International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 4, April 2016

Machine Learning Methods for Medical Diagnosis VIA Web Application

[1] Anz Joseph [2] Ashik Mohamed K [3] Venvas Emmanuvel A [4] Mrs. Paul T Sheeba
[1][2][3] UG student, [4] Assistant Professor & Head
Department of CSE, Loyola Institute of Technology

***Abstract* — In this era of medical field ,this particular industry is developing in a rapid speed . Open information and data explosion in healthcare industry is on a tipping point. Big Data plays a major role in this new change. One of the biggest challenges that the medical industry faces while it steps up digitization is the disparate data, speed of generation of this data and complexity arising out of multiple & non-standard formats. Patient data residing in disparate systems is a roadblock to having the right information at the right time. Clinical Decision Support systems need a single view of the patient for making better diagnosis and treatments. Patient identification and matching is a critical challenge in interfacing to the Electronic Health Record (EHR). Different documents and results from various disparate systems like laboratory, pharmacy, claims systems etc. need to be linked to the correct patient record. At this point when healthcare organizations share patient information internally as well as externally, patient records from numerous disparate databases should be connected effectively to guarantee that the decisions made by the clinicians are based on correct patient records and minimizing duplicate information and overheads. This will help to do better diagnosis process.**

This paper attempts to study the problem disparate systems and proposes a solution by using a social network for medical care and Data mining techniques for better clinical decision support and diagnosis. The main benefits of the proposed system are scalability, cost-effectiveness, flexibility of using and handling of any data source and ease in medical diagnosis.

***Keywords*— Data mining, Patient Matching, Clinical Decision Support System**

I. INTRODUCTION

One of the major challenges in Healthcare is to consolidate disparate patient data into one view [1]. Patient data resides at many places like clinical, billing, laboratory, pharmacy and claim systems. These separate systems contain duplicate information for the same patient. Even there exists variance in patient record formats. The expectation of having a "one patient, one record" level of care continues to rise every year. Generation of a longitudinal patient record is dependent on accurate patient identification. Without accurate patient identification portions of the longitudinal patient health record may be fragmented into isolated episodes (duplicates) or attached to the wrong patient (overlays). This disparate patient view can affect patient care and make healthcare professionals less productive. The need of the hour is to identify, cleanse, match, de-duplicate and merge patient records to create a master index that may be used to generate a complete and single view of a patient [2]. This web application is a promising right direction which is in its early stages for the healthcare sector. Healthcare is a data-rich domain. As more and more information is being collected, there will be expanding interest for big data analytics. Unscrambling the "Big Data" related complexities can facilitate many insights about making the right choices at the right time for the patients. Data with more complexities continue developing in healthcare thus prompting more opportunities for big data analytics. The problem of consolidating patient data from disparate systems can be solved using Big Data Analytic techniques like data mining,web application & social network[3,4,5,6]. These techniques perform matching at multiple levels and discover complex relationship between disparate data sets. These algorithms can be used to identify de-duplicates from the data based on identified keys. Data mining can also be used to match patients' data. Multiple attributes of a patient can be assigned different weights. For any pair of entities, distance is calculated between corresponding attributes. Attribute wise distances are aggregated over all the attributes of a patient record to find the distance between two patient records. Using Query, data about a patient identity can be exploded and multiple records can be generated with Adhar ID. This paper studies the problem of matching patient records from disparate systems and proposes a solution by using Data mining techniques is used for better clinical decision support. The main benefits of the proposed system are scalability, cost-effectiveness, flexibility , ease in medical diagnosis.

II. RELATED WORKS

- Social networks for doctors allows to introduce their new theories. And this will help these communities, providers find medical advice and best-practices, job openings and career tips, research and product information, as well as

opportunity securely communicate with peers. Patient focused networks, often built around a particular condition or disease. But this network wont help medical care units for clinical decision support system and medical diagnosis system.

❖ A medical algorithm is any computation, formula, statistical survey nomogram, or lookup table useful in health care. Medical algorithms include decision tree approaches to healthcare treatment (eg;if symptoms A, B and Care evident, then use treatment X) and also less clear-cut tools aimed at reducing or defining uncertainty.As the medical field is improved day by day, updating the algorithm is not easy , which will provide false information to the user.

III.WHY IS PATIENT DATA MATCHINESO IMPORTANT?

A Clinical Decision Support System (CDSS) is a software designed to help clinicians in making decisions by matching individual patient characteristics to computerized knowledge repositories with the aim of generating patient-specific assessments or recommendations. As decisions are based on higher volumes of data that are more current and relevant, it can make decision support process simpler, quicker and ultimately more accurate. SMARTHealth India, a CDSS system, for cardiovascular diseases is one such example [14]. A good CDSS should have a single patient view for better decision making based upon longitudinal patient records. Many health organizations use Enterprise Master Patient Index (EMPI) systems to maintain accurate, consistent, and up-to-date demographic and vital medical data on the patients visited and managed within its different departments. The patient is assigned a unique identifier that is used to refer to this patient across the enterprise. The goal is to make sure that each patient is represented only once across all the software systems used within the organization [15]. The essential patient data includes name, date of birth, gender, social security number, ethnicity and race, current address and contact information, current diagnoses, insurance information, most recent date of hospital admission and discharge. A 2008 RAND report, "Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the US Health Care System," noticed that the healthcare system in the US could save about $4.5 billion per year by avoiding adverse drug events, usually caused due to incomplete linking information about a patient's medications or allergies [16]. This report further states that in spite of the widespread deployment of EMPIs, the average duplicate medical record rate in the healthcare industry is an unacceptable 8%. If the MPI databases had more than 1 million records, the average duplicate record rate increased to 9.4%. In addition to that, the report identified that the duplicate record rates of the EMPI databases studied were as high as 39.1% [17]. The dup rate is much higher for larger, more complex organizations such as integrated delivery networks and health information exchanges (HIEs). The main problem is data integrity - missing, incorrect, non-standardized, out of date data, address discrepancies, phone number discrepancies, Name discrepancies, aliases and data entry errors. Another causes of duplicate patient records can be multiple information systems and databases, merger and acquisition data consolidation, EMR upgrades or replacements & poor system integration, or no integration. Even a "fat fingers" typing error can prevent matching old and new records. Patient matching is a critical patient safety issue. Comingling of medical information from two or more patients can lead to catastrophic events. Inability to link multiple records for the same patient leads to fragmented, incomplete records resulting in less than optimal outcomes. The clinical effect is showed as inefficient, inadequate or improper care and results in readmissions, privacy violation and low quality short and long haul patient wellness [18]. Interoperability and health information exchange (HIE) is another major requirement to do patient matching. Electronic health data from disparate sources, whether aggregated in a repository or linked 'just in time', must be precisely matched to avoid data fragmentation and inaccurate consolidation.Accurate billing is another benefit of matching patient data. Duplicate medical records may not contain the patient's current healthcare coverage information which can prolong the revenue cycle. Inability to do accurate patient matching results in errors in billing which deferpayments and build customer frustration and dissatisfaction. It further causes regulatory and compliance failures which can result in fines or reduction in reimbursement funds [18].

IV.CHALLENGES IN PATIENT RECORDSMACHINE

In order to overcome patient matching challenges, a healthcare organization should start by getting it right within first. It requires utilizing Healthcare IT efficiently with the right data strategy to generate a single view of patient. Few major challenges in matching patient records from disparate systems are:

A. Lack of nationwide unique patient identifier

Without a nationwide unique patient identifier, accurately matching multiple records for the same patient from disparate sources is a great challenge [1]. The industry's answer to this challenge is the Enterprise Master Patient Index (EMPI) which uses a variety of statistical algorithms to match patient records while simultaneously endeavouring to minimize the number of false positive and false negative matches. But even with these systems, there exists lots of duplicate patient records [16].

B. Disparate Systems

Various healthcare systems maintain their own set of records on their own platforms, and these systems rarely

share information. They hold patient information in disintegrated manner. Due to nature of data captured in healthcare system, it is not possible to hold all information on single existing legacy platform. Data are highly fragmented in this domain. Some major pools of data are provider clinical data, laboratory data, claims system data, patient behaviour and sentiment data and pharmaceutical data. This data resides in many different types of source systems like RDBMS, Flat files, Excel, and Pdfs etc. This lack of communication can have a negative impact on patient's care.

C. Scalability

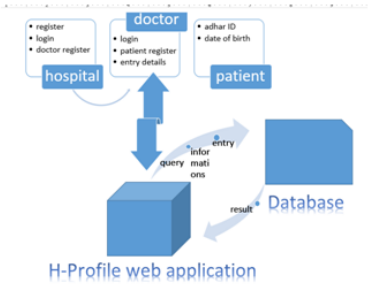
Healthcare is becoming increasingly digitized and fragmented across an endless string of specialties. The evergrowing volume of patient data contained within disparate clinical systems continues to expand. It is very difficult to handle large amount of patient data coming at a high speed using existing technologies due to scalability issues. It causes bigger challenge to match huge volume of patient records. Buying new, better infrastructure incurs huge expenses. This is a limitation with existing solutions and Health IT finds it hard to deal with it.

D. Costly and inflexible existing solutions

Existing out-of-the-box toolsets for patient matching which are available in the market are costly with high licensing fees. Also they are not very flexible and limited to a few fuzzy matching algorithms.

These challenges are roadblocks in the process of effective patient matching. The need of the hour is to have a scalable, cost-effective, flexible solution which can handle any type of data source.

V.PROPOSED SOLUTION OF PATIENT ARCHITECTURE



Modules:

There are four modules in our project. Each module is implemented separately and integrated at the end. The modules are:

❖ ***REGISTRATION***

Registration process is required to get the information from the client, who will use the service. Registration form will have inputs from the user with various data from the user.

❖ ***LOGIN***

In computer security, logging in, (or logging on or signing in or signing on), is the process by which gains access to a database by identifying and authenticating themselves. The user credentials are typically some form of "username" and a matching "password", and these credentials themselves are sometimes referred to as a login, (or a logon or a sign in or a sign on). In practice, modern secure systems also often require a second factor for extra security.

❖ ***MEDICAL ENTRY***

The terms medical record, health record, and medical chart are used somewhat interchangeably to describe the systematic documentation of a single patient's medical history and care across time within one particular health care provider's jurisdiction. The medical record includes a variety of types of "notes" entered over time by health care professionals, recording observations and administration of drugs and therapies, orders for the administration of drugs and therapies, test results, x-rays, reports, etc. The maintenance of complete and accurate medical records is a requirement of health care providers and is generally enforced as a licensing or certification prerequisite.

❖ ***MEDICAL DIAGNOSIS***

Medical diagnosis is the process of determining which disease or condition explains a person's symptoms and signs. It is most often referred to as diagnosis with the medical context being implicit. The information required for diagnosis is typically collected from a history and physical examination of the person seeking medical care. Often, one or more diagnostic procedures, such as diagnostic tests, are also done during the process. Sometimes Posthumous diagnosis is considered a kind of medical diagnosis.

Patient records are dispersed across multiple treatment facilities and geographies that have disparate technologies. False positive medical record matches co-mingle information from two or more different people raising safety issue. False negative medical record matches fail to link multiple records for the same person resulting in a fragmented, incomplete EHR which can compromise outcomes.

Data mining and Big Data can be leveraged to improve the identification of unique patient records in healthcare organization patient files and/or EMPI files stored in hospitals, labs, pharmacies and health plans, which as a

result, will greatly enhance and facilitate health information exchange. The services provided by this web application can provide easier data mining and Big Data. This web application perform matching at multiple levels and discovers complex relationship between disparate data sets.

This paper proposes usingdata mining to match enormous amount of patient records while overcoming most of the challenges mentioned in previous section. The proposed solution is explained with an example of indicative data which will be scalable to any amount of data flow for patient matching improvement. As shown in Fig. 2, let's say 30 patient records from 4 different systems are consolidated after integration, cleansing and transformation e.g. date of birth is converted to same format

Sample Patient Data

Patient ID	Patient Name	Gender	DOB	Address1	Address2
1	Richard Gomez	M	11/16/1996	2934 Encino Pt	San Antonio TX 78259
2	Jim Dobbs	M	6/4/1938	2756 Bulls Bay Hwy	Jacksonville FL 32220
3	Robert Lewis	M	12/24/1971	194 Buckboard Dr	Augusta GA 30907
4	Rick Gomez	M	11/16/1969	2934 Encino Point	San Antonio TX 78259
5	Dharam Patel	M	8/11/1937	84 Prospect Hill Dr	Tewksbury MA 01876
6	Richard Gomez	M	11/6/1996	2935 Encino Pt	San Antonio TX 78259
7	Jim Dobbs	M	4/6/1938	2756 Bulls Bay Hwy	Jacksonville FL 32220
8	Ellis Cornwell	M	11/16/1996	4031 Laurel Lane	Midland TX 79701
9	Elizabeth Lange	F	6/4/1938	1447 Orchard Street	Minneapolis MN 55401
10	Salvador Keels	M	12/24/1971	339 Elliot Avenue	Seattle WA 98115
11	Richard Gomaz	M	11/9/1996	29350 Encino Pt	S. Antonio TX 78259
12	Edith Morgan	F	12/31/1948	2145 Stout Street	Carlisle PA 17013
13	Robert Dorman	M	8/11/1937	80 Snowbird Lane	Omaha NE 68144
14	Nicole Zorn	F	11/2/1986	1559 Shingleton Road	Grand Rapids MI 49503
15	Greg Hill	M	7/5/1987	4384 Elkview Drive	Stuart FL 34994
16	John Boyd	M	4/4/1990	1849 Liberty Street	Plano TX 75074
17	Kenneth Austin	M	7/19/1939	2497 Rockford Road	Mcgill NE 89318
18	Crystal Walker	F	12/28/1937	2437 Church Street	New York NY 10017
19	Mildred Cross	F	6/16/1956	1227 Fowler Avenue	Norcross GE 30093
20	Donald Sorensen	M	10/7/1963	3323 Oak Avenue	Hickory Hills FL 60457
21	Nancy Hughes	F	7/19/1938	60 Losh Lane	Hickory PA 15340
22	Valerie Brown	F	10/4/1975	1867 Ward Road	El Paso TX 79902
23	Sandra Philpps	F	1/4/1971	248 Railroad Street	Jacksonville FL 32207
24	Boyd John	M	4/4/1990	1849 Liberty Street	Plano TX 75074
25	Rich S. Gomez	M	11/16/1996	2934 Encino Point	San Antonio TX 78259
26	Crystal Wallker	F	12/28/1937	247 Church Street	New York NY 10017
27	Mildred Cross	F	6/6/1956	127 Fowler Avenue	Norcross GE 30093
28	Don Sorensn	M	10/7/1963	3323 Oak Avenue	Hick Hills FL 60457
29	Nancy Hughs	F	7/19/1938	601 Losh Lane	Hickory PA 15340
30	Valere Brown	F	1/4/1975	1867 Ward Road	El Paso TX 79902

Fig. 2 Partial dataset showing how the same patient record is

Multiple times after consolidation (*Above stated data is for indicative purpose only) Patient "Richard Gomez" is having 5 similar records after consolidation. Now the challenge is to match his records to remove de-duplication. Fig. 3 shows the steps of proposed Patient data is first extracted from various disparate systems like Clinical, Laboratory, Pharmacy, and Claims Systems . Data is integrated, cleansed and transformed. Effective matching requires the use of probabilistic techniques. Patients (Name, ID & Address) across disparate data sets (silos of Hospitals) via synonyms, phonetics and approximate spellings. Fuzzy matching is an advanced mathematical process in which the similarities between data sets, information, and facts are determined. The result of these similarities is not always true nor false, or 100% certain. In this process, any data type of any length and from any place in a field is compared to find non-exact matches. For every piece of data examined, a probability score is generated using data mining process to determine the accuracy of the match

This can be used to match patients' data. There are multiple attributes of a patient identity e.g. Name, Date of Birth, Address, City, Pin Code etc. One's identity is what differentiates one from other people, but different parts of this identity can differentiate to a greater or lesser extent. For example, people can be better differentiated by names compared to date of birth as many more people were born the same day than people with the same name. Then each attribute identity has an associated weight. Patient attributes are assigned different weights (e.g. Last name match counts for more than a First name match, Date of Birth match counts for more than City match) [2]. These weights are configurable as per the needs of a healthcare organization. For the above stated example.A Social network which will enable to connect all the medical care.Collection of each patient's medical history with unique numbers.Collection of treatment and medicine history for both doctors and Hospitals.Sharing the medical records without revealing the identity of patient's.Formation of medical profile &Diagnosis using the history of medical records.

VI. REDUCING COSTS AND IMPROVING HEALTHCARE EFFICIENCY

While Big Data programs might seem like a hefty investment at the front end, using Big Data analytics internally at hospitals could drive down the cost of operations by detecting inefficiencies in patient identification, by reducing risks of incorrect diagnosis and treatment, by increasing patient satisfaction and by improving Clinical Decision Support systems. In Pharmacy, real-time unified single view of patient will help to identify and reduce instances of customers having multiple contacts with pharma contact channels for the same transaction. By consolidating claims data, lab data, pharmacy data and self-reported patient data, various drug adherence programs can also be developed.

VII.CONCLUSIONS

Today, Patient matching is one of the biggest challenges faced by healthcare industry and is critical for successful Health Information Exchange (HIE) and public health. The challenge starts with getting it right within a healthcare organization first by matching patient records coming from various disparate systems like clinical, billing, laboratory, pharmacy and claim systems. The repercussions of inaccurately matched patient records or unmatched patient records can be life threatening. Combining medical and medication history for two unique individuals can lead to incorrect diagnoses, adverse drug events, unnecessary

diagnostic tests and in the worst scenario, even death. On the other hand, if multiple medical records for a patient are not linked together, the missing medical and medication history for that patient results in a fragmented, inadequate medical record which can also lead to dire consequences. This paper recommends an improved Patient matching process for high volume and high velocity data. It investigates an effective and enduring Big Data Analytics approach using Fuzzy algorithm (Levenshtein Distance) and MapReduce techniques to patient matching of large repositories for better Clinical decision making. The main benefits of this proposed system are scalability, cost-effectiveness, flexibility of using any fuzzy algorithm and handling of any data source. Healthcare organizations can follow this approach internally to ensure their short and long term clinical success, financial stability, and, ultimately, their survival. This solution can be further extended to compare performance of various Fuzzy logic algorithms on MapReduce using different infrastructure resources by large Healthcare Organizations.

REFERENCES

[1] Dooling, J. A., et al. "Managing the integrity of patient identity in health information exchange (updated)." *Journal of AHIMA/American Health Information Management Association* 85.5 (2014): 60.

[2] Office of the National Coordinator for Health Information Technology (February 7, 2014), "Patient Identification and Matching Final Report". Retrieved from http://www.healthit.gov/sites/default/files/patient_identification_match ing_final_report.pdf

[3] Justin Campbell (2009, December 8,). "Legacy Data Conversion: Fuzzy Patient Matching to the EHR", Galen Healthcare Solutions Web log. Retrieved from http://blog.galenhealthcare.com/2009/12/08/legacy-data-conversionfuzzy-patient-matching-to-the-ehr/

[4] Gosh (2013, September 9). "Identifying Duplicate Records with Fuzzy Matching" [Web Log Post]. Retrieved from https://pkghosh.wordpress.com/2013/09/09/identifying-duplicaterecords-with-fuzzy-matching/

[5] Pablo Pazos (2010, March 26). "Increased data quality patronymic", Informatica Medical [Web Log Post]. Retrieved from http://informatica-medica.blogspot.in/2010_03_01_archive.html

[6] Gianmarco De Francisci Morales, Aristides Gionis, Mauro Sozio, "Social content matching in MapReduce" research paper presented at 37th International conference on Very Large Databases (VLDB), 2011

[7] IBM, "Big Data at the Speed of Business," [Online]. Available: http://www-01.ibm.com/software/data/bigdata/, 2012.

[8] Talend. "A total data management approach to big data", White Paper, Oct 2010.

[9] Raghupathi, W., &Raghupathi, V. (2014). "Big data analytics in healthcare: promise and potential". *Health Information Science and Systems*, 2(1), 3.

[10] Image Source: http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare/part1.pdf

[11] Duggal, Reena, Balvinder Shukla and Sunil Kumar Khatri. "Big Data Analytics in Indian Healthcare System – Opportunities and Challenges" research paper accepted at *National Conference on Computing, Communication and Information Processing* (NCCCIP-2015) – May 2015: 92-104

[12] The Data Warehouse Institute. [Online]. Available: http://tdwi.org/portals/big-data-analytics.aspx

[13] Gartner Press Release (October 22, 2012), "Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015", [Online]. Available: http://www.gartner.com/newsroom/id/2207915

[14] Praveen, D. (2013, March). SMARTHealth India: Development and Evaluation of an Electronic Clinical Decision Support System for Cardiovascular Diseases in India. In Medicine 2.0 Conference. JMIR Publications Inc., Toronto, Canada.

[15] Wikipedia, "Enterprise master patient index", [Online]. Available: https://en.wikipedia.org/wiki/Enterprise_master_patient_index, May 2015

[16] Hillestad, Richard, et al (2008). "Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the US Health Care System", Santa Monica, CA: Rand.

[17] AHIMA White paper, "Ensuring Data Integrity in Health Information Exchange", AHIMA HIE Practice Council, July 2012

[18] Dan Carotenuto (2014, January 27). "Improving Patient Matching: Using Healthcare IT And Data Strategy To Create A Single Patient View" [Web Log Post]. Retrieved from http://www.informationbuilders.com/blog/dan-carotenuto/15679

[19] Beth Haenke (2012, October 23). "Record-Matching Integrity: An Algorithm Primer", Health Data Management Web log. Retrieved from http://www.healthdatamanagement.com/blogs/healthcare-algorithmsdefinitions-differences-45144-1.html