Data Publishing Techniques and Privacy Preserving

Ajit Singh
Patna Women's College
India

Abstract

What is Privacy? - According to the definition from Cambridge dictionary, "Someone's right to keep their personal matters and relationships secret". The term privacy is defined as an action where the data is kept hidden from either anonymous user, server to avoid use of malpractice of the data [2], [3]. Healthcare data is considered as a most significant but sensitive data in the world, since it has all private information about patient such as diseases, treatment, prescription, name, address etc. The volume of the data generated in healthcare industry is rapidly growing. In this patient centric world, to get effective results, we need to increase healthcare data utility. With increasing data utility, the privacy of the same data is compromised which is another important challenge that users and healthcare data publishers are facing, since there is no monitory control on data which is published on internet. Hiding sensitive healthcare data from either untrusted users or third-party data publishers is an important concern today. Healthcare Data Publishing is the process where certain transformation (such as anonymization, generalization, suppression etc.) can be applied before publishing healthcare data online. From the available research, it is seen that such transformations are not susceptible to certain attacks like background knowledge, homogeneity etc. This review paper studies all existing Privacy Preserving Data Publishing (PPDP) schemes using data generalization. The literature review also touches recent researches on ARX tool– which is an open source data de-identification tool for analyzing risk and utility factor of healthcare data. The paper finally concludes with feasible research gaps from available literature survey.

1. Introduction

According to the definition from Cambridge dictionary, someone's right to keep their personal matters and relationships secret [18], followed by the article 12 in universal declaration of human rights, No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks [2], [3], [18], [19].

Security and Privacy concerns of any type of data is major issue in today's technology-driven world. The term privacy is defined as an action where the data is kept hidden from either anonymous user, server to avoid use of malpractice of the data. In the world of distributed environment, the processing and storage of such multidimensional data is also done dynamically that too at different dynamic location keeping various transparencies in mind [2], [3]. In such scenario, privacy of data is important concern. Primarily there are some privacy techniques that are used to hide data viz.

Anonymization, Generalization, Perturbation, Role Based Access Control, Encryption etc [2], [3]. Typically, according to author [3], [5], Data has different phases during its lifecycle: storage of data, transition of data, transfer of data and processing of data. Existing privacy preserving techniques are still in maturing stages and strong privacy protection is still an open research problem [3].

Healthcare industry is one of the largest and rapidly developing industries. According to IBM Global Business Services, Executive Report, 2012 [5], the overall healthcare management is changing from disease centered to patient centered [23]. In the last few decades, there have been significant efforts in integrating information and communication technologies (ICT) into healthcare practices [3,]. E-Healthcare is an integration of latest technologies with medical infrastructure which includes the continues monitoring and transfer of health-related issues from patient-centric environment to respective services providers [4], [6]. The volume of the data generated in healthcare industry is rapidly growing. In this patient centric world, to get effective results, we need to increase healthcare data utility.

There has lot of research happened in maintaining privacy of healthcare in various areas viz. IoT-privacy based healthcare system, Machine learning in Healthcare data with privacy, maintaining privacy of healthcare big-data, privacy preserving cloud storage of healthcare data etc.

There are various advent technologies that have been used in healthcare domain. Machine learning algorithms for prediction of certain health parameter / data / diseases / behavior, IoT based Healthcare System [9], [10], [11], [12], [13], [14], [15], [16], Data Analytics in Healthcare Blockchain Implementation in healthcare, Cloud based healthcare systems etc.

Use of wearable technology used in healthcare used for continues patient monitoring, data streaming and sharing, use of data analysis to provide certain health services to the patients. It also uses patients' past records, treatment given by healthcare experts/doctors, prescriptions, allergic details of a patient etc. [7], [8], [9], [10].

Data Publishing is one of the state-of-the-art techniques of publishing healthcare data in tabular format on either public platform such as blogs/websites/public columns OR publishing selective data to selective people which can be only accessed based upon authentication. Since healthcare data contains sensitive information of individual, the privacy concern of the user must treat equally important. There are various ways to achieve privacy viz. Anonymization, Generalization, Suppression, Perturbation etc. before data publishing.

The proposed literature survey, examines the recent Privacy Preserving Data Publishing (PPDP) techniques in healthcare using data generalization. It also identifies the feasible research gaps of applying privacy techniques on healthcare data. The architecture has data tagging techniques invented in [17], [18], [19] and role-based access model for its users according to [20]. By properly authenticating the role of a user, the system displays appropriate healthcare dataset considering the importance of the data. The importance of a data is according to the data tagging system published in [4], [17], [18], [19].

The Structure of the paper is as follows. Section II discusses about the traditional utility and privacy tradeoff of healthcare data [5]. Section III focuses on privacy preserving data publishing (PPDP) techniques and various data format of healthcare data. It also explains the various possible attacks on datasets. The Section IV discusses about the use of ARX tool- an open source de-identification tool which can be used for simulating various existing/ proposed privacy preserving data publishing algorithms. It also explores the recent work done in data publishing using ARX tool. The section V and Section VI holds the major contribution to this paper which includes recent work done in data generalization and data publishing with the variety of dataset used. It also explores the result analysis based upon certain parameters of privacy preserving data publishing. Based upon section V and VI, paper concludes the limitations of the current research by listing few feasible research gaps in privacy preserving data publishing model.

2. Privacy and Utility

There is constant fight between healthcare researchers between data privacy and data utility. To analyze healthcare data using analytics or mining techniques, researchers expect real time datasets from healthcare institutes, organizations or private healthcare agencies. With the use of real time dataset, privacy of individual is always an important concern. There are certain privacy preserving techniques such as data anonymization compromise data utility to certain extent. In this case, data owner can sell his data at a nominal price if the dataset has all privacy measurements.

The central question regarding privacy and utility of data is: Can one pursue higher data utility while maintaining acceptable privacy?

Which extends a further question: "Since acceptable is qualitative term, how do you measure the privacy-utility parameter in quantity and how do you balance the tradeoff between these two? " Which is an important aspect of computing ?

The graphical representation of Data privacy and utility is shown in the figure 1 below.

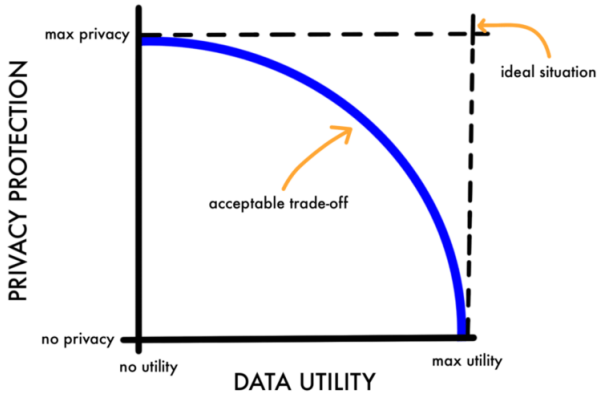

Figure 1. Data Utility- Privacy Tradeoff

3. Privacy Preserving Data Publishing (PPDP)

Healthcare data is considered as a most significant but sensitive data in the world, since it has all private information about patient such as diseases, treatment, prescription, name, address etc. Such dataset of any healthcare organization is susceptible by any attacks which are mentioned in [6]. Data Publishing is one of the state-of-the-art techniques of publishing healthcare data in tabular format on either public platform such as blogs/websites/public columns OR publishing selective data to selective people which can be only accessed based upon authentication. Data publishing is a two-step process viz. Data Collection and Data Publishing. Data collection is process of collecting individual patient information by hospitals, healthcare institutes, various medical departments through online/offline forms, portals and stores them into EHR. Patients and doctors are usually termed as data owners. Data publishing is a process of publishing data by hiding or suppressing some attributes. Any research institutes / organizations / government agencies/third party owners who wish to

do research on healthcare data can be considered as Data recipients. Figure. 2 drawn below show the data publishing process and actors which are involved in data publishing process.

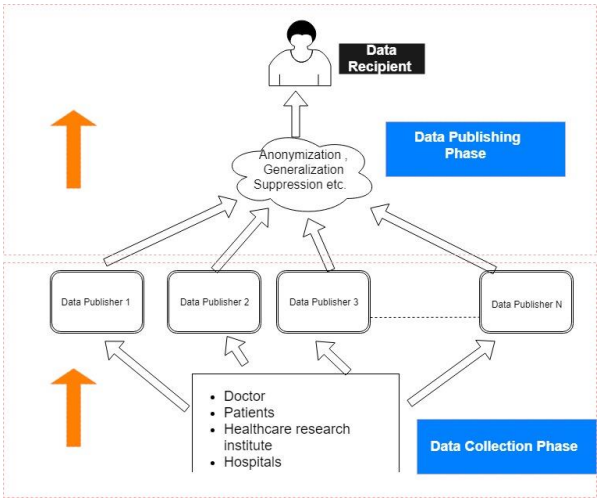

Figure 2. Data Publishing Process in Healthcare

Data publishing happens in two models viz. trusted and untrusted. In untrusted model, publisher attempts to update/manipulate certain sensitive information about the people who owns a data. Since it is a serious privacy issue, record owners can apply certain authentication mechanisms or cryptographic functions on already published data. On the other hand, in the trusted model, it is assumed that the data publisher is trustworthy and will not modify any sensitive information by any means.

As we know, the real time datasets can give sensitive information or could help adversary link records to certain individual, before data publishing some modifications are done. Modification may include:

1) Hiding certain data which is not required but contains sensitive information
2) Anonymizing the dataset by various techniques (listed in further sections of the report)
3) Generalizing and Suppressing certain data values
4) Anatomizing certain data values
5) Perturbing certain data values and many more.

A. Data Publishing Formats: Data is usually published in table format with some specific attributes. In data privacy, attributes are categorized into:

1) **Direct Identifiers (DI):** as the name suggests, individual patient can be directly identified without applying any reverse engineering methods. The example of direct identifiers can be name, email-id, address etc [5].
2) **Quasi Identifiers (QI):** certain attributes in dataset can be combined with other attribute/s which can uniquely identify an individual in the data publishing table. For example, individual gender as an attribute is very difficult to identify but combination of gender, sex, zip code, date of birth can be identified uniquely [4].
3) **Sensitive attributes (SA):** the attributes which supposed to be not published with an individual. The example includes genomic information, special diseases etc.

B. Attacks on Dataset: There are majorly two types of attacks in data publishing – viz. *record linkage attack* and *attribute linkage attack [4]*. Record Linkage attack makes use of quasi identifiers to recognize the value in the dataset. Whereas attribute linkage attack makes use of sensitive attributes which are major entity in the dataset. In such case attacker can easily identify based upon certain sensitive attributes record in the dataset.

In attribute linkage, Homogeneity and background knowledge attacks are the main types. Homogeneity attack happens when there is less diversity between the dataset values. Someone can create a model to leak information easily. In background knowledge attack, there are identification of data with either positive or negative disclosure. In positive disclosure, an adversary can correctly identify some value with the help of quasi identifiers with high probability [5, 6]. In case of negative disclosure, an adversary can correctly delete unwanted information from dataset with high probability. From the available research, background attack is more complex to prevent as compared to homogeneity attack.

C. Existing PPDP Techniques:

1) **K-Anonymity:** A table is said to be k-anonymous if it is indistinguishable from minimum (k-1) records with every quasi identifiers that it is selected. In order to implement this, the table is generalized based on the similar values of quasi identifiers [4]. Each group is then termed as equivalence class and has at least k records. The k –anonymity model introduces record linkage attacks [8], [9]. In [5], [3], [4], authors show that k-anonymity does not provide sufficient protection against attribute linkage [5].
2) **L-Diversity:** A k-anonymous equivalence class is said to be l-diverse if there are hminimum l values available for a particular sensitive attribute. L-diversity is the step which is taken beyond the shortcomings of k-anonymity. However, it is susceptible to attacks such as skewness and similarity attacks. As shown in [8], when the overall

distribution is skewed, satisfying the l-diversity does not prevent attribute linkage [8].

3) **T-Closemess:** A k-anonymous and l-diverse dataset is said to have t-closeness if the distance between the sensitive attribute in the equivalent class is maximum or less than some threshold value t [6], [7]. While implementing k-anonymity and l-diversity model data publisher must effectively choose the value of k and l. In t-closeness, the value "t" is just an abstract distance between two distributions, which could have different meanings in different contexts [8], [14]. Though t-closeness have several limitations which are documented in [6].

4. Arx Tool and Recent Researches

ARX tool is open source de-identification tool which used to analyze healthcare data. ARX makes use of the7information which is used during data de-identification process and it gives the various solutions which fits in their usage scenario [46]. Many biomedical scientists are using this simulation tool for data de-identification model for implementation of different privacy models.

Following are the recent researches of de-identification of healthcare data using ARX tool which is tabulated in Table I. Some of the research papers are also available in website of ARX.

Table 1. Recent researches on usage of arx on biomedical data

Citation, year and dataset used	Parameters used for evaluation	Description of research
[7], 2017, 32k U.S. Census records With Demographic attributes	• Population based risk • Dataset based risk • Safe harbor	• The objective of the research is to integrate the game theoretic model into ARX • Minimization of re-identification risks and minimization in reductions of data utility was the main objective of the research. • The game theoretic model typically outperforms HIPAA Safe Harbor • Implementation details are less important when a) A population table is used. b) The benefit of publishing is significantly higher than the potential loss
[7, 8], 2018, 1) *US Census (USC)*, an excerpt of records from the 1994 U.S. Census database which is often used for evaluating anonymization algorithms [8]. *2) Crash statistics (CS)*, a database about fatal traffic accidents, *3) Time use survey (TUS)*, a dataset consisting of responses to a survey on individual time use in the U.S. 4) *Health interviews (HI)*, a database of records from a survey on the health of the U.S. population [4, 5].	**Analysis of Anonymization Operator :** • Low generalization • Medium generalization • High generalization **Analysis of Optimization functions :** • Distribution of Quality • Probability of mass function • Distribution of output quality **Analysis of search strategy** : • Information Loss **Analysis of Quality of output** : • Precision • Loss • Discernibility • Entropy **Analysis of utility of output:** • Accuracy of output	• In this research, authors have presented a data publishing algorithm that satisfies the differential privacy model [14, 15]. • One of the features which the research has is: The transformations performed are truthful i.e. the dataset does not use any input or out perturbation of external data [63, 66]. Records are randomly selected from the given dataset which ensures that the unique feature of certain biomedical aspect remains hidden. • Authors have shown this by integrating six well-known data quality models listed in column 2. • One of the future research of this algorithm is that it is only applicable to low or medium dimensional data.
[7], 1) *US Census (USC)*, an excerpt of records from the 1994 U.S. Census database which is often used for evaluating anonymization algorithms [8]. *2) Crash statistics (CS)*, a database about fatal traffic accidents, *3) Time use survey (TUS)*, a dataset consisting of responses to a survey on individual time use in the U.S.	**Data Quality parameters :** • Execution Time • Entropy • Loss **Data Privacy Parameters** : • Low generalization • Medium generalization • High generalization	• The paper covers an improvisation approach of reducing re-identification attacks on healthcare data. • Authors have presented a method for implementing controlled data sharing environments and analyze its privacy properties. • Authors have also presented a de-identification method which is specifically suited for sanitizing health data which is to be shared in such environments [10, 11, 16]. • The basic idea which was aimed by author is to reduce the probability that a record in a dataset has characteristics which are unique within the underlying population [10, 11, 16].

4) *Health interviews (HI)*, a database of records from a survey on the health of the U.S. population [14, 15].		
[55], Patient Discharge Dataset	• Average Equivalence Class Size (AECS) • Discernibility • Precision • Loss • Ambiguity • Kullback-Leibler (K.-L.) Divergence • Non-Uniform (N.-U.) Entropy	• In this research authors have compared multiple de-identification and quality models on biomedical data. • Results of the experimental analysis of quality models shows that different models are best suited for specific applications • However from the result analysis, Non-Uniform Entropy is particularly well suited for general purpose use.
[1, 4], 1) 1994 US census database (ADULT), 2) 1998 KDD Cup (CUP), NHTSA crash statistics (FARS), 3) data from the American Time Use Survey (ATUS) 4) Integrated Health Interview Series (IHIS) [56].	• Average Equivalence Class Size (AECS) • Discernibility • Precision • Loss • Ambiguity	• In order to maintain the data quality as well as privacy, authors have described the **Lightning algorithm,** utility-driven heuristic search strategy implemented on ARX tool. • The authors have compared their work with existing heuristic based and globally optimal search algorithms. • The algorithm outperforms on several parameters listed in column 2.

5. Second and following pages

There are lots of improvisation happened on privacy preserving data generalization techniques. The table given below has existing data generalization techniques with applicable data and limitations. Total 11 generalization techniques have been studies and some limitations have been also listed in the table II.

Based on Table II, the parameters by which the techniques are covered are also being compared. Many of the cases, the parameters are not common; therefore, for better understanding only common and general parameters are taken into the considerations. The observations from the comparing parameters are also noted down in the section below.

The generalization techniques are listed according the chronological order of the starting year from 2002 to 2015. 5 generalization techniques have been implemented on categorical attributes whereas 6 generalization techniques have been implemented on numerical attributes. Corresponding drawbacks and attacks are also listed down from the following table.

Table 2. Existing data generalization techniques and their limitations

Techniques	Types of Attributes	Type of Recording	Drawbacks	Attacks covered
Generalization and suppression [2]	Categorical	Global	• Complexity issue • Individual record can be re-identified	• Temporal attack • Complementary attack
k-Minimal Generalization and Suppression [5]	Categorical	Global	• Not applicable to all set of data. Applicable only single dimensional data.	• Re-identification attack • Background knowledge attack.
Full domain generalization hierarchy [66]	Categorical	Global	• Data Utility is an issue • Highest information loss is observed.	Not addressed
Top- Down Specialization [7]	Categorical	Global	• Computational cost is more (CPU Time)	• Background knowledge attack.
Incognito [8]	Categorical	Global	• Distortion ratio is higher.	• Linkage attacks
Pre-defined generalization hierarchy [9]	Numerical	Local	• Results are not accurate in case of both i.e. categorical and numerical data.	Not addressed

Cell level generalization [7]	Numerical	Local	• Information loss is more.	Not addressed
Hierarchy free model [7]	Numerical	Local	• Requires pre-existing hierarchies.	Not addressed
OTF [2]	Numerical	Local	• Less information loss but privacy is compromised.	Not addressed
IOTF [3]	Numerical	Local	• Does not construct k-anonymous and l-diverse data [1].	Not addressed
DCHT [7]	Numerical	Local	• For larger size of dataset the productivity of the accurate results are not good [7, 8].	Not addressed

From the above techniques, the parameters by which the various methods are compared are shown in the table below. From the table III, the parameter "Information Loss" is common among all existing generalization techniques.

Table 3. Parameters of existing generalization techniques

Paper	Parameters
[22]	Not Addressed
[65]	Distortion Precision **Information Loss**
[66]	Minimum Absolute Distance, Minimum Relative Distance, Minimum Distribution, Minimum Suppression [6, 8].
[67]	Anonymity Error(AE), Baseline Error(BE), Upper Error(UE), Efficiency and Scalability, **Information Loss**
[68]	Performance based upon Varied Quasi Identifier Size , Performance based upon fixed quasi-identifier size and varied values of k [3, 5].
[69, 71]	Parameter based upon achieved optimizing the classification Matric - CM, **Parameter based upon achieved optimizing the general loss metric- LM [1, 5, 7].** Tree classifier errors using transformed data(using 10-fold cross validation) [7, 8].
[70]	The effects of crossover for different populations sizes [3], The effect of population size and duplicate values [3, 6].
[71]	Discrete Normal Distribution
[72]	**Information Loss Percentage**
[73]	**Information Loss Percentage**
[74]	**Information Loss Percentage**

The recent data generalization techniques that have been proposed are listed in table below [4], [5], [6], [7]. The objective of the data generalization from the available research is listed below:

- Preserving Privacy of the healthcare data
- Unbiased dataset values with less distortion ratio
- Maintaining quality of the data which can be useful for learning.

Table 4. Recent data generalization techniques

Paper and dataset used	Parameters used	Inference
[4], ADULT dataset (Real world dataset)	• Distortion Ratio • Maximum number of node at each level. • Discernibility Penalty	• In this paper, authors have proposed three different types of generalization techniques on ADULT dataset namely- CGH, DBGH, CBGH. The existing generalization techniques and their limitations are also discussed [1].

		• Conventional (CGH) method was observed ensuring privacy by limiting the number of nodes at each level as compared to the other generalization hierarchies [1]. • Cardinality-based generalization hierarchy (CBGH) appeared as having minimum distortion ratio, as well as Discernibility Penalty. Putting all these facts under consideration, it can be said that CBGH is best amongst all the newly proposed and existing techniques [1]. • **Intelligent mechanism to produce generalization hierarchies and set optimum ranges according to the dataset automatically [1].** • **Generating generalization hierarchies and sharing big data in the distributed framework [1].** • **Generating generalization hierarchies and set ranges in Textual datasets [1].**
[6, 8] US census dataset from the UC Irvine machine learning repository – ADULT Dataset [5, 9].	• Entropy Leakage • Privacy Leakage • Distribution Leakage	• The research targets privacy characterization and measurement issue in healthcare data publishing [4]. • The misjudgment of measuring privacy in the previous available literature is being rectified by introducing new analysis parameters: entropy leakage, privacy leakage and distribution leakage. • Those parameters are compared with most well-known PPDP techniques k-anonymity, l-diversity and t-closeness [5, 6]. • **Better optimization of data generalization techniques can improve the thrust of privacy**. • **Better optimization of the chosen set of quasi-identifiers with an objective of minimizing distribution and entropy leakages within the published table or specific classes of higher privacy concerns [10, 12].**
[5], Metro100k Dataset	• Utility Loss • Time	• In this research, author have proposed new lookup table brute force (LT-BF) mechanism which ensure privacy of individual user within the given trajectory. • Data quality based on *LKC*-privacy model in the scenarios which the generalization technique is applied to anonymize the trajectory data efficiently [27, 56]. • **The future work will be improvisation of algorithm to perform incremental data transformation [7]. (dynamic data addition into the dataset)**

6. Recent Researches on Data Publishing Schemes

There are various researches happened in past 2 year in case of data publishing, cited in [7], [8]. According to major researchers: There is no control over a data once it is published. Hence, some transformation needs to be applied to healthcare data (includes suppressing some values, anonymizing some values, apply generalization etc.). while applying transformation, the quality of data may be degraded, hence the recent research focuses upon the data quality as well as data privacy. The paper which was cited, are the current PPDP scheme-based papers where the methods that has been invented by researchers are being tested by data privacy parameters such as MRE, MAE, re-identification risk value as well as data quality parameter such as entropy, loss etc. Some other parameters like time, efficiency is also considered in the research.

Table 5. Recent researches on data publishing schemes

Paper and dataset used	Parameters used	Inference
[57] • Taxi Trajectory Prediction (Real time dataset),	• Mean Relative Error (MRE) [6]. • Mean Absolute Error (MAE) [6].	• Authors proposed distributed agent based privacy preserving framework, called DADP, that introduces a new level of multiple agents between the users and the untrusted distributed server [3, 5]. • Global w-event €-differential privacy mechanisms are used on crowd sourced distributed data to check the effectiveness of DADP.

• Bike System Company in Montreal (Nice ride) [4].		• **Only two parameters viz. MAE and MRE are considered for evaluation. In the future research the same method can be evaluated based upon some more parameters.**
[6, 8], US census dataset from the UC Irvine machine learning repository [5] – ADULT Dataset	• Entropy Leakage • Privacy Leakage • Distribution Leakage	• The research targets privacy characterization and measurement issue in healthcare data publishing. • The authors have proposed multi-variable privacy characterization and unique quantification model. • The misjudgment of measuring privacy in the previous available literature is being rectified by introducing new analysis parameters: entropy leakage, privacy leakage and distribution leakage. • Those parameters are compared with most well-known PPDP techniques [5, 6] : - k-anonymity, - l-diversity and - t-closeness [5] in healthcare dataset. • **Authors have opened the wide area of research whether only 2 matrices are sufficient enough to measure the privacy of data publishing schemes.** • **Better optimization of data generalization techniques can improve the thrust of privacy**. • **The optimization factor can be improvised with set of quasi identifier sets with minimum threshold of distribution and entropy leakage in the published table for achievement of better privacy**.
[60], dataset is not disclosed.	• Efficiency • Time	• In this research authors have proposed Personal Healthcare k-anonymity Encryption Model (PHKEM) which ensures the privacy on cloud data publishing. • The authors of this research have applied typical k-anonymity with EQI partitioning with AES encryption to preserve personal health records to prevent unauthorized access. • **Limited parameters are considered for evaluation of privacy and utility matrices on cloud** • **The comparison with other PPDP schemes can be done for better results and clarifications.**
[61], Retail Dataset , ADULT Dataset, TPC-E Dataset	**Efficiency parameters :** • Computation Time • Accuracy **Correlation Identification parameters :** • Dependency loss ratio • Complexity reduction ratio	• In this paper, authors have developed mechanism for maintain privacy and utility of high dimensional data. • The research done in [8] , had limitation of non-applicability of Safe-Pub algorithm to high dimensional data. • SVM and random forest classification is applied upon dataset, which shows average, 80% and 60% accuracy respectively.
[62] , POS, WVI, WV2 (Real time datasets)	Data Quality parameters : • Number and size of chunks Data performance parameters : • Time Data privacy parameters : • Average relative error	• In this research authors have given set of new algorithms for privacy oriented for set-valued data on cloud environment. From the available research and experimentation, it is clear that - existing privacy preserving techniques are not applicable to real case scenario in cloud. • In data publishing stage, data partitioning process is implemented (EQI partitioning) which uncouple certain record terms that participate in identifying combinations in the record [5, 6]. • In data querying stage, interactive differential privacy strategy is applied on the set valued data to ensure data is not retrieved using statistical queries. • **In the proposed work, Loss of information is not addresses in the paper, however the data utility is equally important while maintaining privacy in the cloud.**
[63], • World Cup dataset • ECML/PKDD: Taxi Trajectory Prediction dataset	• Mean Relative Error (MRE) • Mean Absolute Error (MAE)	• In this paper, authors have investigated the problem of real-time spatio-temporal data publishing in social networks with privacy preservation [5, 6]. • Its key components including adaptive sampling, adaptive budget allocation, dynamic grouping, perturbation and filtering, are seamlessly integrated as a whole to provide privacy-preserving statistics publishing on infinite time stamps [5, 6].

		• Experimental results show that the proposed schemes outperform the existing methods and improve the utility of real-time data sharing with strong privacy guarantee [5], [6].

7. Research Gaps in PPDP Schemes

Based upon the several research papers on privacy preserving data publishing and data generalization techniques [74], the following research gaps can be formulated. The list of research gaps in the chronological order is tabulated below:

A. User Based Categorization of Data: The existing data publishing algorithms focuses upon different categorization of the data based upon different level of generalization. The parameters like high, medium, low level generalization have been performed and tested to ensure that different kind of data can be visible to preserve its privacy.
The same problem can be looked upon on different view where user can be categorized based upon it's role and authenticity viz. role based access model (RABC). If the system allows us to check the credibility and authenticity of the data as well as user before data publishing, then the appropriateness of the PPDP can be maintained.

However, such work can go into the category of empirical kind of research where the new parameters are required to test the system. The parameters like loss, entropy etc. may not be sufficient to test the validity of the system.

B. Uniform Model for PPDP with Role based access model: From the available literature review, there is no uniform model which incorporates the level of data generalization and role of user simultaneously. Several data publishing techniques which have been studied are only based upon the type of data which is there in the healthcare repository.

C. Re-identification Attacks Countermeasures: Most of the data publishing techniques have still a good probability of re-identifying the particular tuple from healthcare dataset. No research fruitfully guarantees the re-identification attack will happen. Few of the research papers where privacy achieved is very high, has lots of information loss.

D. Maintaining variable privacy utility threshold depending upon the priority of health-care data: Privacy-Utility is always a concern in data publishing. Several latest literature surveys which are cited in this report have agreed on the fact that – both privacy and utility cannot be achieved with highest threshold. There is certain research where privacy parameters have outperformed with maximum information loss and vice versa. Maintaining trade-of between privacy and utility is the major challenge from available literature.

E. Computational complexity of data publishing algorithms: There are several literatures based upon the data publishing strategy have more computational complexity (CPU Cycles, Generalization time etc.). Some algorithms outperform better only if the size of dataset is small. Some algorithms are fruitful only for low or medium dimensional data. There are no fruitful schemes where multi-dimensional sparse data.

8. Conclusion

Privacy and Utility are the qualitative terms and one cannot only achieve privacy or utility by compromising other. In the patient centric world and in the era of ICT, the information of health is available everywhere. The technologies like Data Mining, Artificial Intelligence etc. made analysis of health data faster and reliable. On the other side, due to availability of personal health data the privacy of individual is always on risk since there is no monitory control on how to protect the data from untrusted entities. In Privacy preserving data publishing, through the data is transformed to some form before publishing in such a way that untrusted users couldn't easily identify an individual from available dataset table. However, PPDP schemes have still research gaps such as highest probability of re-identification risk, Computational complexity of existing data publishing algorithms, non-uniformity of the PPDP model, lack of user-based categorization of the data etc.

In near future, the research gaps mentioned in section VII can be addressed. To conclude, the paper touches all the aspect of privacy preserving data publishing techniques majorly in the healthcare industry.

9. References

[1] Definition of " privacy" from the Cambridge Academic Content Dictionary © Cambridge University Press.

[2] UN General Assembly, Universal Declaration of Human Rights, 10 December 1948, 217 A (III), available at: https:// www. refworld. org/ docid/3ae6b3712c.html [accessed 9 April 2019].

[3] Vora, J., DevMurari, P., Tanwar, S., Tyagi, S., Kumar, N., & Obaidat, M. S. (2018, July). Blind signatures based secured e-healthcare system. In 2018 International Conference on Computer, Information and Telecommunication Systems (CITS) (pp. 1-5). IEEE.

[4] Pussewalage, H. S. G., & Oleshchuk, V. A. (2016). Privacy preserving mechanisms for enforcing security and privacy requirements in E-health solutions. International Journal of Information Management, 36(6), 1161-1173].

[5] Burke, J. (2013). Health analytics: gaining the insights to transform health care (Vol. 71). John Wiley & Knowl.

[6] Senthilkumar, S. A., Rai, B. K., Meshram, A. A., Gunasekaran, A., & Chandrakumarmangalam, S. (2018). Big data in healthcare management: a review of literature. Am. J. Theor. Appl. Bus, 4, 57-69].

[7] Sharma, S., Chen, K., & Sheth, A. (2018). Toward practical privacy-preserving analytics for iot and cloud-based healthcare systems. IEEE Internet Computing, 22(2), 42-51.

[8] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352.

[9] Luo, E., Bhuiyan, M. Z. A., Wang, G., Rahman, M. A., Wu, J., & Atiquzzaman, M. (2018). Privacyprotector: Privacy-protected patient data collection in IoT-based healthcare systems. IEEE Communications Magazine, 56(2), 163-168.

[10] Elmisery, A. M., Rho, S., & Aborizka, M. (2017). A new computing environment for collective privacy protection from constrained healthcare devices to IoT cloud services. Cluster Computing, 1-28.

[11] Banerjee, S., Hemphill, T., & Longstreet, P. (2018). Wearable devices and healthcare: Data sharing and privacy. The Information Society, 34(1), 49-57.

[12] Dimitrov, D. V. (2016). Medical internet of things and big data in healthcare. Healthcare informatics research, 22(3), 156-163.

[13] Ould-Yahia, Y., Bouzefrane, S., & Boucheneb, H. (2018, April). Towards privacy and ownership preserving of outsourced health data in IoT-cloud context. In 2018 International Symposium on Programming and Systems (ISPS) (pp. 1-6). IEEE.

[14] Rahman, F., Bhuiyan, M. Z. A., & Ahamed, S. I. (2017). A privacy preserving framework for RFID based healthcare systems. Future generation computer systems, 72, 339-352.

[15] Yang, Y., Zheng, X., Guo, W., Liu, X., & Chang, V. (2019). Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. Information Sciences, 479, 567-592

[16] Tao, H., Bhuiyan, M. Z. A., Abdalla, A. N., Hassan, M. M., Zain, J. M., & Hayajneh, T. (2019). Secured data collection with hardware-based ciphers for iot-based healthcare. IEEE Internet of Things Journal, 6(1), 410-420.

[17] Sweeney, L., Crosas, M., & Bar-Sinai, M. (2015). Sharing sensitive data with confidence: The datatags system. Technology Science.

[18] Bar-Sinai, M., Sweeney, L., & Crosas, M. (2016, May). DataTags, data handling policy spaces and the tags language. In 2016 IEEE Security and Privacy Workshops (SPW) (pp. 1-8). IEEE [44].

[19] Baxter, R. (2017). Using DataTags to classify personal data under GDPR.

[20] Sandhu, R. (1998). Role-Based Access Control, Advanced in Computers. Academic Press., 46.

10. Acknowledgment

I am expressing my gratitude to everyone who helped me throughout this literature review paper. I am thankful for their encouragin support, invaluably constructive guidance and positive helpduring this work. I am sincerely grateful to them for sharing their valuable and research views on a number of issues related paper. I am also thankful to Retd. Professor Bal gangadhar Prasad, Patna University, for his support and encouragement and constantly supporting me throughout the research work.

