Technical
White Paper

Deterministic patient matching

Watson Health analytics and informatics

Watson Health™

IBM

Contents

Introduction

Correctly matching records is necessary for accurate quality measures, financial reporting, and population management, but determining which healthcare records belong to each patient is a challenging problem. Some data sources contain limited patient identifiers (e.g., laboratory feeds or insurance claims) and at other times demographic data elements conflict (e.g., asynchronous databases or basic human data entry errors).

Because multiple data sources are typically aggregated in the IBM® Explorys Platform, there are multiple ways to link records together. In addition to a probabilistic matching algorithm documented in the Watson Health™ whitepaper entitled, “Probabilistic Person Matching”, the IBM Explorys Platform can use a deterministic algorithm to focus on the enterprise master patient index (EMPI) values on records from Watson Health partners.

Additional source systems like laboratory feeds and claims datasets often don't have access to the EMPI assignment systems, therefore, some type of backup matching algorithm will be necessary to fold data sources together. For example, all but the last row in the table below might belong to the same patient even though not all have an EMPI value, so a backup approach to collating records is necessary.

Name	Sex	Birth date	Address	MRN	EMPI
Janice Doe	F	1987-04-02	123 Cedar, Cleveland OH 44106	A12345	975312468
Janice Doe	F	1987-04-02	123 Cedar, Cleveland OH 44106	–	975312468
Jan Doe	F	1987-04-12	–	XYZ987	975312468
Janice Doe	F	1987-04-02	987 Main, Columbus OH 43210	–	

Note: The names and information that appear in the figures in this paper are used fictitiously for sample purposes only, and any resemblance to actual persons is entirely coincidental.

Deterministic matching

If the EMPI field is populated on a demographic record, then the deterministic matching algorithm simply checks for matches on that field. If the EMPI field is not populated, then a backup matching algorithm is applied to find records that should still be matched.

Backup matching

Watson Health matches patients from disparate systems into the IBM® Explorys Data Grid using proprietary algorithms coupled with elements of the New York State Identification and Intelligence System(NYSIIS). The NYSIIS transformation of names is essentially a phonetic algorithm devised by New York State that builds on the traditional SOUNDEX algorithm to transform names into phonetic coding to facilitate the matching of both first and last names.

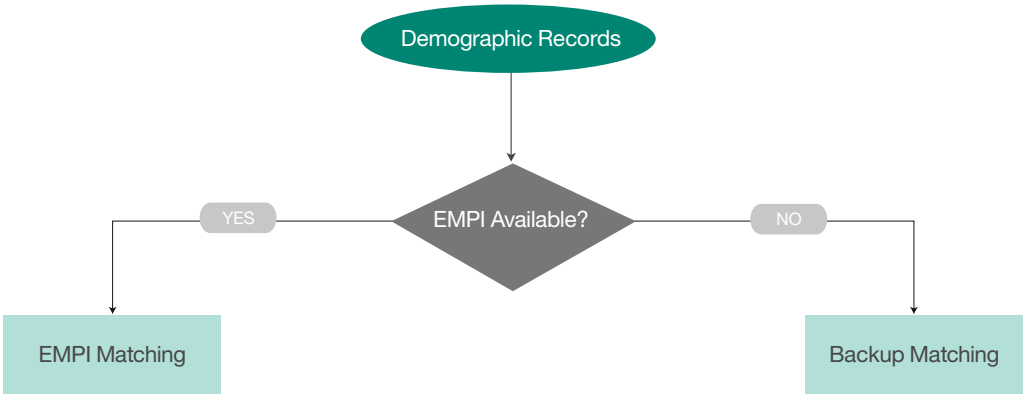

Figure 1: Flowchart of deterministic matching

The NYSIIS algorithm has 11 basic rules that replace common pronunciation variations with standardized characters, remove repeated characters, and replace all vowels with the letter "A." Because the NYSIIS algorithm retains information on the sequence of vowels, it has higher discriminating power than SOUNDEX. For example, the NYSIIS transformations for "Shaun" and "Sean" are both "SAN."

The IBM® Explorys Patient Matching Engine algorithm consists of a cryptographic hash function (SHA-512, designed by the National Security Agency) of five data elements:

– First name (using NYSIIS)
– Last name (using NYSIIS)
– Date of birth
– Gender
– Zip Code (3 digit)

Note that the term "hash" simply refers to a mapping of longer more complex data into smaller datasets of a fixed length. The National Institute of Standards and Technology (NIST) website has additional information on their Federal Information Processing Standards Publications (FIPS PUBS) website.

Backup matching algorithm performance

Matching performance will depend on the particular characteristics of the datasets being aggregated. Stale address information and incomplete demographic information can affect backup matching, while both type I and type II errors in the EMPI assignments in data sources will propagate with the deterministic matching algorithm.

Watson Health tested its deterministic backup matching algorithm above against a mature Enterprise Master Patient Index (EMPI) in use for more than 5 years. Although not a "Gold Standard", the EMPI at the Watson Health customer site was the identifier chosen for all clinical, operational, and financial transactions and also the identifier used for local health information exchange (HIE).

Performance Matrix	EMPI Match	EMPI Non-match
Backup Match	> 98.4%	< 1.6%
Backup Non-match	< 0.02%	> 99.9%

This table shows the comparison of an EMPI and the backup matching algorithm. This is based on an actual population of more than 5M patient records. Moreover, it is estimated that the IBM Explorys Patient Matching Engine algorithm has a false-positive rate of 0.015% based on manual inspection of 10,000 patient records (= 0.01).

Manual review was performed on instances where multiple customer EMPI records existed for records grouped together by the backup matching algorithm. In many cases it is shown that these were indeed the same patients for whom the EMPI failed. Thus, the IBM Explorys Patient Matching Engine algorithm improved upon the EMPI, matching an additional 2 percent of patients that had been missed by the presence of an EMPI alone.

The probabilistic patient matching algorithm enables customers to combine data from multiple sources in a scalable, configurable, and accurate way, and plays an important role in improving the value of the IBM Explorys Platform.

References

1. New York State Identification and Intelligence System (NYSIIS)
2. Federal Information Processing Standards Publications (FIPS PUBS)
3. Watson Health "Probabilistic Person Matching" Algorithm whitepaper

About IBM Watson Health

In April 2015, IBM launched IBM Watson Health and the Watson Health Cloud platform. The new unit will work with doctors, researchers and insurers to help them innovate by surfacing insights from the massive amount of personal health data being created and shared daily. The Watson Health Cloud can mask patient identities and allow for information to be shared and combined with a dynamic and constantly growing aggregated view of clinical, research and social health data.

For more information on IBM Watson Health, visit: ibm.com/watsonhealth.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2016

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed or misappropriated or can result in damage to or misuse of your systems, including to attack others.

No IT system or product should be considered completely secure and no single product or security measure can be completely effective in preventing improper access. IBM systems and products are designed to be part of a comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM does not warrant that systems and products are immune from the malicious or illegal conduct of any party.

Note: The names and information that appear in the figures in this paper are used fictitiously for sample purposes only, and any resemblance to actual persons is entirely coincidental.

HPW03026-USEN-02

Watson Health™