VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Anonymization and Pseudo-Anonymization for E-Healthcare

Review-3

Slot: F1

Submitted by:

Name:	Registration Number:
Rakshith Sachdev	18BCI0109
Hrithik Ahuja	18BCE2154
Rohan Allen	18BCI0247

Abstract:

Electronic health records (EHR) contain data that can very easily identify and expose sensitive information about the users that malicious parties can exploit and use to their advantage. However, analysis must be performed in order to analyse these records and provide useful results and conclusions from the given records. Protecting the privacy of medical data is extremely vital given the sensitivity of this information. It can lead to severe consequences if the health data is compromised as it is extremely sensitive. At the same time preserving the utility of medical records is also necessary so that the patients health record can be used in surveys, analysis, etc to improve the quality of healthcare provided. Our project attempts to delicately balance both these priorities so that neither data privacy nor utility is compromised.

To accomplish this, anonymization methods are utilized. These strategies help accomplish utility while saving the security of the information. In the clinical business, speculation is the most well-known technique to accomplish this. In this undertaking we actualize an utility-protecting technique for the security saving information distributing (PPDP). The strategy is separated into three principle steps or classes. The first arrangements with the utility-safeguarding model. At that point we embed the fake records. At long last, the fake records are classified. This applies full space speculation. Past techniques like concealment and migration accompany the downside of not being versatile to huge datasets. With all the measurements, our proposed strategy shows a lower data misfortune than the current existing strategies while keeping up the utility.

Keywords: *Data Privacy; Utility-preserving; Data Anonymization; Grouping; k-anonymity; Medical privacy; privacy preserving data publishing (PDPP);*

Introduction:

Making electronic wellbeing records (EHRs) public to the majority may uncover touchy data and hence bargain the security and personality of a person. Typically, wellbeing records are anonymized before distributing, accordingly fulfilling protection models, for example, k -anonymity. Speculation is the most generally utilized anonymization calculation which prompts enormous data misfortune. Ensuring the protection of clinical information is very fundamental given the affectability of this data. It can prompt serious results in the event that it falls into some unacceptable hands. Simultaneously safeguarding the utility of clinical records is likewise fundamental with the goal that this data can be utilized in overviews, investigation, and so on to improve the nature of medical services gave. Our undertaking endeavours to carefully adjust both these needs so neither information protection nor utility is undermined.

Key Terms:

Anonymization:

It is a technique of deliberately modifying, supressing and falsifying data so that no person can be uniquely identified. It is generally performed to generalize data and to prevent outliers. It is an irreversible process. Completely destroys the original data. High data privacy but low data utility. Ex: Suppression, Perturbation, Generalization, k-anonymity, etc.

Pseudo-Anonymization:

It is the way toward subbing the personality of an information esteem with a reliable irregular worth. It very well may be turned around later on with the assistance of extra data. Low information security however high information utility. Ex: Tokenization.

Explicit Identifiers:

These are personally identifiable information like Aadhaar Number, Full Name, Passport Number, etc. This information is extremely sensitive and must be completely masked.

Quasi Identifiers:

These are touchy data which as one can be utilized to extraordinarily distinguish an individual. For Example: Address, Age, Date of Birth, Zip Code, and so on This must be anonymised to secure the person's protection however not totally concealed out to guarantee information utility.

Sensitive Data:

These are amazingly delicate and basic information that can be utilized to extraordinarily recognize an individual. For Example: Salary, Religion, Political Affiliation. This information must not be upset to much as it is imperative for information investigation. Some Pseudo-Anonymization methods can be utilized to substitute qualities and subsequently ensure the security of a person. Safeguarding the connection between the Quasi Identifiers and the Sensitive information is the most significant and troublesome undertaking to be accomplished.

Non-Sensitive Data:

This information is totally futile to the assailant. It isn't inclined to any information spillages, and so forth and need not be expressly ensured by any anonymization or pseudo-anonymization calculation.

Motivation:

There is a rising strain to disseminate singular patient records for optional purposes, for example, research and factual investigations. For instance, research financing associations are firmly promising proprietors of assets to give out information gathered by their undertakings. The normal favourable circumstances from sharing individual patient data for wellbeing research reasons include: it ensures there is responsibility in results and that detailed examination results are legitimate, it permits analysts and specialists to make/expand on crafted by different scientists all the more adequately and to perform singular patient data and information meta-investigation to sum up proof, and it decreases the issues on research subjects through the reuse of e effectively present patient records. At commonly, be that as it may, tolerant protection concerns have been considered as a key boundary for making singular patient information accessible for sharing and exploration.

Aim:

To build up another utility-saving anonymization calculation and to show that the utility of EHRs anonymized by the proposed technique is fundamentally in a way that is better than those anonymized by past methodologies.

Objectives:

To anonymize and protect EHRs

To preserve utility of EHRs

To design an algorithm to balance both privacy and utility of EHRs

To compare information loss between proposed and existing algorithm

Literature Review:

Paper 1:

Towards the Analysis of How Anonymization Affects Usefulness of Health Data in the Context of Machine Learning

Brief Description:

This paper utilizes various machine learning algorithms to model and predict health outcomes based on the ever-growing dataset of patients. It also uses multiple privacy preserving and anonymization techniques to ascertain the privacy of the patients sensitive and confidential health information before applying the machine learning algorithms. Obviously, anonymization of data leads to information loss. This affects the utility of the data. This in turn also affects the predictions and modelling of the machine learning algorithm. Henceforth this paper is proposed to analyze the intracricies among anonymization and data misfortune helped by data misfortune measurements. Henceforth, we can investigate the deficiencies of the current frameworks and plan to additionally diminish data misfortune with the goal that AI calculations can be prepared on these datasets and produce predictable outcomes a greater part of the time. This has a significant application in the field of clinical examination. Prescient displaying can be utilized to make drugs, antibodies, screen the strength and vitals of the patients in an amazingly effective way. This will enormously improve the life expectancy and personal satisfaction of individuals.

Algorithms Used:

k-anonymity, l-diversity, t-closeness, Classification Metric (CM), Information Loss Metric/Iyengars Loss Metric (ILM), Discernability Metric (DM), Minimal Distortion (MD).

Dataset Used:

Anonymized Electronic Health Records

Conclusion:

These Machine learning calculations can be utilized to anticipate wellbeing data significantly more precisely. This will go far in improving individuals' lives.

Paper 2:

Privacy Protection with Pseudonymization and Anonymization In a Health IoT System: Results from OCARIoT

Brief Description:

This paper expects to handle the different inadequacies and fragilities of wellbeing information in an Internet of Things (IOT) based framework. It joins various cryptography, anonymization and pseudo anonymization calculations and procedures. Every one of these techniques are applied to all the sorts of information for example Information being used, Data very still and Data moving. The OCARIoT application gathers information of younger students and their folks and is a unique application that interfaces with different substances, for example, instructors, guardians, medical care experts, innovation suppliers, the legislature, and so forth It is a multi-faceted apparatus used to log data about youngsters from each part of their lives. Ensuring this data is incredibly essential given the basic kid and wellbeing data. Information is anonymized and pseudo-anonymized likewise with the goal that lone the concerned specialists can sort out the data. For instance, educators ought not have the option to figure out a youngster's information as that would empower them to handily recognize the kid. This assignment is very dull given the interconnected idea of this IOT based framework. The paper focuses on that it is basic that all aspects of this IOT based framework is altogether examined to guarantee that protection and security principles are kept up. A break in any aspect of this environment will bargain the entire application given its interconnected nature. A profound security technique with complete comprehension of individual parts and their connections is an absolute necessity for OCARIoT to be a triumph.

Algorithms Used:

k-anonymity, substitution, permutation, differential privacy, Noise addition.

Dataset Used:

Anonymized and Pseudo-Anonymized Information of children and their parents.

Conclusion:

This application holds lots of valuable information, which if used wisely can improve children’s healthcare, education and their quality of life in general.

Paper 3:

Anonymization algorithm for security and confidentiality of health data set across social network

Brief Description:

Informal communication is the new typical these days. Huge measures of information are communicated through interpersonal organizations day by day. This paper proposes another protection saving calculation to anonymize the information so as to forestall re-recognizable proof just as safeguard the utility of information so it tends to be used in information investigation, information mining, and so forth This tale calculation called SOCnet means to save the protection of graphical information. (Long range informal communication Data can be pictured as far as a chart). The information included are the character of the individual, the connection or connections between individuals in the associated snare of individuals and the real substance or data. The character of the individual is totally concealed as this can be utilized to translate the first personality of the individual. The connections and connections between hubs are bunched together arbitrarily so as to anonymize them. At last, the information itself is anonymized and scrambled so no touchy data is undermined or available to an unapproved individual. Presently this information can be broke down and tried to reveal new and significant bits of knowledge that will upgrade the organization's net revenues over the long haul. They can utilize this information to improve their product and thus upgrade client fulfillment and they can likewise promote items as indicated by the client's determination. This information can likewise be offered to different organizations for their advantage. Eventually this security ensured information is an incredible asset which can be utilized to measure the interests of people in general and consequently utilize this to the organization's favorable position.

Algorithms Used:

Socnet, Random Perturbation, clustering.

Dataset Used:

Dataset of a social network.

Conclusion:

If the dataset is privacy-protected and anonymized properly companies can use this data to increase their profit margins without getting into legal troubles which will harm their reputation irreparably.

Paper 4:

A Tool for Optimizing De-identified Health Data for Use in Statistical Classification

Brief Description:

Various Biomedical Journals and Papers are distributed every year each containing new and creative patterns about the clinical practice and examination. The New England Journal of Medicine and The Lancelot are the most esteemed and widely acclaimed. In distributing information valuable to humankind, they additionally uncover datasets containing individuals' touchy data. This paper is distributed to elevate information anonymization and to extoll its protection benefits which are multifold. Simultaneously as well, much data ought not be lost with the goal that this information can be utilized in prescient displaying and examination. Data misfortune measurements and characterization measurements are likewise used to guarantee this. ARX an open source anonymization instrument for de-distinguished datasets is likewise reached out for building powerful measurable classifiers and furthermore dissecting their exhibition altogether. The fundamental target is that these measurable classifiers which are security ensured can be shared openly among clinical practioners and specialists for their utilization. First the informational indexes are anonymized utilizing k-obscurity. At that point the data misfortune is determined utilizing different measurements like Information Loss Metric/Iyengars Loss Metric (ILM). When the data misfortune is restricted to an adequate level information disclosure and prescient displaying are performed with the goal that wellbeing results can be anticipated incredibly precisely. For instance utilizing the patients vitals like circulatory strain, pulse and so on, this paper had the option to precisely decide the expense and term of a patients remain at a medical clinic.

Algorithms Used:

k-anonymity, Classification Metric (CM), Information Loss Metric/Iyengars Loss Metric (ILM), Discernability Metric (DM), Minimal Distortion (MD).

Dataset Used:

Patient Health Records

Conclusion:

This idea if implemented correctly by hospitals, will substantially boost their profits as they can use data to predicts the cost and duration of a patients stay without compromising their personal information.

Paper 5:

Privacy-Preserving Access Control in Electronic Health Record Linkage

Brief Description:

Trading Electronic Health Records with various Datasets is by and large encouraged on the grounds that it prompts better assessment and improved public care and clinical results.For occasion, interfacing wellbeing records with budgetary records to think about the associations among bulkiness and monetary status. This data assembled can be utilized for all around assessments which can be used by the lawmaking body to upgrade the money related achievement of its inhabitants. This paper presents the Semantic Linkage K-Anonymity (SLKA) computation which plays out the linkage of record of a component for instance just the patient beginning with one dataset then onto the following without compromising the security of the patient by anonymising the blends and record joins which are seen as unsafe as those are irregularities. The information of the patients are pseudo-anonymised by subbing the equivalent with unpredictable characteristics. The arranging exists and can be pulled back at whatever point required. These mappings are not appeared to the public plainly. Information adversity estimations are in like manner applied to calculate the information hardship and to ensure that those are saved to an adequate level. As the datasets are scattered and have a spot with different affiliations, ensuring their security and anonymizing them fittingly is the most noteworthy thing. This will ensure that no assailant can perform record-linkage attacks using his/her experience data.

Algorithms Used:

Semantic Linkage K-Anonymity (SLKA), k-anonymity, l-diversity, t-closeness, Classification Metric (CM), Information Loss Metric/Iyengars Loss Metric (ILM).

Dataset Used:

Datasets from Australia’s Diabetes Portal and their National Survey.

Conclusion:

If implemented properly this record linkage can be used by governments to enhance the welfare and socio-economic indicators of its population without compromising anyone’s personally identifiable information.

Paper 6:

E-Health Care Solutions Using Anonymization

Brief Description:

The information is put away into cloud by cloud clients to appreciate the great organizations, servers, administrations and applications from a mutual pool of configurable computing assets. Favourable circumstances of cloud computing omnipresent organization access, transaction of danger, area independent asset pooling. Sensitive information model individual medical records may must be scrambled by information proprietors before moving operations to the business public cloud to secure information protection and battle spontaneous gets to in the cloud and past. One of the security protecting procedures that control the data, making the information identification proof hard to anyone aside from the proprietors is anonymization. It is not quite the same as that of information encryption. Anonymization of information eliminates distinguishing ascribes like names or government backed retirement Numbers from the information base. The remarkable highlights are offered by the framework including effective key administration, particularly for recovery at crises, productive key administration and auditability, for abuse of health information. The proposed strategy looks to create security innovation and security ensuring foundations to encourage the advancement of a health data framework with the goal that people can effectively ensure their own data .

Algorithm used:

Cloud User (CU) , Cloud Service Provider (CSP) & Cloud Server (CS) , Third party Auditor (TPA) , K- Anonymity, L-Diversity, TCloseness, P-Sensitive and M-invariance

Dataset Used:

Data uploaded on cloud

Conclusion:

The utilization of cloud computing is a significant advancement on the world. Numerous IT organizations have begun utilizing cloud design in view of its pay-more only as costs arise idea. Security is one most significant factor that everybody thinks before transferring information's in the cloud. In this paper, the proposed is a mix of anonymization and encryption to empower making sure about of transitional datasets. A protection safeguarding looking through calculation is likewise proposed to recognize which halfway datasets are to be encoded instead of scrambling all the middle of the road datasets in order to empower cost

effective security. Any dataset that will be downloaded by the client, he ought to get confirmed by the TPA.

Paper 7:

A Methodology for the Pseudonymization of Medical Data

Brief description:

E-health empowers the sharing of patient-related information at whatever point and any place important. Electronic health records (EHRs) guarantee to improve correspondence between medical care suppliers, subsequently prompting better nature of patients therapy and diminished expenses. In any case, as profoundly sensitive patient data gives a promising objective to attackers and is likewise much of the time requested by insurance agencies and managers, there is expanding social and political weight with respect to the anticipation of medical information abuse. This business locales this issue and presents an approach that shields medical records from unapproved access and lets the patient as information proprietor choose who the approved people are, i.e., who the patient unveils her health data to. In this way, the approach forestalls information revelation that contrarily impacts the patient's life (e.g., by being denied medical coverage or work). The structure furnishes medical services suppliers with an exceptional arrangement that ensures information security (e.g., as per HIPAA) and permits essential and auxiliary utilization of the information simultaneously. The security examination demonstrated that the philosophy is secure and ensured against regular intruder situations.

Algorithm used:

Pommerening approach

Dataset used:

Patient Health Records

Conclusion:

This research utilizes a mix of theoretical investigative, relic building and artifact evaluating research draws near. The article begins with a point by point investigation of existing pseudonymization assurance systems, for example, encryption, anonymization and

pseudonymization, by looking at and breaking down related work (calculated scientific methodology). In view of these outcomes and the recognized weaknesses, a pseudonymization procedure is characterized and assessed by methods for a danger examination. At last, the exploration results are approved with the plan and execution of a model.

Paper 8:

Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption

Brief Description:

E-health permits better correspondence between medical care suppliers and higher accessibility of clinical information. Nonetheless, the disadvantage of interconnected frameworks is the expanded likelihood of unapproved admittance to profoundly delicate records that could bring about genuine oppression the patient. This article gives an outline of genuine protection dangers and presents a pseudonymization approach that safeguards the patient's security and information secrecy. It permits (direct consideration) essential utilization of clinical records by approved medical care suppliers and privacy preserving (non-direct consideration) auxiliary use by analysts. The arrangement additionally addresses the recognizing idea of hereditary information by expanding the fundamental pseudonymization approach with queryable encryption .

Algorithm used:

K-Anonymity , Query mechanism

Dataset used:

Dataset is referenced with pseudonyms

Conclusion:

Pseudonymization is a promising strategy to satisfy the necessities of information stockpiling and access for essential use just as security saving auxiliary use, yet when all is said in done requires an adequately enormous number of people and records to be powerful. We likewise need to pressure the way that fruitful pseudonymization (just as anonymization) requires dependable depersonalization, which can be very troublesome, if certainly feasible, for particular sorts of health information. Particularly information including hereditary data should be taken care of with unique consideration because of the recognizing nature.

Subsequently, we introduced a pseudonymization approach that is reasonable for pseudonymizing clinical records. Whenever required, the essential methodology can be reached out to deal with queryable (specific) record encryption to deal with hereditary information. For this situation, exceptionally delicate information parts can be scrambled and still save question usefulness, while depersonalized and enormous information, for example, clinical pictures, can be left decoded, however are as yet secured by pseudonymization.

Paper 9:

An IoT-Based Anonymous Function for Security and Privacy in Healthcare Sensor Networks

Brief Desciption:

Security has become a basic aspect of the present figuring world with respect to the universal idea of the IoT substances by and large and IoT-based medical care specifically. In this paper, research on the calculation for anonymizing sensitive data about health informational collection traded in the IoT climate utilizing a remote correspondence framework has been introduced. To safeguard the security and security, during the information meeting from the clients collaborating on the web, the calculation characterizes records that can't be uncovered by giving assurance to client's protection. Also, the proposed calculation incorporates a safe encryption measure that empowers health information namelessness. Moreover, the paper gave an investigation utilizing numerical capacities to legitimate the calculation's obscurity work. The outcomes show that the anonymization calculation ensures health highlights for the considered IoT framework applied in setting of the medical care correspondence frameworks.

Algorithm Used:

Wireless approach , Anonymization algorithm

Dataset used:

User health dataset used from HDS

Conclusion:

This paper proposed the improvement of a hypothetical methodology that guarantees the security and protection of sensitive information for the thought about IoT climate. The proposed calculation gave required security highlights, for example, protection or secretly for the client's information that is sent inside the medical care organization. At the point when the client sends his data to be utilized by the outsider by means of a given health organization,

the encryption cycle is initially executed utilizing a key from the key pair and the framework demand a reaction to the outsider where the anonymization work produces an incentive to anonymize the scrambled informational collection. In the work, we indicated that our proposed plot ensures the namelessness work where the calculation registers the conditions and afterward executes the anonymization technique on the medical care information. What's more, we showed that the calculation fulfills the computational intricacy necessities of the execution, all things considered.

Paper 10:

PAX: Using Pseudonymization and Anonymization to Protect Patients' Identities and Data in the Healthcare System

Brief description:

Electronic health record (EHR) frameworks are very helpful for dealing with patients information and are generally spread in the medical area. The primary issue with these frameworks is the means by which to keep up the security of sensitive patient data. Due to not completely shielding the records from unapproved clients, EHR frameworks neglect to give security to ensured health data. Feeble safety efforts likewise permit approved clients to surpass their particular benefits to get to clinical records. Accordingly, a portion of the frameworks are not a dependable source and are unfortunate for patients and medical care suppliers. Subsequently, an authorisation framework that gives protection while getting to patients information is needed to address these security issues. In particular, security and protection safety measures ought to be raised for explicit classes of clients, specialist guides, doctor analysts, crisis specialists, and patients' family members. As of now, these clients can break into the electronic frameworks and even abuse patients protection as a result of the benefits allowed to them or the lacking security and protection components of these frameworks. To address the security and protection issues related with explicit clients, we build up the Pseudonymization and Anonymization with the XACML (PAX) measured framework, which relies upon customer and worker applications. It gives a security answer for the protection issues and the issue of safe-access choices for patients information in the EHR. The aftereffects of hypothetical and trial security examination demonstrate that PAX gives security includes in safeguarding the protection of medical services clients and is sheltered against known assaults.

Algorithm Used:

Elliptical Curve Digital Signature Algorithm (ECDSA) , RSA encryption algorithm

Dataset used:

Patients dataset

Conclusion:

To guarantee the arrangement of security and protection, this paper proposes a PAX authorisation framework that underpins pseudonym, anonymity, and XACML. In particular, the proposed framework utilizes an irregular pseudonym to separate individual data about patients information, obscurity to conceal subjects' data, and XACML to make appropriated admittance control strategies to approve subjects solicitations to items records in HER. This paper accomplishes the security and protection safeguarding by using the alias namelessness methods, which can decrease the pointless time utilization and the weight on the worker. Security examinations utilizing the hypothetical strategy or formal strategies (BAN and AVISPA) exhibit that PAX is sheltered, keeps up the protection of medical services clients and mitigates the danger of entering contrasted with existing exploration. We accept that the PAX framework gives a security significant level that keeps up patients' security, and the framework particularly shields patients' data from backhanded clients, who have been viewed as a genuine security danger to any medical care framework since they can do inward assaults utilizing the benefits conceded to them.

Paper 11:

Centralized and Distributed Anonymization for High-Dimensional Healthcare Data

Brief Description:

Sharing medical information has become a significant prerequisite in medical services framework the board. Nonetheless, inappropriate sharing and utilization of clinical information can risk understanding protection. In this article, examinations have been done on the security worries of sharing patient data between the Hong Kong Red Cross Transfusion Service (BTS) and public clinics. It sums up information and protection necessities to the issues of incorporated anonymization and decentralized anonymization, recognizing the key difficulties that make conventional information anonymization strategies irrelevant. Also, the creators have proposed another security model called LKC protection to conquer the difficulties, and show two anonymization calculations that accomplish LKC security in both incorporated and decentralized situations. It gives a security answer for the protection issues and the issue of safe-access choices for patients' information. Examinations with genuine information show that the anonymization calculation can hold the significant data of unknown information for information investigation and is versatile to anonymize huge and multidimensional datasets.

Algorithms used:

LKC Privacy anonymization

Dataset used:

Patient dataset from the Hong Kong Red Cross Transfusion Service (BTS) and public hospitals.

Conclusion:

This application holds bunches of important data, which whenever utilized shrewdly can improve kids' medical services, training and their personal satisfaction by and large. On the off chance that the dataset is security ensured and anonymized appropriately organizations can utilize this information to build their overall revenues without getting into lawful inconveniences which will hurt their reputation irreparably. In particular, the proposed framework utilizes an arbitrary pseudonym to separate individual data about patients' information, anonymity to shroud subjects' data.

Paper 12:

Quantifying the costs and benefits of privacy-preserving health data publishing

Brief Description:

Cost benefit analysis is a need for settling on great business choices. In the business places, organizations need to make benefit from augmenting information utility of distributed information while having a commitment to secure individual protection. In this paper, the creators have measured the compromise among security and information utility in health information distributing as far as financial worth. The creators haves proposed a logical cost model that can help health information and data custodians (HICs) settle on better choices on sharing individual explicit health data with others. The creators have analyzed important cost factors related with the estimation of anonymised information and the conceivable harm cost because of potential protection breaks. The model proposed by this paper controls a HIC to locate the best benefit of distributing health data and could be used for both perturbative and nonperturbative anonymization techniques. The creators show that their methodology can distinguish the best an incentive for various security models, including K-secrecy, LKC-protection, and differential protection, under different anonymization calculations and security factors through legitimate tests on genuine information.

Algorithms used:

K-anonymity, LKC-privacy, and e-differential privacy

Dataset used:

Public patient datasets under the HIC pool

Conclusion:

On the off chance that the dataset is security ensured and anonymized appropriately organizations can utilize this information to build their overall revenues without getting into lawful inconveniences which will hurt their reputation irreparably.

Paper 13:

Publishing data from electronic health records while preserving privacy: A survey of algorithms

Brief Description:

The spread of Electronic Health Records (EHRs) can be exceptionally valuable for a wide scope of clinical investigations, crossing from clinical preliminaries and trials to pandemic control research, yet it must be acted such that saves patients' security. This paper examines the significant security dangers that information distributing involves just as dislikes the protection models that have been intended for information insurance. They additionally study more than 45 security calculations for distributing quiet explicit data. This paper talks about promising bearings for future investigation in information protection.

This isn't immediate on the line, on the grounds that the dispersed data should be secured against a few protection breaks, while staying helpful for resulting examination undertakings. In this paper, the creators have introduced a study of calculations that have been proposed for distributing organized patient information, in a security saving strategy. They audit in excess of 40 calculations, acquire experiences on their activity, and feature their favorable circumstances and hindrances. They likewise give a conversation of some encouraging bearings for future examination around there.

Algorithms used:

k-Anonymity, k-Mapping, r-Presence, *l*-Diversity, p-Sensitive-k-anonymity, rho-Uncertainty

t-Closeness

Dataset used:

User health dataset used from HDS

Conclusion

This paper proposed the improvement of a hypothetical methodology that guarantees the security and protection of delicate information. The proposed calculation gave required security highlights, for example, protection or privately for the client's information that is sent inside the medical services organization.

Paper 14:

Anonymizing Healthcare Records: A Study of Privacy Preserving Data Publishing Techniques

Brief Description:

The progressions in IT industry are pushing over the ordinary limits of the medical care organizations and associations to produce persistent data through different ways, for example, brilliant wearable devices, home checking gadgets, tele-health and so forth The amount of patient data produced can be utilized for the advancement of the network, clinical analysis, and other health related discoveries. All contemporary health administering offices and guidelines including ISO 22600-1:2014, Health Insurance Portability Accountability Act (HIPAA) and European Data Protection Act requires understanding agree so as to give out the clinical data to drug associations, protection offices, and exploration associations. Privacy Preserving Data Publishing (PPDP) strategies are generally utilized in anonymizing the patient data before distributing the patient information to these elements. This paper endeavors to show out an examination on the generally utilized PPDP techniques for anonymizing health information and talks about the latest patterns in the period of huge information. This exploration likewise gives some contribution on the relative examination of these strategies.

Algorithms used:

PPDP techniques for anonymizing health records

Dataset used:

Patient records from HIPAA association

Conclusion:

The article begins with a definite investigation of existing pseudonymization security systems, for example, encryption, anonymization and pseudonymization, by looking at and dissecting related work (theoretical systematic methodology). In view of these outcomes and the distinguished deficiencies, a pseudonymization approach is characterized and assessed by methods for a danger examination.

Paper 15:

Utility-preserving anonymization for health data publishing

Brief Description:

Distributing crude electronic health records (EHRs) might be thought of as a break of the protection of patients since they as a rule contain sensitive information. A typical technique utilized for the security safeguarding information distributing is to anonymize the records before distributing, and along these lines fulfill protection strategies and models, for example, k-obscurity. Among different anonymization techniques, speculation is the most broadly utilized in clinical/health records handling. Speculation unavoidably causes information misfortune and loss of utility, and along these lines, different procedures have been spoken about in this paper to decrease information misfortune. Be that as it may, existing generalization based health record anonymization methods can't dodge unreasonable information misfortune and save information utility.

The creators of this paper propose an information utility-safeguarding anonymization method for privacy preserving data publishing (PPDP). To help safeguard utility of information, the proposed method involves three sections: (1) utility-protecting module, (2) fake record addition, (3) index of the fake information. The creators of the paper additionally propose an anonymization calculation utilizing the proposed procedure. Their anonymization calculation applies full-area speculation calculation. They additionally assess their procedure in examination with presence strategies on two angles, information misfortune estimated through different quality measurements and blunder pace of investigation result.

Algorithms used:

Generalisation and K-anonymity of datasets

Dataset used:

Public medical records available for statistical use

Conclusion:

The creators of this paper propose another information utility-saving anonymization method and an anonymization calculation utilizing the proposed way. Through preliminaries and experimentations on different health records and public datasets, the creators have demonstrated that the utility of EHRs anonymized by the proposed procedure is path better than those anonymized by past methodologies.

Modules Used:

• Jupyter Notebooks
– Description: to store and run ipynb files.
• Python
– Description: programming language used to write the code
• Python libraries
– Various libraries used like pandas, numpy, sklearn etc.

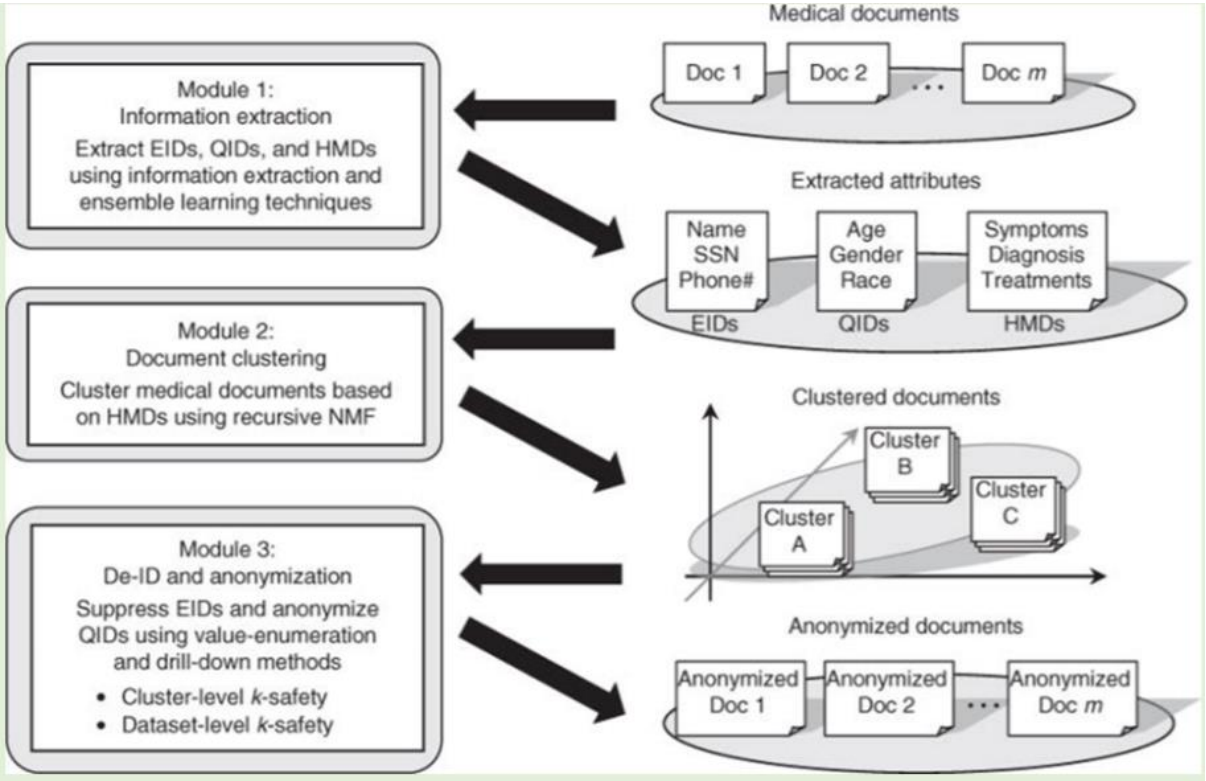
Contributions:

Paper 1-5: Rohan

Paper 6-10: Hrithik

Paper 11-15: Rakshith

FlowChart:



So the above shown flowchart describes the whole project in a nutshell showing the model one of the projects is information extraction so in this we extract the four types of information present in any data set. Four types of data Are explicit identifiers, Quasi identifiers, sensitive data and non-sensitive data. Non-sensitive data is irrelevant because it does not affect the privacy of the data set whereas the other three types of data have to be privatised. The explicit identifiers are information that is extremely sensitive and that can be used to uniquely identify a person therefore it is extremely crucial that this information is completely suppressed for example your name your age your other number are examples of explicit identifiers. then we have caused identifiers. This is an example of this is your age, your nationality, your health data etc, has to be anonymous rather than be suppressed. It has to be anonymized so that utility is preserved and that the linkage between the sensitive data and causing identifiers have to be maintained. This is achieved bi techniques like generalisation, k-anonymity etc. We can achieve this for example in generalization we can generalize the age so that instead of printing out the age as 25 we can give a range of say 20-30. This is extremely crucial so that each record will not be unique and will be somewhat similar which will promote and make it harder to re identify data and the level of privacy we have chosen for this project is 20. After all these changes are made in the final data set can be used for data analysis so that medicines, drugs can be produced and medical journals and research work can be conducted in a proper manner that will improve the health and livelihood of the general population.

Algorithm:

Algorithm 1: Anonymization Algorithm

Input : Original data O, Generalization rule G, Privacy parameter k, utility parameter h
Output: Anonymized data AT, Catalog for counterfeit records C

```
1  Create hierarchical lattice hl for all possible generalization cases, except for the case where the degree of generalization is more than h.
2  min = Maximum value of RCE
3  for each node n_i ∈ hl do
4      TempC = ∅, C = ∅
5      T̂* = generalization(O, n_i)
6      E_m ← list of equivalent class in T̂*
7      for j = 1 to |m| do
8          if |E_j| < k then
9              for j = 1 to |m| do
10                 addCounterfeitRecords(E_j, TempC);
11             end
12         end
13     end
14     C = Grouping(T̂*, TempC)
15     result = CalcuateRCE(T̂*, C)
16     if min > result && result ≠ null then
17         AT = T̂*
18         min = result
19     end
20 end
21 return AT and C
```

First according to this anonymization algorithm, we have to generalize the data and generalize all the data with respect to the privacy parameter k and utility parameter h. So first we go through all the records using a for loop. Find all the records which have unique attributes. If the number of records exceed 20 then they are generalized by broadening the quasi-identifiers and by adding counterfeit data. Utility parameter h is maintained by leaving the sensitive data as it is. For Ex. Making the age 20-20 instead of 23. This is called grouping into equivalence classes. The number 20 is chosen arbitrarily and can be changed whenever required. This twenty is the privacy parameter k. As now less than 20 records of the 900 or so total records are unique, it will be much tougher for an adversary to re-identify data.

To come to the literature review we can see the existing models and the methods that are used normally so generalization is normally used for privacy but this causes a lot of information loss as it generalizes the specific entries to a more broader and multi-use entry for example gender when we enter a male and female we can generalize it to male or female for all the entries and so this method is not very preferred as it decreases the utility and many of the existing techniques that are normally used for privacy preserving deal with mainly structured data and the current privacy methods used by the electronic health records normally focus on the detection of the patient identify us that is the values or entries that can be used to identify the patient using outside knowledge outsourcing knowledge etcetera and also the removal of the particular of those particular patient identifiers but when they remove the patient identify us the privacy might not be might be inadequate or in some cases the utility and quality won't be as good so now the so now the proposed models so to make it better we propose a we propose a utility preserving anonymization for privacy preserving data publishing that is a PPDP for short so to successfully preserve the utility of the data the proposed method has our proposed method has three parts the first ring the utility of preserving model that is the model. The main objective of aim of a project is to have maintained that the privacy and the utility of the data is maintained so that is the main aim and the problem with this is the fact that it's very hard to do this it's very tough and tedious task to achieve this so on one hand you have on one hand you have the patient whose personal personally identifiable information has to be a completely suppressed completely masked so that his you know his information is confidential so this is very important as health data is very sensitive and masking risk helps muskiness helps to safeguard is privacy safeguard is security and on the other hand the data can't be completely suppressed it can be completely masked as you know healthcare companies as pharmaceutical companies they require this health data to conduct research to conduct research to you know produce new drug supplies new vaccines to basically enhance the healthcare sector so this is very important that both aspects are met the privacy and the utility so um this is very tough to achieve and this is where the government comes in so the Health Ministry or the government should take to it that both parties are satisfied so all the stakeholders satisfied so like obviously the patient he wants maximum privacy he wants all his details to remain confidential and on the other hand you have the healthcare company of the pharmaceutical companies who want to utilize the data to the maximum do you know achieve their goals to make new medicines so this is a very delicate situation which has to be handled well so that you know all parties are satisfied so moving on yes so the objectives of the project like I said is to anonymize and protect EHRs to protect to preserve the utility of EHRs to discipline an algorithm to balance both the privacy and the utility so this is the main concept as I talked about so this is the main objective of our algorithm and obviously when this this algorithm will have some information loss and compare that to the original data.

Result Analysis with Code:

The work needed to be done is to set up a data set, with the end goal that it can later be utilized for AI purposes (for example classification, regression, clustering) without containing any non-utility data.

Our dataset contains the data of patients who are currently seeking medical treatment from certain hospitals. The columns contain serial number or patient ID, Insured or not, number of visitors allowed, Patient name, sex, Age, room number, bill in thousands, doctor reference number, condition whether is stable, critical or quarantined, etc.

PatientID	Insured	numVisitors	Name	Sex	Age	RoomNum	Bill (in thousand)	docRef	Condition
1	0	3	Braund, Mr. Owen Harris	male	22	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	373450	8.05		S

For our project we have used Jupiter notebook as our programming environment and have imported the required libraries for performing appropriate mathematical functions on columns to peruse regression, classification and clustering.

```
In [1]: import pandas as pd
        import numpy as np
        import scipy.stats
        %matplotlib inline
        import matplotlib.pyplot as plt
        from sklearn.preprocessing import LabelEncoder
        # get rid of warnings
        import warnings
        warnings.filterwarnings("ignore")
        # get more than one output per Jupyter cell
        from IPython.core.interactiveshell import InteractiveShell
        InteractiveShell.ast_node_interactivity = "all"
        # for functions we implement later
        from utils import best_fit_distribution
        from utils import plot_result

In [2]: df = pd.read_csv("health_data.csv")

In [3]: df.shape
        df.head()

Out[3]: (891, 10)

Out[3]:
```

	PatientID	Insured	numVisitors	Name	Sex	Age	RoomNum	Bill (in thousand)	docRef	Condition
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	373450	8.0500	NaN	S

Df.shape and head are commands used to print first 5 rows of the table.

```
In [4]: df.drop(columns=["PatientID", "Name"], inplace=True) # dropped because unique for every row
        df.drop(columns=["RoomNum", "docRef"], inplace=True) # dropped because almost unique for every row
        df.dropna(inplace=True)

In [5]: df.shape
        df.head()

Out[5]: (713, 6)
```

Out[5]:

	Insured	numVisitors	Sex	Age	Bill (in thousand)	Condition
0	0	3	male	22.0	7.2500	S
1	1	1	female	38.0	71.2833	C
2	1	3	female	26.0	7.9250	S
3	1	1	female	35.0	53.1000	S
4	0	3	male	35.0	8.0500	S

Here we have removed all the columns with private information that can directly be used to identify any patient like name, patientID, etc.

```
In [7]: encoders = [(["Sex"], LabelEncoder()), (["Condition"], LabelEncoder())]
        mapper = DataFrameMapper(encoders, df_out=True)
        new_cols = mapper.fit_transform(df.copy())
        df = pd.concat([df.drop(columns=["Sex", "Condition"]), new_cols], axis="columns")

In [8]: df.shape
        df.head()

Out[8]: (713, 6)
```

Out[8]:

	Insured	numVisitors	Age	Bill (in thousand)	Sex	Condition
0	0	3	22.0	7.2500	1	2
1	1	1	38.0	71.2833	0	0
2	1	3	26.0	7.9250	0	2
3	1	1	35.0	53.1000	0	2
4	0	3	35.0	8.0500	1	2

Here we have changed values of sex and condition column to 0,1 and 0,1,2 respectively so that person trying to identify won't know 0 represents male or female.

```
In [9]: df.nunique()

Out[9]: Insured                2
        numVisitors            3
        Age                   88
        Bill (in thousand)   220
        Sex                    2
        Condition              3
        dtype: int64

In [10]: categorical = []
         continuous = []

In [11]: for c in list(df):
             col = df[c]
             nunique = col.nunique()
             if nunique < 20:
                 categorical.append(c)
             else:
                 continuous.append(c)

In [12]: categorical

Out[12]: ['Insured', 'numVisitors', 'Sex', 'Condition']

In [13]: continuous

Out[13]: ['Age', 'Bill (in thousand)']
```

Here we check for number of unique values in each column to help us to classify the columns as categorical or continuous. All values above 20 unique items are classified as continuous and the rest are categorical.

```
In [14]: for c in categorical:
             counts = df[c].value_counts()
             np.random.choice(list(counts.index), p=(counts/len(df)).values, size=5)

Out[14]: array([0, 0, 0, 0, 0])

Out[14]: array([1, 3, 3, 3, 2])

Out[14]: array([1, 0, 0, 0, 1])

Out[14]: array([2, 1, 2, 1, 2])

In [15]: # https://stackoverflow.com/a/37616966/1820480

In [16]: best_distributions = []

In [17]: # for c in continuous:
         #     data = df[c]
         #     best_fit_name, best_fit_params = best_fit_distribution(data, 50)
         #     best_distributions.append((best_fit_name, best_fit_params))

In [18]: best_distributions

Out[18]: []

In [19]: best_distributions = [
             ('fisk', (11.744665309421649, -66.15529969956657, 94.73575225186589)),
             ('halfcauchy', (-5.537941926133496e-09, 17.86796415175786))]
```

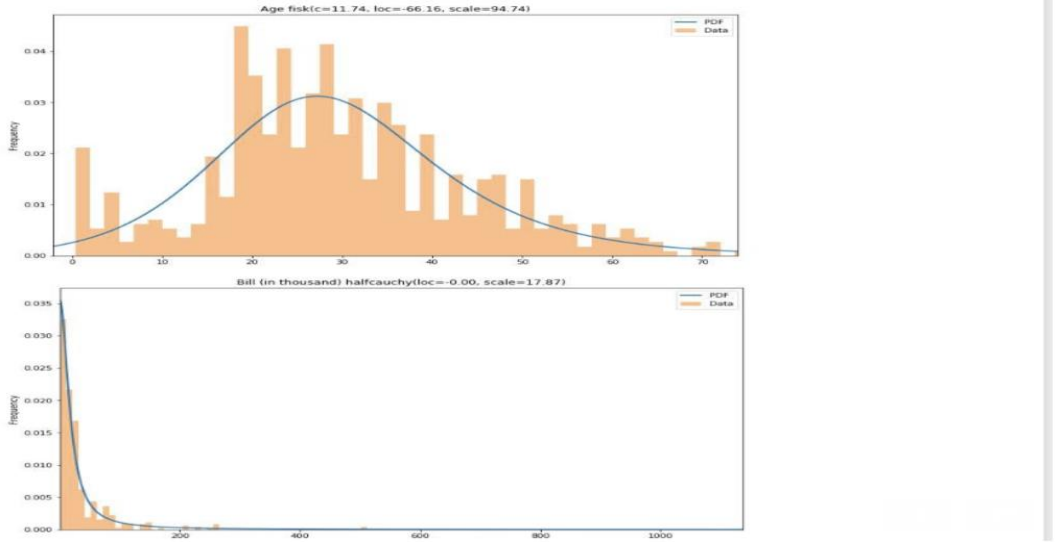
For every categorical value, we need to find out number of unique values, and then we find a concrete function to find number of values for each respective unique value. For every continuous value we find the best values for putting for each unique value. Using machine learning we will then use it to make new datasets or predict more values to create larger datasets for research purposes in the fields of data set training and statistical researches.

We then change the column names to 0,1,2 so that only the owner will know which column represents which values and it will be impossible for an external person to reidentify any certain individual or patient.

Hence are data becomes very secure yet the data does not lose its utility.

```
In [19]: best_distributions = [
             ('fisk', (11.744665309421649, -66.15529969956657, 94.73575225186589)),
             ('halfcauchy', (-5.537941926133496e-09, 17.867964151757866))]

In [20]: plot_result(df, continuous, best_distributions)
```



We plot the data in the graph and use a regression line through the graph to recognise a certain pattern which we can use to predict next set of values to increase our dataset or create new ones using a simple linear regression algorithm as used.

```
In [21]: def generate_like_df(df, categorical_cols, continuous_cols, best_distributions, n, seed=0):
             np.random.seed(seed)
             d = {}

             for c in categorical_cols:
                 counts = df[c].value_counts()
                 d[c] = np.random.choice(list(counts.index), p=(counts/len(df)).values, size=n)

             for c, bd in zip(continuous_cols, best_distributions):
                 dist = getattr(scipy.stats, bd[0])
                 d[c] = dist.rvs(size=n, *bd[1])

             return pd.DataFrame(d, columns=categorical_cols+continuous_cols)

In [22]: gendf = generate_like_df(df, categorical, continuous, best_distributions, n=100)

In [23]: gendf.shape
         gendf.head()

Out[23]: (100, 6)

Out[23]:
```

	Insured	numVisitors	Sex	Condition	Age	Bill (in thousand)
0	0	1	1	0	25.406552	9.474289
1	1	3	0	2	51.812626	11.859376
2	1	1	1	2	12.387505	19.327654
3	0	2	1	2	54.595218	43.251377
4	0	3	1	2	46.181993	10.322591

```
In [24]: gendf.columns = list(range(gendf.shape[1]))

In [25]: gendf.to_csv("output.csv", index_label="id")

In [26]: gendf.shape
         gendf.head()

Out[26]: (100, 6)

Out[26]:
```

	0	1	2	3	4	5
0	0	1	1	0	25.406552	9.474289
1	1	3	0	2	51.812626	11.859376
2	1	1	1	2	12.387505	19.327654
3	0	2	1	2	54.595218	43.251377
4	0	3	1	2	46.181993	10.322591

We clearly then store the new values and data created from the regression algorithm into new list or vectors and we check to test the accuracy of our new found or created data and we then print it using head and shape functions from sklearnpandas library.

We then save the new data in the new columns into a new output.csv file that is downloadable and can be used.

The only problem with this approach of ours is that once we changed everything by clustering and suppression like sex to 0 and changing column names, we cannot change it back to original form so the solution is to keep backups before starting this procedure. But the data is completely anonymised from quasi identifiers keeping sensitive data in check and secure.

References:

[1] F. Carmona, J. Conesa and J. Casas-Roma, "Towards the Analysis of How Anonymization Affects Usefulness of Health Data in the Context of Machine Learning," *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, Cordoba, Spain, 2019, pp. 604-608, doi: 10.1109/CBMS.2019.00126.

[2] S. L. Ribeiro and E. T. Nakamura, "Privacy Protection with Pseudonymization and Anonymization In a Health IoT System: Results from OCARIoT," *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece, 2019, pp. 904-908, doi: 10.1109/BIBE.2019.00169.

[3] N. Bruce and H. J. Lee, "Anonymization algorithm for security and confidentiality of health data set across social network," 2014 International Conference on Information and Communication Technology Convergence (ICTC), Busan, 2014, pp. 65-70, doi: 10.1109/ICTC.2014.6983085.

[4] F. Prasser, J. Eicher, R. Bild, H. Spengler and K. A. Kuhn, "A Tool for Optimizing Deidentified Health Data for Use in Statistical Classification," 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Thessaloniki, 2017, pp. 169-174, doi: 10.1109/CBMS.2017.105.

[5] Y. Lu, R. O. Sinnott, K. Verspoor and U. Parampalli, "Privacy-Preserving Access Control in Electronic Health Record Linkage," 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, 2018, pp. 1079-1090, doi: 10.1109/TrustCom/BigDataSE.2018.00151.

[6] C.K.Chu, S. M. Chow, W.G.Tzeng, J. Zhou andR. H. Deng, "KeyAggregate Cryptosystem for Scalable Data Sharing in Cloud Storage",Volume 25, Issue 2,pp. 1-11, 2014

[7] S. Märkle, K. Köchy, R. Tschirley, H.U. Lemke, The PREPaRe system—patient oriented access to the personal electronic medical record, in: Proceedings of the 17th International Congress and Exhibition on Computer Assisted Radiology and Surgery, ser. International Congress Series, no. 1256, 2001, pp. 849–854.

[8] Chaudry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S.C., Shekelle, P.G.: Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. Annals of Internal Medicine 144(10), 742–752 (2006)

[9] Höller, J.; Tsiatsis, V.; Mulligan, C.; Karnouskos, S.; Avesand, S.; Boyle, D. From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence; Elsevier: Amsterdam, The Netherlands, 2014.

[10] Calvillo-Arbizu, J.; Roman-Martinez, I.; Roa-Romero, L.M. Standardized access control mechanisms for protecting ISO 13606-based electronic health record systems. In

Proceedings of the 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, Spain, 1–4 June 2014; pp. 539–542.

[11] Adam, N. R. and Wortman, J. C. 1989. Centralized and Distributed Anonymization for High-Dimensional Healthcare Data. ACM Comput. Surv. 21, 4, 515--556. 10.1145/76894.76895

[12] Khokhar, Rashid Hussain, et al. "Quantifying the costs and benefits of privacypreserving health data publishing." 50 (2014): 107-121.

[13] Gkoulalas-Divanis, Aris; Loukides, Grigorios; Sun, Jimeng.” Publishing data from electronic health records while preserving privacy: A survey of algorithms”. Journal of Biomedical Informatics, ISSN: 1532-0464, Vol: 50, Page: 4-19

[14] Jayabalan, Manoj; Rana, Muhammad Ehsan. “Anonymizing Healthcare Records: A Study of Privacy Preserving Data Publishing Techniques”. Advanced Science Letters, Volume 24, Number 3, March 2018, pp. 1694-1697(4) 10.1166/asl.2018.11139

[15] Lee, H., Kim, S., Kim, J.W. et al. Utility-preserving anonymization for health data publishing. BMC Med Inform Decis Mak 17, 104 (2017). https://doi.org/10.1186/s12911017-0499-0