A NOVEL METHOD FOR LINKING EXISTING HEALTH-RELATED DATA AND MAINTAINING PARTICIPANT CONFIDENTIALITY

[1]Rupali S. Baviskar, [2]Swati S. Patil, [3]Vrushali S. Patil, [4]Mayuri D. Patil

Students
Computer Engineering Department
SSBT's College of Engineering and Technology, Jalgaon, Maharashtra

***Abstract*: Develop a Secure System for a record linking of existing individual health care information and maintaining participant confidentiality. Apply of individual health-related information faces manyproblems: Either asingular personal symbol, like Social Security variety, isn't on the marketor non-unique person recognizable info, like names, square measure privacy protected and cannot be accessed. An answer to shield privacy in probabilistic record linkages is to cipher this sensitive information and to beat these challenges, develop the Privacy Preserving Probabilistic Record Linkage (P3RL) methodology. The Privacy conserving Probabilistic Record Linkage methodology apply a tripartite protocol, with 2 sites assembling individual data associate degreed an freelance sure linkage center because the third partner.**

Keywords: Security, Encryption, Probabilistic record linkage,AES Algorithm

I. INTRODUCTION

Medical databases of individuals typically contain identifiers like surnames, given names, date of birth, and address data. The matter of finding records those represent an equivalent individual in separate databases while not revealing the identity of the people is termed "privacy-preserving record linkage". Initially, the plain answer for privacy-preserving record linkage looks to be the secret writing of the identifiers with a typical scientific discipline procedure. The aim is to explain a replacement methodology for the calculation of the similarity between two encrypted strings to be used in probabilistic record linkage procedures .Record linkage of existing individual health care information is Associate in Nursing economical thanks to answer necessaryepidemiological analysis queries. Reprocess of individual health-related information faces many problems: Either a singularpersonal symbol, like Social Security variety, isn't out there or non-unique person identifiable info, like names, square measure privacy protected and can't be accessed. An answer to guard privacy in probabilistic record linkages is to encipher these sensitive info. Therefore, customary secret writing strategies cannot be applied. To overcome these challenges, use a tendency to developed the Privacy protective Probabilistic Record Linkage (P3RL) technique.

II. LITERATURE SURVEY

The content of the paper focuseson the research and contributions ofvarious sources. The sources include:The paper describes theprocess for generation linkage table and secure the personnel information patient. The linkage table used for indexing the encrypted information.The paper describes theoperation of Genetic Algorithm. It alsogives an overview about functioning ofGenetic algorithm with AES algorithmand its internal operators. Among threeoperators, it gives a deep idea ofcrossover and row /column shuffling ofbits.

III. PROPOSED SYSTEM

The proposed system enhancesdata security of personnel information by providing the encryption and decryption.The proposed methodconsists of three main steps: pre-processing, encryption and probabilistic record linkage.Data pre-processing and encryption are done at the sites bylocal personnel. To guarantee similar quality and format of variables and identical encryption procedure at each site, the linkage center generates semi-automated pre-processing andencryption templates. To retrieve information (i.e. data structure) for the creation of templates without ever accessing plain person identifiable information, we introduced a novelmethod of data masking. Sensitive string variables are encrypted using Advanced EncryptionAlgorithm, which enables calculation of similarity coefficients. For date variables, develop special encryption procedures to handle the most common date errors. The linkage center performs probabilistic record linkage with encrypted person identical information and plain non-sensitive variables.

A. METHODS

1. Setting

Figure.1 offers a summary of the P3RL technique. P3RL is appropriate for settings wherever sites collect individual health-related knowledge on a similar persons while not a typical unique symbol (ID) and with rules limiting access to non-unique PII. The tendency to use a 2 site example (site A&B). In P3RL, probabilistic record linkage with encrypted non-unique PII (e.g. names, DOB, date of death [DOD], address) and plain linkage variables (e.g. gender, marital status, nationality) are used to combine the data from sites A and B. P3RL utilizes a trusted linkage center (site C) . The linkage center is an independent partner with stringent ethical guidelines and up-to-date privacy safeguards. It performs P3RL using only linkage variables (encrypted PII and plain demographic variables without health-related information). Site C is not involved in data collection or analyses, has no direct access to the individual records at site A or B and never sees PII in plain text. This is in keeping with Kelmanetal's best practice protocol . The output from P3RL is a link table containing only mapped site IDs (e.g. site A ID 1234 = site B ID 789). The link table is used to combine the records from the individual sites into a dataset tailored for the defined research objectives.

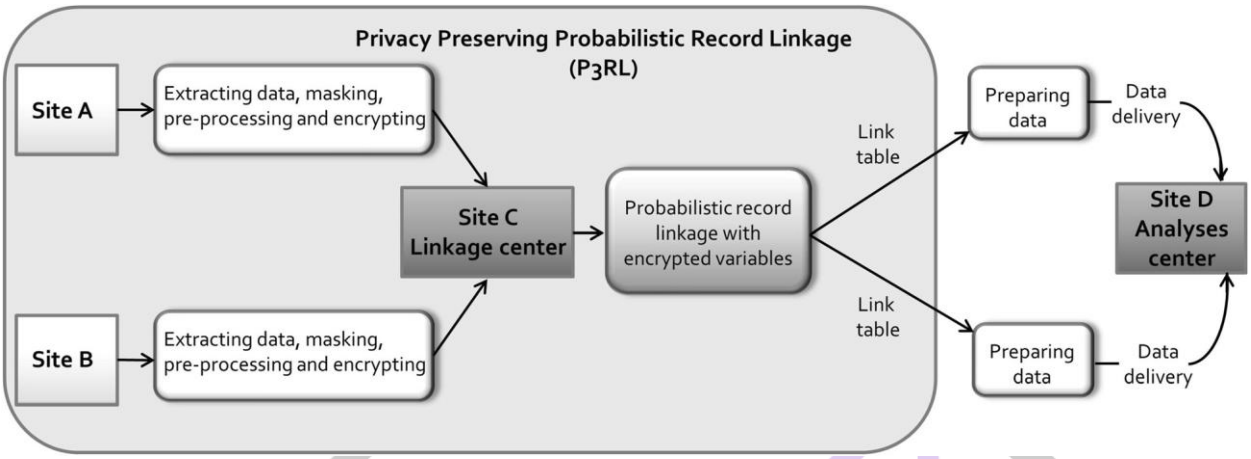
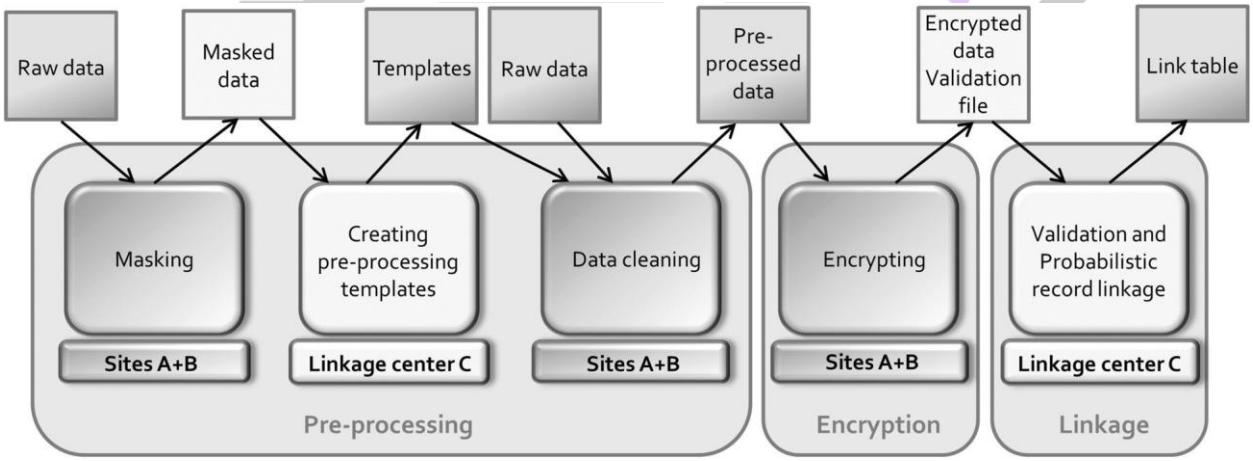


Figure 2 shows the flow of data between sites and thesites responsible for the individual steps included in P3RL method. P3RL consists of three main steps: preprocessing,encryption and probabilistic record linkage.Data pre-processing and encryption are done at the datacustodian sites (site A and B) by authorized localpersonnel. Creating pre-processing and encryption templates,encryption validation and probabilistic recordlinkage are done at the linkage site (site C).



2. Masking

Masking is used to disclose the individual site data structures to site C without revealing PII. The masked data are used to create the site-specific pre-processing templates. Data for building pre-processing templates are exported to site C as masked alone or masked and additionally shuffled depending on site restrictions.

3. preprocessing

Pre-processing is a crucial step in record linkage. The aim of pre-processing is to harmonize the linkage variables at each site to make them directly comparable and thuseasier to link. Pre-processing includes three steps: masking (site A and B), creatingpre-processing templates (site C) and data cleaning (site A and B).Using masking, thelinkage center creates custom pre-processing templatesbased on the data structure at each site. The templates are supplied to the individual sites allowing them to perform standardized data cleaning procedures that result in linkage variables with similar data quality and harmonized formats.The aim of masking is to alter the plain text variables

So, they are no longer readable. Masking replaces numeric characters between 1 and 9 with 9, lower case alpha characters a to z with z and upper case A to Z with Z. Some characters are left untouched. For example, first characters of fields, numeric character zero, special or language specific characters (e.g., -()) and spaces are unchanged. Masking of linkage variables is performed at the individual sites based on a pre-determined sample number of records or the entire population (depending on project-specific restrictions). Masking informs the data cleaning procedures by hinting at data errors, like numbers in name fields, characters in a numeric field or special codes for missing data (e.g., 9, 99, , .) and reveals language of text, number of names(surnamesfirst names) in a single variable, separators, special characters (e.g. language specific) and date formats.

4. Preprocessing template

The aim of using templates for pre-processing is to guarantee similar quality and format of linkage variables after numbers instead of names for month, types of delimitersand leading zero for numeric month and day. Although not particular to P3RL, checking the expected order of numeric day and month is not possible if both are less than 13.

5. Data Cleaning

Data cleaning is required because data from independent sources may differ in many aspects. For example, the format of variables may differ or string variables such as names can be inconsistent due to typographical errors, use of nicknames or abbreviations, changes due to marriage or pre- and postfixes. Therefore, the application of consistent data cleaning rules is crucial for any data warehouse generally and for record linkage particularly. In our P3RL workflow data cleaning is based on pre-processing templates and takes place at sites A and B, before encryption. This step is critical as non-pre-processed linkage variables result in a decreased linkage proportion because true matches are more frequently missed .

6. Encryption

The aim of encryption is to protect participant privacy and data confidentiality. Encryption is done at the individual sites using an automated encryption tool developedin-house specifically for our P3RL projects. All linkage variables deemed to be confidential (e.g. names, DOB) are encrypted while all other non-sensitive PII (e.g. marital status) are not. In this paper we focus solely on encryption of name and date variables. Nevertheless, the basic method of P3RL is applicable to other variable types. Levels of security and variables to be encrypted will differ from project to project.

7. Record linkage table

The last step in the P3RL method is probabilistic record linkage. However, before linkage begins site C must verify that the encryption was performed uniformly at site A and B. If the validation files from site A and B do not match the encryption must be redone before linkage can begin.

IV. RESULTS

All the results regarding enhanced data security with cryptography project is explained in Result. All the analysis is done by tester is explain further. The encryption is used to securely communicate data in open networks. As each type of data has its own structures, different techniques should be used to protect confidential data. The AES algorithm used for encryption in cryptography such as easily data can be retrieved using ASCII values for numerical representation. The pro-posed system combines cryptographic algorithm to protect data over network thereby enhancing the data security. The given information is related to heath data. First, each and every character of the
message is converted to the numerical value that is into ASCII value. The ASCII value for each character.The AES algorithm is used to encryption and decryption of heath related data for security purpose.

V. CONCLUSION AND FUTURE SCOPE

Privacy Preserving Probabilistic Record Linkage facilitates the linkage of existing data-setsin health related research settings using automated pre-processing and encrypting to fullyprotect personal identifying information.
Record linkage has a long tradition in both the statistical and the computer scienceliterature. The current survey approaches to the record linkage problem in a privacy-aware setting and contrast these with the more traditional literature and identify several importantopen questions that pertain to private record linkage from different perspectives. A novelmethod for linking an existing health related data and maintaining participant confidentiality
can put on Cloud in future and it can compulsory to do membership to PHR owner foruploading. data on server.

REFERENCES

[1] Alavi M, Law MG, Grebely J, Thein HH, Walter S, Amin J, et al. Lower lifeexpectancy among people with an HCV notification: a population-basedlinkage study. J Viral Hepat. 2014;21(6):e10–8.

[2] Bachteler T, Schnell R, Reiher J: An empirical comparison of approaches toapproximate string matching in private record linkage. Proceedings ofStatistics Canada Symposium 2010: Social Statistics: The Interplay among Censuses, Surveys and Administrative Data; 2011:290–295.

[3] Galhardas H, Florescuand D, Simon E, Shasha D. An Extensible Frameworkfor Data Cleaning. In: 16th International Conference on Data Engineering(ICDE): 2000; San Diego. 2000.

[4] Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approachto achieve private medical record linkage in light of public resources. J AmMed Inform Assoc. 2013;20(2):285–92.