JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SCRUB SYSTEM FOR MEDICAL RECORDS

Aruna Kamble (Faculty)
Department of CSE
Bharati Vidyapeeth College of Engineering
Belapur, Navi Mumbai Maharashtra, India
aruna.s.kamble@gmail.com

Pushkar Wathore (Student)
Department of CSE
Bharati Vidyapeeth College of Engineering
Belapur, Navi Mumbai Maharashtra, India
pushkarwathore14@gmail.com

***Abstract*—Medical sector has a vast amount of data and every tiny bit of it can be sensitive enough to cause trouble for a person. This report aims to help the privacy protection of the medical and healthcare data so that hospitals can share their vital data for research without hesitating of any misuse of data. The algorithm presented in this report makes data clearer and also, masks the direct identifiers along with listing down the quasi- identifiers that might be risky for the sharing of data.**

Keywords: Data scrubbing, Personally-identifiable data, Data masking, Medical care, Confidentiality of patients, Data privacy

INTRODUCTION

Medical industry is booming in recent years with all the advancements in the knowledge of doctors, equipment and medicines. But one of the major issues which still exist is the sharing of medical data with the responsibility of maintaining the confidentiality of the patients. The need is to supply only the required data to the researchers who are working on the medicine or doing the statistics and mask the data which can be classified as identifiable or unnecessary. Main objectives include:-

(1) Medical institutions should not be hesitant to share their data involving their patients to the government or research institutions.

(2)Protecting the confidentiality of the patients by scrubbing the unnecessary data.

(3) Scrubbing the data in such a manner that the dataset becomes clearer to the required researcher and the utility is maintained.

(4) Scrubbing mechanism will use a selective search algorithm for identifying explicit identifiable attributes.

LITERATURE SURVEY

[1] Probably the greatest test confronting therapeutic informatics is the sharing and spread of medicinal records while keeping up a pledge to persistent privacy (Sweeney 1996). Review examination, diminished organization costs, improved therapeutic consideration and the advancement of electronic medical record frameworks are a portion of the advantages conceivable when the substance of restorative records are checked on by experts. The issue isn't as straightforward as scanning for the patient's name and supplanting all events with a pseudo-name. Distinguishing data is regularly covered up in the shorthand notes of clinicians and in letters traded between specialists.

[2] Content based patient therapeutic records are a crucial asset in medicinal research. So as to protect understanding privacy, be that as it may, the U.S. Medical coverage Portability and Accountability Act (HIPAA) necessitates that wellbeing data (PHI) be expelled from therapeutic records before they can be dispersed. Manual de-distinguishing proof of huge therapeutic record databases is restrictively costly, tedious and inclined to mistake, requiring programmed strategies for huge scope, mechanized de-ID (Neamatullah, I., Douglass, M.M., Lehman, L.H. et al. Computerized de-recognizable proof of free-content restorative records. BMC Med Inform Decis Mak 8, 32 (2008)). In any de- distinguishing proof framework, there is a solid chance that the product may experience PHI that are missing in the broad word references of known PHI and that are likewise not recognized fair and square.

[3] Electronic medicinal record (EMR) frameworks have empowered social insurance suppliers to gather nitty gritty patient data from the essential consideration space. Simultaneously, longitudinal information from EMRs are progressively joined with biorepositories to create customized clinical choice help conventions. Rising arrangements urge specialists to scatter such information in a deidentified structure for reuse and coordinated effort, however associations are reluctant to do so in light of the fact that they dread such activities will risk tolerant security. Specifically, there are worries that lingering segment and clinical highlights could be misused for reidentification purposes (Tamersoy and Loukides 2012).

[4] The expanding health information infrastructure offers the promise of latest medical knowledge drawn from patient records(Behlen and Johnson 1999). during this article, the authors analyze the interests of patients and institutions in light of public policy and institutional needs. They conclude that the multicenter study, with Institutional Review Board approval of every study at each site, protects the interests of both. "Anonymity" is not any panacea, since patient records are so rich in information that they will never be truly anonymous. The authors find that computer security tools are needed to administer multicenter patient records studies and describe simple approaches which will be implemented using commercial database products. this text presents, first, an analysis of the policy problems with multi-institutional patient records research, then discusses these issues in light of implementation experience in an inter-institutional “virtual repository” project. Privacy and confidentiality of the patient record has attracted extensive debate and analysis, including discussion of research. The institution's patient records contain information not only about patients but also about physicians and therefore the institution. While public policy and professional ethics set conditions and procedures for the utilization of patient data, the participation of provider institutions in research is strictly voluntary, and a search program must meet the standards and serve the interests of every participating institution. Thus, a system of sharing patient records must recognize and protect the interests of both the patients and therefore the contributing institutions.

[5] The use of medical records and human tissues in biomedical research within the U.S. is roofed under the Standards for Privacy of Individually Identifiable Health Information and therefore the Common Rule. In response to a congressional mandate within the insurance Portability and Accountability Act of 1996 (HIPAA), the Department of Health and Human Services (HHS) issued the HIPAA Privacy Rule regulations in December 2000. The Privacy Rule permits covered entities (i.e., health plans, health care clearinghouses, or health care providers who transmit health information in electronic form in reference to a transaction that HHS has adopted standards) to use and disclose data that are removed of patient identifiers without obtaining an authorization and without further restrictions on use or disclosure because data removed of those identifiers are not any longer protected health information (PHI) and, therefore, aren't subject to the Privacy Rule. The utilization of those strategies in multicenter research studies is paramount in importance, given the necessity to share EHR data across multiple environments and institutions while safeguarding patient privacy(Kushida, Nichols, Jadrnicek, Miller, Walsh, Griffin 2012). Systematic literature search using keywords of deidentify, de- identify, deidentification, de-identification, anonymize, anonymization, data scrubbing, and text scrubbing. Search was conducted up to June 30, 2011 and involved 6 different common literature databases. a complete of 1,798 prospective citations were identified, and 94 full-text articles met the standards for review and therefore the corresponding articles were obtained. Search results were supplemented by review of 26 additional full-text articles; a complete of 120 full-text articles were reviewed.

[6] Text-based patient medical records are an important resource in medical research. so as to preserve patient confidentiality, however, the U.S. insurance Portability and Accountability Act (HIPAA) requires that protected health information (PHI) be far away from medical records before they will be disseminated. Manual de- identification of huge medical history databases is prohibitively expensive, time-consuming and susceptible to error, necessitating automatic methods for large-scale, automated de-identification. We describe an automatic Perl-based de- identification software package that's generally usable on most free-text medical records, e.g., nursing notes, discharge summaries, X-ray reports, etc. The software uses lexical look-up tables, regular expressions, and straightforward heuristics to locate both HIPAA PHI, and an extended PHI set that has doctors' names and years of dates. This gold standard corpus was wont to refine the algorithm and measure its sensitivity. to check the algorithm on data not utilized in its development, we constructed a second test corpus of 1,836 nursing notes containing 296,400 words. The algorithm's false negative rate was evaluated using this test corpus.A wide range of medical research – from epidemiology to the planning of decision support systems – relies on medical records (Saeed M, Lieu C, Raber G, Mark 2002). For both legal and ethical reasons, it's necessary to preserve patient confidentiality. Within the us the insurance Portability and Accountability Act (HIPAA) specifies 18 specific categories of data that has got to be far away from medical records to be utilized in research. These categories of protected health information (PHI) include names, geographic locations (more precise than a state), elements of dates except years, Social Security numbers, telephone and fax numbers and medical history numbers, among others.

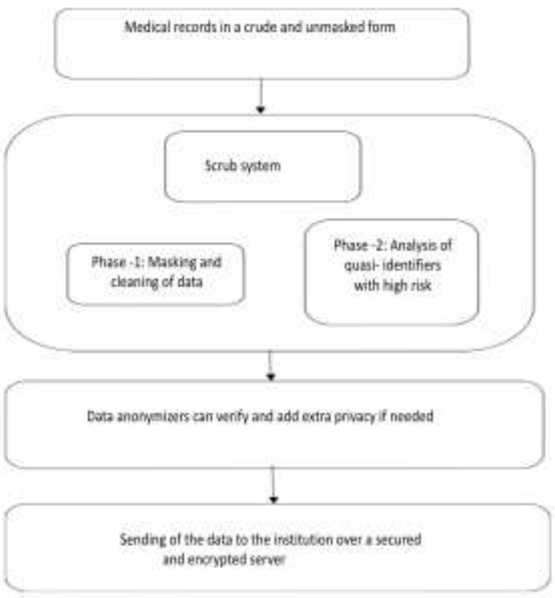
[7] In the last years, the need to de-identify privacy-sensitive information within Electronic Health Records (EHRs) has become increasingly felt and extremely relevant to encourage the sharing and publication of their content in accordance with the restrictions imposed by both national and supranational privacy authorities. In the field of Natural Language Processing (NLP), several deep learning techniques for Named Entity Recognition (NER) have been applied to face this issue, significantly improving the effectiveness in identifying sensitive information in EHRs written in English. However, the lack of data sets in other languages has strongly limited their applicability and performance evaluation. To this aim, a new de-identification data set in Italian has been developed in this work, starting from the 115 COVID-19 EHRs provided by the Italian Society of Radiology (SIRM): 65 were used for training and development, the remaining 50 were used for testing. The data set was labelled following the guidelines of the i2b2 2014 de-identification track. As additional contribution, combined with the best performing Bi-LSTM + CRF sequence labeling architecture, a stacked word representation form, not yet experimented for the Italian clinical de-identification scenario, has been tested, based both on a contextualized linguistic model to manage word polysemy and its morpho-syntactic variations and on sub-word embeddings to better capture latent syntactic and semantic similarities. Finally, other cutting-edge approaches were compared with the proposed model, which achieved the best performance highlighting the goodness of the promoted approach.

METHODOLOGY

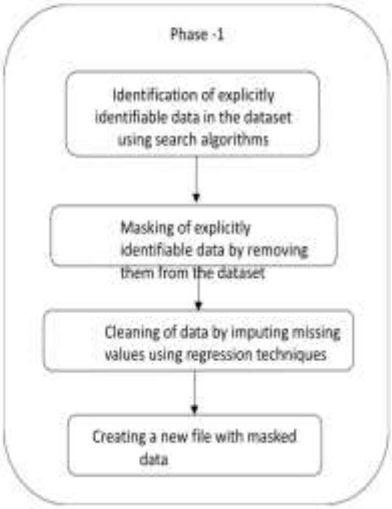
A data scrubber will help in protecting the confidentiality of patients and maintaining their privacy. It will additionally focus on improving the quality of data, thus improving the overall utility of the data and making the data more accurate.

The application uses Python and Data Analysis to find the explicit Identifiers and Secondary Quasi-identifiers by improvising already used methods like "Search and replace". It also helps clean the dataset by using Single Imputer, Bayesian Ridge Algorithm.

ARCHITECTURE DIAGRAM

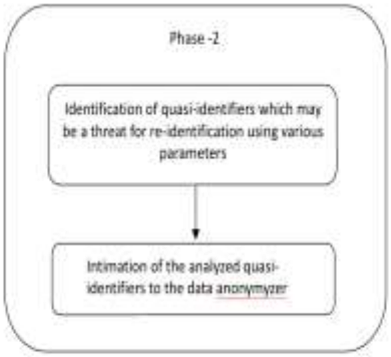

Medical records in a crude and unmasked form
Scrub system
Phase -1: Masking and cleaning of data
Phase -2: Analysis of quasi- identifiers with high risk
Data anonymizers can verify and add extra privacy if needed
Sending of the data to the institution over a secured and encrypted server

fig(1) - Overall design of the scrub system.


Phase -1
Identification of explicitly identifiable data in the dataset using search algorithms
Masking of explicitly identifiable data by removing them from the dataset
Cleaning of data by imputing missing values using regression techniques
Creating a new file with masked data

fig(2) - Phase 1 of the Scrubber.

In phase -1, as shown in fig(2), we are masking the explicit identiifiers on the result based on the search from a list of explicit identifers and then cleaning the data making it more useful as we are imputing the missing values using regression techniques


Phase -2
Identification of quasi-identifiers which may be a threat for re-identification using various parameters
Intimation of the analyzed quasi-identifiers to the data anonymyzer

fig(3) - Phase 2 of the Scrubber.

In phase-2, as shown in Fig(3), we are identifying the primary quasi-identiifiers on the result based on the search from a list of quasi identifers. We then calculate uniqueness and influennce of each oh these primary quasi-identifiers.

Uniqueness denotes how many unique values does a column contain, as more uniqueness will result in more risk of re-identification.

Influence of a column denotes how much does removing of that column affect the entire dataset as we need to maintain a good level of utitlity.

ADVANTAGES OVER PREVIOUS SYSTEM

With the gaps identified in the literature review we have resolved them, Where Basic search and replace method has been replaced with a search algorithm based on various parameters that lead to more accurate results. We are using regression techniques to classify the data that is missing in the dataset in the cleaning of the data. Our algorithm masks the direct identifiers and forms a new dataset so that these are not visible to the receiver of the data.

RESULT

INPUT:

The dataset taken in this paper is named as ‘ohit-ehr-payments-to-providers-july-2018.csv’ and Figure 1 shows top 26 values with some of the columns in the dataset

PROCESS: Initialling masking of explicit identifiers takes place, then the cleaning of data takes place with influence and uniqueness of the column. Consecutively, we are taking additional verification of the organization as inputs in order to modify our threshold risk and calculate results more accurately.

OUTPUT: Creation of output directory with masked.csv, Our source code creates a directory called Output_files and creates a csv file of the name given by the user in it.

In the phase-2 of the algorithm, the keywords given are the column names of the data which were analyzed as risky to the privacy of the data.

Now, the data anonymizer can easily use algorithms like k-anonymity on these columns to anonymize the data.

We are not anonymizing all the columns and hence, data can be of more use. This helps in keeping a good balance between privacy as well as utility of the data after anonymization.

CONCLUSION AND FUTURE WORK

Our project is mainly focused on the medical dataset. This would be very helpful for the health care institutions as they will no longer worry about supplying the patients’ information to the research institutions which calculate statistics for diseases and develop medicines. The scrubbing mechanism will ensure that no unnecessary information is supplied to the researcher and none of it gets into the hands of the people that can harm the patients. This system along with a safely encrypting system can be used to transfer data with complete security.

Slight modifications in this system can make it useful in other industries where there is a need for huge data transfers with privacy like educational institutions, voter’s records, census records, etc. Hence, this system is very useful for the maintaining the privacy of the society.

Various businesses can install this system to safely send their databases to their clients along with maintaining the privacy of their clients as well as the company.

REFERENCES

[1] **"Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression"**, L. Sweeney, 1996.

[2] **"Automated de-identification of free-text medical records"**, Neamatullah, I., Douglass, M.M., Lehman, L.H. et al. Computerized de-recognizable proof of free-content restorative records. BMC Med Inform Decis Mak 8, 32 (2008)

[3]**"Anonymization of Longitudinal Electronic Medical Records"**, Tamersoy and Loukides, 27 January 2012-IEEE.

[4]**"Security architecture for multi-site patient records research."**, Behlen and Johnson, 1999.

[5]**"MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring"**, Saeed M, Lieu C, Raber G, Mark, 25 Sept. 2002.-IEEE.

[6]**"Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies"**, Kushida, Nichols, Jadrnicek, Miller, Walsh, Griffin, July 2012.

[7]**"A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records"**, Rosario Catelli; Francesco Gargiulo; Valentina Casola; Giuseppe De Pietro; Hamido Fujita, 25 Jan. 2021.- IEEE.

LINKS:

- **Dataset:**https://data.chhs.ca.gov/dataset/cfaaae24-55b9-417e-89bb-eaf5a5318023/resource/41464f02-0a76-49f3-ae83-026f988eae3b/download/ohit-ehr-payments-to-providers-july-2018.csv
- https://ieeexplore.ieee.org/abstract/document/9335570
- https://ieeexplore.ieee.org/abstract/document/1166854
- https://ieeexplore.ieee.org/abstract/document/6140575
- https://pubmed.ncbi.nlm.nih.gov/8947683/
- https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf
- https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-8-32
- https://www.ncbi.nlm.nih.gov/pubmed/8947683
- https://www.blue-pencil.ca/data-cleansing-what-is-it-and-why-is-it-important/
- https://www.digitalvidya.com/blog/data-cleaning-techniques/
- https://towardsdatascience.com/machine-learning-for-data-cleaning-and-unification-b3213bbd18e
- https://www.dataquest.io/blog/machine-learning-preparing-data/
- https://techcrunch.com/2019/03/17/medical-health-data-leak/

  https://www.medicinenet.com/script/main/alphaidx.asp?p=a_dict
- https://www.researchgate.net/publication/348772529_A_Novel_COVID-19_Data_Set_and_an_Effective_Deep_Learning_Approach_for_the_De-Identification_of_Italian_Medical_Records