, August 2010, Scott Schumacher, Ph.D. and Lorraine Fernandes, RHIA