# Privacy-preserving healthcare informatics: a review

*Kah Meng* Chong[1*]

[1]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman (UTAR), 43000 Kajang, Selangor, Malaysia

**Abstract**. Electronic Health Record (*EHR*) is the key to an efficient healthcare service delivery system. The publication of healthcare data is highly beneficial to healthcare industries and government institutions to support a variety of medical and census research. However, healthcare data contains sensitive information of patients and the publication of such data could lead to unintended privacy disclosures. In this paper, we present a comprehensive survey of the state-of-the-art privacy-enhancing methods that ensure a secure healthcare data sharing environment. We focus on the recently proposed schemes based on data anonymization and differential privacy approaches in the protection of healthcare data privacy. We highlight the strengths and limitations of the two approaches and discussed some promising future research directions in this area.

## 1 Introduction

*EHR* systems are increasingly adopted as an important paradigm in healthcare industry to collect and store patient data, which include sensitive information such as demographic data, medical history, diagnosis code, medications, treatment plans, hospitalization records, insurance information, immunization dates, allergies and laboratory and test results. The availability of such big data has provided unprecedented opportunities to improve the efficiency and quality of healthcare services, particularly on improving the patient care outcomes and reducing medical costs. *EHR* data are published to allow useful analysis that are required by healthcare industries [1] and government institutions [2-3]. Some key examples may include large-scale statistical analytics (eg. study of correlation between diseases), clinical decision making, treatment optimization, clustering (eg. epidermics control) and census survey. Driven by the potential of *EHR* systems, a number of *EHR* repositories have been established, such as National Database for Autism Research (*NDAR*), *UK* Data Service, ClinicalTrials.gov and *UNC* Health Care (*UNCHC*).

Although the publication of *EHR* data is enormously beneficial, it could lead to unintended privacy disclosures. Many conventional cryptography technologies have been deployed to primarily protect the security of the *EHR* systems, such as access control, authentication and encryption. However, these technologies do not provide guarantee on privacy preservation. That is, the sensitive information of patient could still be inferred from the published data by an adversary. Various policies and guidelines are developed to restrict

*Corresponding author: kmchong@utar.edu.my

the type of publishable data and agreements on the usage and storage of data. For instance, *US* Health Insurance Portability and Accountability Act (*HIPAA*) [4-5], *EU* General Data Protection Regulation (*GDPR*) [6-7] and Personal Data Protection Act [8]. The limitations of this approach are: i) A high trust level is required on the data recipient that they follow the rules and regulations provided by the data publisher. Yet, there are adversaries who attempt to attack the published data to reidentify a target victim. ii) The sensitive data might be carelessly published due to human error and fall into the wrong hands, which eventually leads to the privacy breach of individual. Nevertheless, policies and governmental acts do not provide computational guarantee for preserving privacy of patient and thus cannot fully prevent such privacy violations. The need of protecting individual data privacy in a hostile environment while allowing accurate analysis on the patient has driven the development of effective privacy models in protecting healthcare data.

In this paper, we present the privacy issues in healthcare data publication and elaborate on relevant adversarial attack models. We focus on data anonymization and differential privacy and discuss the limitations and strengths of the proposed approaches. Finally, we conclude the paper and highlight the future research direction in this area.

## 2 Privacy threats

In this section, we first discuss privacy-preserving data publishing (*PPDP*) and the properties of healthcare data. Then, we present the major privacy disclosures in healthcare data publication and show the relevant attack models. Finally, we present the privacy and utility objective in *PPDP*.

### 2.1 Privacy-preserving data publishing

Privacy-Preserving Data Publishing (*PPDP*) provides technical solutions that address privacy and utility preservation challenges of data sharing scenarios. An overview of *PPDP* is shown in Fig. 1, which includes a general data collection and data publishing scenario. During the data collection phase, data of record owner (patient) are collected by the data holder (hospital) and stored as *EHR*. In the data publishing phase, the data holder releases the collected data to the data recipient (e.g. the public or a third party such as insurance industry and medical center) for further analysis and data mining task. However, some of the data recipients (adversary) are not honest and attempt to obtain more information about record owner beyond the published data, which includes the identity and sensitive data of record owner. Hence, *PPDP* serves as a vital phase that sanitizes personal sensitive information to avoid privacy violations.

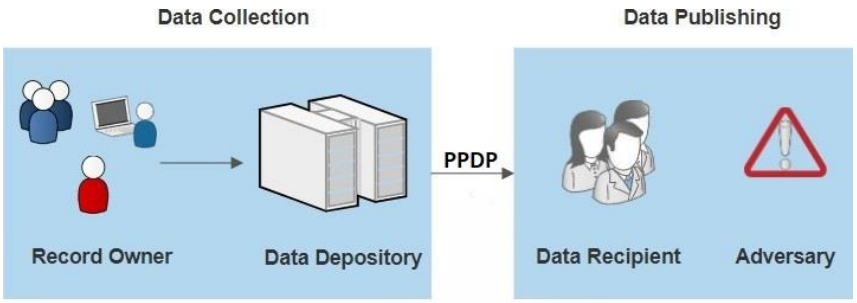


**Fig. 1.** Overview of *PPDP*

## 2.2 Healthcare data

Typically, healthcare data are relational data in tabular form. Each row (tuple) corresponds to one record owner and each column corresponds to a number of distinct attributes, which can be grouped into the following four categories:

- **Explicit identifier (*ID*):** It is a set of attributes that uniquely identifies a record owner, such as name, social security number, national IDs, mobile number and driving license number.
- **Quasi-identifier (*QID*):** It is a set of attributes that cannot uniquely identify a record owner, but potentially identify the target if combined with some auxiliary information. For example, date of birth, gender, address, zip code and hobby.
- **Sensitive attribute (*SA*):** It is a sensitive personal information that the record owner intends to keep private from unauthorized parties. Example includes diagnosis code, genomic information, salary, health condition, insurance information and relationship status.
- **Non-sensitive attribute (*NSA*):** It is an attribute that do not violate the privacy of the record owner if it is disclosed. All attributes that are not categorized as *ID*, *QID* and *SA* are classified as *NSA*. Examples include cookie IDs, hashed email address and mobile advertising IDs generated from EMR.

The attribute can be further divided into numerical attribute (e.g. age, zip code and date of birth) and non-numerical attribute (e.g. gender, job and disease). Table 1 shows an example dataset, in which the name of patients is naively anonymized (by removing the names and social security numbers).

**Table 1.** An example of different types of attributes in a relational table.

| Name | Quasi-identifier | | | Sensitive Attribute |
|---|---|---|---|---|
| | Age | Zip code | Gender | Disease |
| 1 | 23 | 96038 | Male | Diabetes |
| 2 | 28 | 96070 | Female | Diabetes |
| 3 | 26 | 96073 | Male | Diabetes |
| 4 | 37 | 96328 | Male | Cancer |
| 5 | 33 | 96319 | Female | Mental Illness |
| 6 | 33 | 96388 | Female | Diabetes |
| 7 | 43 | 96583 | Male | Diabetes |
| 8 | 49 | 96512 | Female | Cancer |
| 9 | 45 | 96590 | Male | Cancer |

## 2.3 Privacy disclosures

A privacy disclosure is defined as a disclosure of personal information that users intend to keep private from an entity which is not authorized to access or have the information. There are three types of privacy disclosures:

- **Identity disclosure:** Identity disclosure, also known as reidentification, is the major privacy threat in publishing healthcare data. It occurs when the true identity of a targeted victim is revealed by an adversary from the published data. In other words, an individual is reidentified when an adversary is able to map a record in the published data to its corresponding patient with high probability (record linkage).

For example, if an adversary possesses the information that *A* is 43 years old, then *A* is reidentified as record 7 in Table 1.

- **Attribute disclosure:** It occurs when an adversary successfully links a victim to their *SA* information in the published data with high probability (attribute linkage). This *SA* information could be a *SA* value (eg. disease in Table 1) or a range that contains the *SA* value (eg. medical cost range).
- **Membership disclosure:** It occurs when an adversary successfully infers the existence of a targeted victim in the published data with high probability. For example, the inference of an individual in a Covid-19-positive database poses a privacy threat to the individual.

### 2.4 Attack models

Privacy attacks could be launched by matching a published table containing sensitive information about the target victim with some external resources modelling the background knowledge of the attacker. For a successful attack, an adversary may require the following prior knowledge:

- **The published table, $T'$.** An adversary has access to the published table $T'$ (which is often an open resource) and knows that *T* is an anonymized data for some table T.
- ***QID* of a targeted victim.** An adversary possess partial or complete *QID* values about a target from any external resource and the values are accurate. This assumption is realistic as the *QID* information is easy to acquire from different sources including real- life inspection, external demographic data and voter list.
- **Knowledge about the distribution of the *SA* and *NSA* in table *T***. For example, an adversary may possess the information of *P*(disease=diabetes, age>50) and may utilize this knowledge to make additional inferences about records in the published table *T*.

Generally, privacy attacks could be launched due to the linkability properties of the *QID*. Now, we discuss the relevant privacy attack models for identity and attribute disclosure.

- **Linkage attack [9-13]**: adversary may reidentify the identity and discover the *SA* values of a targeted record owner by matching the auxiliary *QID* values with the published table $T'$. For example, if Table 1 is published without modification and suppose *A* possess the knowledge that *B* lives in zip code 96038, then *A* infers that *B* belongs to record 1 (identity disclosure) and has diabetes (attribute disclosure).
- **Homogeneity attack [9, 14]:** This attack discloses the *SA* values of a target when there is insufficient homogeneity in the *SA*. That is, the combination of *QID* is mapped to one *SA* value only. For example, suppose *A* knows that *B* is 28 years old, which belongs to the first equivalence class (an equivalence class is a cluster of records with the same *QID* values) in Table 2 (record 1, 2 and 3). Since these records have the same disease, *A* infers that *B* suffers from diabetes.
- **Background knowledge attack [9, 14]:** This attack utilizes logical reasoning and additional knowledge about a target to breach the *SA* values. For example, suppose *A* knows that *C* is 43 years old and lives in the zip code 96583, which belongs to the third equivalence class in Table 2 (record 7, 8 and 9). Nevertheless, the records show that *C* may have either diabetes or cancer. Based on *A*'s background knowledge that *C* is a person who likes sweet foods, *A* infers that *C* is diabetic.

- **Skewness attack [15]:** When the overall distribution of *SA* in the original data is skewed, *SA* values can be inferred. The *SA* values have different degrees of sensitivity. For instance, a victim may not mind being known as diabetic as it is a common (majority) disease. However, one would mind being known to have mental illness. According to Table 3, the probability of having mental illness is 33.3%, which is much higher than that of real distribution (11.1% in Table 1). Thus, this imposes a privacy threat that, anyone in the equivalence class have 33.3% possibility of being inferred to have mental illness, as compared with 11.1% of the overall distribution.
- **Similarity attack [14-15]:** This attack discloses *SA* values when the semantic relationship of distinct *SA* values in an equivalence class is close. For example, suppose that an adversary infers the possible salary of a target victim are 2K, 3K and 4K. Although the numbers represent distinct salary, they are all categorized in the range [2K,4K]. Hence, an adversary could infer that the target has low salary when the *SA* values are semantically similar.

**Table 2.** Published data of Table 1.

| No. | Quasi-identifier | | | Sensitive Attribute |
|---|---|---|---|---|
| | Age | Zip Code | Gender | Disease |
| 1 | < 30 | 960** | * | Diabetes |
| 2 | < 30 | 960** | * | Diabetes |
| 3 | < 30 | 960** | * | Diabetes |
| 4 | < 40 | 963** | * | Cancer |
| 5 | < 40 | 963** | * | Mental Illness |
| 6 | < 40 | 963** | * | Diabetes |
| 7 | < 50 | 965** | * | Diabetes |
| 8 | < 50 | 965** | * | Cancer |
| 9 | < 50 | 965** | * | Cancer |

**Table 3.** Another published data of Table 1.

| No. | Quasi-identifier | | | Sensitive Attribute |
|---|---|---|---|---|
| | Age | Zip Code | Gender | Disease |
| 1 | < 30 | 960** | * | Diabetes |
| 2 | < 30 | 960** | * | Mental Illness |
| 3 | < 30 | 960** | * | Cancer |
| 4 | < 40 | 963** | * | Cancer |
| 5 | < 40 | 963** | * | Mental Illness |
| 6 | < 40 | 963** | * | Diabetes |
| 7 | < 50 | 965** | * | Diabetes |
| 8 | < 50 | 965** | * | Cancer |
| 9 | < 50 | 965** | * | Mental Illness |

## 2.5 Privacy and utility objective of *PPDP*

*PPDP* allows computational guarantees on the prevention of privacy disclosures while maintaining the usefulness of the published data. From the privacy aspect, the identity of patients and their corresponding *SA* values should be concealed from the public. For instance, it is permissible to disclose the information that there exist diabetic patients in the hospital, but the published data should not disclose which patients have diabetes. Utility preservation is another aspect of *PPDP*, which emphasizes publishing data that is "almost similar" to the original data. Given that $M$ is an arbitrary data mining process, the output of $M(T)$ and $M(T')$

should be almost similar: the difference between $M(T)$ and $M(T')$ should be less than a threshold $t$. In most *PPDP* scenarios, the data mining process M (the usage of the published data) is unknown at the time of publication. This process $M$ could be a simple census statistic or some specify analysis and data exploration, such as pattern mining, association rules and data modelling. Privacy and utility are two contradictory aspects: publishing a high utility data implies less privacy protection to the record owner and vice versa.

## 3 Privacy models

In this section, we present some well-established privacy models that are used to ensure privacy in healthcare data. Particularly, we focus on data anonymization and differential privacy as two mainstream *PPDP* technologies which are different in their data publishing mechanisms.

### 3.1 Data anonymization

Fig. 2 shows a data publishing scenario in data anonymization. An original database is modified before being published as an anonymized database, which is generated by deploying generalization and suppression on the original database. The anonymized database could be studied in place of the original database. Some common data anonymization models to prevent privacy disclosure include $k$-anonymity [10-14], $l$-diversity [9], $t$-closeness [15] and $\delta$-presence [16].


**Fig. 2**. Data anonymization scenario.

a) ***k*-anonymity:** $k$-anonymity was developed to address identity disclosure. It requires that, for one record in the table that has some *QID* value, there exists at least $k$-1 other records in the table that have the same *QID* value. Hence, each record is indistinguishable from at least $k$-1 other records with respect to the *QID* value in a $k$-anonymous table. For example, Table 2 and 3 are 3-anonymous tables. In $k$-anonymity, any individual cannot be reidentified from the published data with a probability of higher than $1/k$. Other variations of $k$-anonymity include clustering anonymity [11], distribution-preserving $k$-anonymity [12], optimization-based $k$-anonymity [13], $\theta$ -sensitive-$k$-anonymity [14], *(X,Y)*-anonymity [17], *(α, k)*-anonymity [18], *LKC*-privacy [19] and random $k$-anonymous [20] which prevent identity disclosure by hiding the record of a target in an equivalence class of records with the same *QID* values. Although $k$-anonymity model protects against identity disclosure, it is vulnerable against attribute disclosure. Homogeneity attack and background knowledge attack is possible by deducing the sensitive attribute values from the published data. To provide protection on the sensitive attribute value, $l$-diversity and $t$-closeness were proposed.

b) ***l*-diversity:** $l$-diversity requires every *QID* group to contain at least $l$ distinct sensitive attribute values. For example, Table 3 is a 3-diverse table where there are at least 3 distinct sensitive attribute values for every *QID* group. This method depends on the range of the sensitive attribute values. If the number of distinct sensitive attribute values is lower than the desired privacy parameter $l$, some fictitious data are added to achieve $l$-diversity. This further

leads to excessive modification and may produce biased results in statistical analysis. In addition, *l*-diversity does not prevent attribute disclosure when the overall distribution of the sensitive attribute is skewed. Skewness attack and similarity attack are still possible to disclose the *SA* values in *l*-diversity. *k*-anonymity and *l*-diversity were combined to propose $\tau$-safe (*l, k*)-diversity [21].

c) ***t*-closeness:** To address these vulnerabilities, *t*-closeness was proposed, which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table. That is, the distance between the distributions is less than a threshold. This property prevents an adversary from making an accurate estimation of the sensitive attribute values and thus preventing attribute disclosure. However, only *SA* values are modified while all the *QID* values remain unchanged in this model. Hence, it does not prevent identity disclosure. Furthermore, *t*-closeness deployed brute-force approach to examine each possible partition of the table to find the optimal solution. This process takes an enormous computation time complexity of $2^{O(n)O(m)}$.

d) **$\delta$-presence:** To address membership disclosure, $\delta$-presence was proposed to limit the confidence level of an adversary in inferring the existence of a targeted victim in the published data to at most $\delta$%.

There is a significant amount of precedent for different parameter value of the privacy models, which could be used as benchmarks for efficient data publishing. However, the choice of the privacy parameter value is flexible and depends on the desired privacy and utility objectives of the data publication, provided that "an acceptable privacy level" is guaranteed.

### 3.2 Differential privacy

Fig. 3 shows a data publishing scenario in differential privacy. Differential privacy [22-25] involves a query answering process, which a data recipient may ask a query to the database and the result of that query is probabilistically indistinguishable regardless of the presence of a record in the database. That is, given two databases that differ in exactly one record, a differentially private mechanism provides two randomized outputs that have almost similar probability distributions. In other words, an adversary could not infer the existence of a targeted victim in the published database with high probability. Randomized noise derived from Laplace distribution is added to the result of the query to achieve privacy.

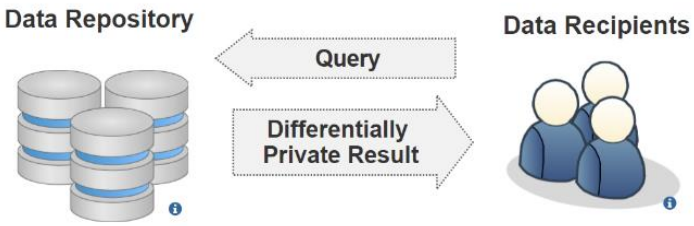


**Fig. 3.** Differential privacy scenario.

This is a stronger privacy-enhancing technique that addresses all privacy invulnerabilities data anonymization approaches and it makes no assumption about the background knowledge of any potential adversary. However, it has some privacy and utility limitations. Firstly, the original data could be estimated with high accuracy from repeated queries. If an adversary performs a series of repeated differential privacy queries (*k* times) on a published

database, then the original data could be disambiguated with high probability. Hence, Laplace noises must be injected $k$ times to guarantee that the published data is invulnerable against $k$ times of such queries. When $k$ is large, the utility of the published data is degraded significantly. In a differentially private database, a maximum of $q$ times queries is allowed to ask the database. This parameter $q$ is called the privacy budget. The privacy of a database cannot be guaranteed if more than $q$ times queries are made to the database. Thus, the database would stop answering further queries and provide no data utility after $q$ times of queries.

Differential privacy preserves utility for low-sensitivity queries such as counting, range and predicate queries, as the presence or absence of a single record changes the result slightly by one. However, a differentially private database could provide extremely inaccurate results for high-sensitivity queries. Examples of high-sensitivity queries include computation of sum, maximum, minimum, averages and correlation. Hence, a differentially private database is expected to provide highly biased results for more complex queries, such as computation of variance, skewness and kurtosis.

## 4 Conclusion

Although healthcare data provide enormous opportunities to various domains, preserving privacy in healthcare data still poses several unsolved privacy and utility challenges. In this paper, we have provided a general overview of healthcare data publishing problems and discussed the state-of-the-art in data anonymization and differential privacy. We highlighted the practical strengths and limitations of these two privacy-enhancing technologies.

For future research direction, it may be of interest to develop a standardization of privacy protection for privacy policy compliance as one of the subjects of future research.
Healthcare data holders are required to comply with a number of privacy policies to protect the privacy of a user. This may require the data holders to install systems and processes in place to maintain compliance. However, there is no clear indication of which privacy model and protection level should be adopted. In addition, what constitutes to "an acceptable privacy level" is not explicitly or clearly defined in any current privacy laws. Furthermore, it is of interest to design a privacy model that considers data publication in a distributed and dynamic environment, where there are multiple data holders who publish their data independently to a data pool with the possibility of data overlapping. The problem is on how to anonymize and analyze the aggregated data that consists of anonymized data from each publisher. Furthermore, data are collected and published continuously in a dynamic *EHR* system (such as wearable healthcare devices). The information contained in profiles could be updated from time to time and required to be reflected in the anonymized data.

## References

1. S. Senthilkumar, B.K. Rai, A.A. Meshram, A. Gunasekaran, S. Chandrakumarmangalam, Am. J. Theoret. Appl. Bus. **4**(2), 57 – 69 (2018)
2. M.A. Dudeck, T.C. Horan, K.D. Peterson, K.A. Bridson, G. Morrell, D.A. Pollock, J.R. Edwards, Am. J. Infect. Control **39**(10), 798–816 (2011)
3. K. Powell, Q. Li, C. Gross, K.A. Bridson, M. Dudeck, J. Edwards, S. Magill, *Ventilator-associated events reported by US hospitals to the National Healthcare Safety Network, 2015-2017*, in B48. Crit. Care: Meas. Measure-quality Improv. Implement. Best Pract., A3419 – A3419, (2019)
4. G. Cohen，M.M. Mello, J. AM. Med. Assoc. **320**(3), 231 – 232 (2018)

5. O. Obeng, S. Paul, *Understanding HIPAA compliance practice in healthcare organizations in a cultural context*, in AMCIS Proceedings of Information Security and Privacy (SIGSEC), 1 – 5 (2019)
6. P. Voigt, A.V.D. Bussche, The EU General Data Protection Regulation (GDPR): a practical guide, 1st ed. (Cham: Springer International Publishing, 2017)
7. C. T. Piri, A. Rohunen, J. Markkula, Comput. Law Secur. Rev. **34**(1), 134 – 153 (2018)
8. P. Carey, Data protection: a practical guide to UK and EU law, 5th ed. (UK: Oxford University Press, 2018)
9. A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, *l-diversity: privacy beyond k-anonymity*, in 22nd International Conference on Data Engineering (ICDE), 24 – 36 (2006)
10. L. Sweeney, Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**(5), 557 – 570 (2002)
11. F. Liu, T. Li, Secur. Commun. Netw. **5**, 1 – 8 (2018)
12. D. Wei, K.N. Ramamurthy, K.R. Varshney, Stat. Anal. Data Min. **11**(6), 253 – 270 (2018)
13. Y. Liang, R. Samavi, Comput. Secr. **93**, 1 – 18 (2020)
14. R. Khan, X. Tao, A. Anjum, T. Kanwal, A. Khan, C. Maple et al., Electronics **9**(5), 716, 1 – 24 (2020)
15. N. Li, T. Li, S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, in IEEE 23rd International Conference on Data Engineering, 106 – 115 (2007)
16. M.E. Nergiz, M. Atzori, C.W. Clifton, *Hiding the presence of individuals from shared databases*, in SIGMOD, 665 – 676 (2007)
17. K. Wang，B. Fung, *Anonymizing sequential releases*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 414 – 423 (2006)
18. R.C.W. Wong, J. Li, A.W.C. Fu, K. Wang, *(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 754 – 759, (2006).
19. N. Mohammed, B. Fung, P.C. Hung, C.K. Lee, *Anonymizing healthcare data: a case study on the blood transfusion service*, in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1285 – 1294 (2009)
20. F. Song, T. Ma, Y. Tian, M.A. Rodhaan, IEEE Access **7**, 75434 – 75445 (2019)
21. H. Zhu, H.B. Liang, L. Zhao, D.Y. Peng, L. Xiong, IEEE Access **7**, 687 – 701 (2018)
22. C. Dwork, *Differential privacy: A survey of results*, in International Conference on Theory and Applications of Models of Computation, 1 – 19 (2008)
23. A. Alnemari, C.J. Romanowski, R.K. Raj, *An adaptive differential privacy algorithm for range queries over healthcare data*, in IEEE International Conference on Healthcare Informatics, 397 – 402 (2017)
24. H. Li, Y. Dai, X. Lin, *Efficient e-health data release with consistency guarantee under differential privacy*, in 17th International Conference on E-health Networking, Application & Services, 602 – 608 (2015)

25. O. Gutierrez, J.J. Saavedra, M. Zurbaran, A. Salazar, P.M. Wightman, *User-centered differential privacy mechanisms for electronic medical records*, in International Carnahan Conference on Security Technology, 1 – 5 (2018)