Studying Genotype-Phenotype Attack on k-anonymised Medical and Genomic Data

Muzammil M. Baig [1] **Jiuyong Li** [1] **Jixue Liu** [1] **Hua Wang** [2]

[1] School of Computer and Information Science
University of South Australia,
Mawson Lakes, South Australia 5095,
Email: {muzammil.baig jiuyong.li jixue.liu}@unisa.edu.au

[2] Department of Maths and Computing
University of Southern Queensland ,
Toowoomba, Queensland, 4350,
Email: wang@usq.edu.au

Abstract

Personal data of patients is largely collected at hospitals, clinics, labs etc. This data consists of medical and genomic record. Such patient data is shared for various health and research purposes. The utility of such sharing is worthwhile and its benefits are now well documented. It includes early diagnostic of some diseases like Phenylketonuria that can cause high chances of recovery. Population health analysis, derived from collaborative sharing of patient data, help government agencies to draft proper policies to raise the standard of living of people. On the other side of picture, many patients fear about the misuse of their personal data. This fear caused social (sometimes legal) requirement to properly safeguard the personal data before sharing. Various generalization techniques were suggested to anonymize the both types of patient data i.e. medical and genomic. Generalization based privacy protection technique, k-anonymity is considered to be one of important practices to anonymize the patient data. Due to rapid technological advancements, it is possible that the medical and genomic data of same patient(s) can be publically available from different sources. Such a scenario has created new privacy threats to patient data. Genotype-Phenotype attack is one of these threats. This research paper shows that how k-anonymised medical and genomic data is subject to genotype-phenotype attack.

Keywords: k-anonymity, genotype-phenotype attack, privacy compromise.

1 Introduction

Health organizations largely collect the personal information from patients. The range of this personal information has changed with time. Anciently it consists of few personal particulars like name, date of birth, address. Technological advancements added new particulars and dimensions to patient personal information. Presently, patient information is divided into two major groups i.e. medical information and genomic information. Medical information is the set of those attributes that is directly asked from patient like name, date of birth, address, nationality etc. Moreover it also includes the medical state of patient like disease information and results of various laboratory test(s). Genomic information consists of DNA sequence of the patient. Healthcare organizations now can store gigantic medical and genomic information due to decreasing cost of archiving and DNA sequencing.

Large scale storing of medical and genomic information is very useful for research purposes. It provides bio-medical research community the opportunities that were severely limited or nonexistent before. The analysis of medical information through modern data mining techniques helps the researchers to come-up with interesting facts. It can lead to find out that which population is prone to which disease. The genomic data, also called DNA sequence, is the blueprint of species. The study of genomic data can help to find out the reasons behind analytical results of medical information. Genomic data is a precious resource for biomedical research. It helps medical researchers to understand causes of diseases and to find effective ways to cure the diseases. For example, the evolution of medical paradigms like personalized health care is a result of genomic analysis (M.West et al. 2006). Other examples include the discovery of individuals genotype influence on the metabolism of pharmaceuticals (Roses 2000, Burnett et al. 2003), and the formulation of new medicines based on subsequent results. Therefore, sharing medical and genomic data is crucial for modern biomedical research.

Patient medical and genomic data may be maliciously used while it is very helpful for biomedical research. Biological research has shown that the genomic data holds identity and sensitive information of an individual, such as gender and disease traits (Altschul 1990, Caenazzo 1998). Employers and insurance companies may utilize such private information for discrimination. For example, employers may refuse job to someone who has the trait of mental illness. Insurance companies may charge higher premium for some individuals. Due to this and many other reasons many people fear that information gleaned from their genomic data will be misused, abused, to influence their employment and insurance status, or simply cause social stigma (Clayton 2003, Xiao & Tao 2007, Hall & Rich 2000, Rothstein 1997). Genomic data, unlike most medical information, contains more specific information that can relate family

Copyright ©2009, Australian Computer Society, Inc. This paper appeared at the Eighth Australasian Data Mining Conference (AusDM 2009), Melbourne, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 101, Paul J. Kennedy, Kok-Leong Ong and Peter Christen, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Table 1: Possible cases of 'Match Records'

	Anonymised Medical Data				Anonymised Genomic Data		
	Quasi Identifiers			Sensitive Attributes	Quasi Identifiers		Sensitive Attributes
Case #	**Age**	**Sex**	**Zip**	**Phenotype**	**Age**	**Sex**	**Genotype**
1	25 – 35	Male	500*	A	25 – 35	Male	A
2	25 – 35	Male	500*	B	35 – 40	Male	B
3	25 – 35	Male	500*	C	25 – 30	Male	C
4	25 – 35	Male	500*	D	40 – 45	Female	D

members too. Due to the sensitive nature of medical and genomic data there are many social pressures to protect the privacy of patients data.

Due to the serious nature of such issues, US Government adopted the Health Insurance Portability and Accountability Act (HIPA) (DHHS 2002) and Genetic Information Non-discrimination Act (GINA) to restrict the access of medical/genomic data by employers and insurance companies (GINA 2008). In the European Union, the Data Protection Act (DPA) of 1998 imparts strict regulations on the processing of personal data, of which genomic data is a part of (NHS 2000). So, if there is no proper assurance of privacy preservation, not only patients will be hesitant to provide but many data holders will be unwilling to share patient data. Such a situation can be harmful for new or continuing bio-medical research. In recognition of the problem, the protection of privacy for patient data is considered a major challenge for the biomedical research community (Altman & Klein 2002, Vaszar et al. 2003).

Realizing the need of privacy protection of patient information many researchers presented their research work to preserve the patients privacy both in medical and genomic data. Initially lot of research work purposed in the domain of medical data and lot of anonymization techniques were purposed like k-anonymity (Sweeny 2002), L-diversity (Machanavajjhala et al. 2007), m-Invariance (Xiao & Tao 2007), (alpha,k) (Wong et al. 2006). Subsequently ample developments in DNA sequencing caused ease accessibility to genomic databases. There are currently more than 100 genomic databases available online over the World Wide Web (Michael & Galperin 2007); most of them are freely available. Such huge availability of genomic data also raised the need of privacy protection. It motivated the researchers to purpose the techniques to preserve the privacy of genomic data like familial data protection (Gaudet et al. 1999), gene trustee (Burnett et al. 2003), generalization lattices (Malin 2005*b*).

Existing medical and genomic data anonymization techniques were designed without considering the availability of both types of data of the same patient. It is possible that two or more locations can publically release the patient data of same patient (Malin & Sweeney 2004). Such situation caused genotype-phenotype attack, discussed in detail in section 3. Generalization based techniques like k-anonymity (Sweeny 2002) are largely discussed in literature. In this paper we proved that k-anonymization technique for patient data do not assure the privacy of patient against genotype-phenotype attack.

In rest of the paper, some related work in section 2. highlights the k-anonymity model for patient data publishing. Section 3. discusses genotype-phenotype attack and weaknesses of k-anonymity against this inference. Experimental results of genotype-phenotype attack on real work data are in section 3. Conclusion and possible future work is summed in section 5.

2 Related Work

Medical data was the first constituent with respect to the privacy of patient data. In the beginning, explicit identifying features of a patient, such as name, address, telephone number and social security number, were removed from the medical data before publishing. This practice was based on the assumption that anonymity is maintained because the resulting data look anonymous. Advocates of such practice claimed that such anonymization sufficiently protects the identity of the patient to whom the medical data corresponds. At a glance, such claims appear to be true. How can one learn the identity of such anonymised data, when there is no register linking anonymised data to identity?

But in her pioneer work Sweeny (Sweeny 2002) showed that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on 5-digit ZIP, gender, date of birth. Further, by linking such anonymised medical data with voter registration list the medical record of governor of Massachusetts was successfully identified.

These analyses arose the need of some proper anonymization mechanism for the publishing of medical data. Sweeny also defined the k-anonymity modal in her pioneer work (Sweeny 2002). It states that a record in the quasi-identifier should be identical to at least (k-1) other records in a data set, and therefore, no individual is identifiable.

3 Genotype-Phenotype Attack on k-anonymity Tables

In this section we formally define genotype-phenotype attack. First we explain some preliminary definitions commonly used in k-anonymization.

3.1 Preliminary Definitions

A data set discussed in k-anonymization has the following attributes.

1. Identifier attribute (A_{ID}) is the attribute that identifies every record in medical data. An example of identifier attribute is social security numbers (SSNs) or names of patient.

2. Quasi-identifier (A_{QI}) attributes is a set of attributes ($A_{QI}{}^1$, $A_{QI}{}^2$,...,$A_{QI}{}^n$) in a table that potentially identifies individuals in the table. Normally, a quasi-identifier attribute set is specified by domain experts. For example zip code, age, sex of patient.

3. Sensitive attribute A_s is a set of attributes that contains private information of patients. It includes disease information etc.

Data sanitation before the publication will remove identifier attribute(s). However, the quasi-identifiers

Table 2: Medical and genomic tables before anonymization

	Identifier attribute	Quasi-identifiers			Sensitive attributes (Phenotype)		
Record #	**Patient Name**	**Age**	**Gender**	**Zip**	**Disease**	**Disease State**	**Treatment**
1	Alex	50	Male	5000	Refsum	4	Medicine – A
2	Adam	55	Male	5095	Huntington	3	Medicine – A
3	Noreen	40	Female	5001	Galactosemia	3	Medicine – C
4	Kathy	45	Female	5007	Fragile X	1	Medicine – F

(a) Patient Medical table

	Identifier Attribute	Quasi Identifiers		Sensitive Attribute
Record #	**Patient Name**	**Gender**	**Age**	**DNA**
1	Alex	Male	50	catg....
2	Richard	Male	47	actg....
3	Noreen	Female	40	ttag....
4	Tanya	Female	39	aacc....

(b) Patient Genomic table

can potentially reveal identity information of individuals. K-anonymization is used to generalize the values in the quasi-identifier attributes and make them indistinguishable. An equivalence class $E_c(A)$ with respect to an attribute set (A) is the set of all records in the table containing identical values for attribute set (A). In anonymization process, attribute set (A) refers to the quasi-identifiers. So proper notation of equivalence class is to be $E_c({A_{QI}}^n)$. A table is k-anonymous with respect to quasi-identifiers if the size of equivalence class is 'k' or more. The k-anonymity property can be defined as $\forall_i |E_i| \geq k$. K-anonymization is a process to modify a table to a view that satisfies the k-anonymity property with respect to the quasi-identifiers $({A_{QI}}^n)$. Where 'k' indicates the strength of anonymization, larger the k, the stronger the privacy protection.

3.2 Genotype-Phenotype Attack

DNA sequences are very important biological information for studying human diseases. DNA sequences also contain rich privacy information of individuals. They can be used to indentify not only a person but also whole family tree. Presently there is tremendous increase in the availability of DNA databases and more than 100 genomic databases are available online over the World Wide Web (Michael & Galperin 2007). Researchers proved that genomic data not only itself can exploit the privacy (Malin & Sweeny 2002) but it can also be linked with some other existing patient information to compromise the privacy of data holder (Malin 2000, 2001).

Clinical phenotypic abnormalities are the *phenotype* of a patient. For example, disease, disease stage, medication etc. Marker genes for some diseases and known hotspots of mutation and other characterized mutations in genes are the *genotype* of a patient. For example, some diseases (like Refum and Galactosemia) can be identified from mutation in specific gene.

Based on biological knowledge, phenotype can be linked to genotype with certain precision. Such precision is improving with improved knowledge in genetic knowledge. Genes and gene mutations are linked to diseases or other medical conditions, so that DNA sequence is linked to medical record where patient is potentially identified from quasi-identifiers.

Genotype-phenotype attack already discussed by Malin (Malin 2005*a*). The DNA sequences are published with little personal identifiable information and are assumed secure for privacy. However, the link of genotype-phenotype between medical and genomic records associates anonymous DNA record to a medical record, where quasi-identifiers contain rich personal identifiable information, and hence no more privacy security for DNA anymore. Genotype-phenotype attack is similar to the quasi-identifier based linking attack used in Sweeneys earlier work with health data re-identification (Sweeney 2000) on non-generalized data. If medical and gene data have not been anonymised, the chance of individuals being identified is great as discussed in (Malin 2000). In more practical scenario, quasi-identifiers of both medical and genomic data sets are k-anonymised before public release. When quasi-identifiers of both medical and genomic data sets are k-anonymised, will the genotype and phenotype attack exists? In the next subsections, we will mainly analyze the cause of the identification and demonstrate the risks in anonymised data.

Formally, we assume that there are two tables D_M and D_G. Where D_M contains patients' medical information including $\{Age,\ Sex,\ Zip,\ Medical\ conditions\ (phenotype)\}$. D_G contains patients' genome information including $\{Age\ Sex,\ DNA\ sequence\ (genotype)\}$. Some records in both tables belong to the same group of patients, i.e. $D_M \cap D_G \neq \emptyset$. The genotype-phenotype attack is simulated as record match, which is defined as follows:

Definition 1: (Match) Match records (T_{Match}) is the pair of record that exists in both medical and genomic tables; when the genotype of a genomic record matches with phenotype of a medical record.

$$D_M(Phenotype) = D_G(Genotype)$$

One of the following situations occurs for each quasi-identifier attribute $({A_{QI}}^n)$ of the match records (T_{Match}) in both tables:

1. Values of quasi-identifier(s) of match records are same in both tables (medical and genomic).
2. Generalization ranges of quasi-identifier of match records of both tables (medical and genomic) overlap each other.
3. One of the generalized value interval of a quasi-identifier of matched records (either genomic or medical) is the sub-set of another one's.
4. Values of quasi-identifier do not match.

Table 3: Medical and genomic table publically released

	Quasi-identifiers			Sensitive attributes (Phenotype)		
Record #	**Age**	**Gender**	**Zip**	**Disease**	**Disease State**	**Treatment**
1	50 – 55	Male	50**	Refsum	4	Medicine – A
2	50 – 55	Male	50**	Huntington	3	Medicine – A
3	40 – 45	Female	500*	Galactosemia	3	Medicine – C
4	40 – 45	Female	500*	Fragile X	1	Medicine – F

(a) 2-anonymised Patient Medical table

	Quasi Identifiers		Sensitive Attribute
Record #	**Gender**	**Age**	**DNA**
1	Male	45 – 50	catg....
2	Male	45 – 50	actg....
3	Female	37 – 42	ttag....
4	Female	37 – 42	aacc....

(b) 2-anonymised Patient Genomic table

Table 4: Publically released Patient Genomic Record after genotype extraction

	Quasi identifier		Sensitive attribute	Extracted Attributes (Genotype)		
Record #	**Gender**	**Age**	**DNA**	**Disease**	**Disease State**	**Medicine**
1	Male	45 – 50	catg....	Refsum	4	Medicine – A
2	Male	45 – 50	actg....	Sickle Cell	1	Medicine – B
3	Female	37 – 42	ttag....	Galactosemia	3	Medicine – C
4	Female	37 – 42	aacc....	Refsum	3	Medicine – F

In the first three cases, privacy of the patient can be compromised. Now we formally define the privacy compromise.

Definition 2: (Privacy Compromise) if any equivalence class $E_c({A_{QI}}^n)$ in k-anonymised table no more satisfies the k-anonymity property then the privacy of the records in that specific equivalence $E_c({A_{QI}}^n)$ class is said to be compromised.

One trivial privacy compromise is that when both data sets have different k, after the matching, the anonymity level is reduced to a smaller k for the matching records. In this paper, we consider a non-trivial case that both data sets have the same k. Table 1. depicts the four possible cases of match records; first three can cause the privacy compromise.

In real world situation, the genotype-phenotype matching is more complicated. A genotype may not match a phenotype with 100% accuracy. There can be multiple matched and unmatched cases. One phenotype in the medical data set may match many genotypes in the genome data set, one genotype in the genomic data set may match many phenotypes in the medical data, or genotype and phenotype has many matches but with different patients. These cases will lead to a low successful rate for genotype-phenotype attack. However, privacy breaching is an individual event and one breaching may cause tremendous damage for a person. Therefore, average possibility is not applicable to privacy protection. Instead, lower bound should be considered in privacy protection. In the following discussions, we will show that there is real possibility that an individual can be identified by genotype-phenotype attack on the k-anonymised medical and genome tables.

Table 2. shows the sample medical and DNA record of different patients. We have a realistic assumption that both medical record and genomic data is held by two different locations. It is already proved (Malin & Sweeney 2004) that different locations can release the medical and genomic data of same patient(s).

Before the public release, both tables are k-anonymised. Table 3. depicts the publically released view of both tables. Both tables satisfy the k-anonymity property where k = 2.

Both data holders anonymize and release medical and genomic data independently. K-anonymity model is simple to understand and easy to implement. Now we apply genotype-phenotype attack on Table 3(a) and Table 3(b).

First, the genotype of patients is extracted from the DNA sequences of Table 3(b). Some traits like single gene diseases, disease state, medicine and dosage of medicine used can be extracted from DNA sequence. These extracted traits depict the 'genotype'. Malin has already demonstrated there are minimum of 40 standardized diseases (via ICD-9 codes) to which DNA mutations in the genome are directly related (Malin 2000). These traits are the result of continuous efforts of bio-research community that are busy to find relationships between genomic mutations and the ability to process drugs and treatments (Altman & Klein 2002, Vaszar et al. 2003). These bio-research efforts also include activities like GWAS[1] and TCGA[2]. Table 4. depicts the extended view of the patient genomic Table 3(b) after 'genotype' extraction.

Figure 1. illustrates the graphical representation of generalization hierarchies of *Age* attribute in both tables (medical and genomic).

With this information in hand, we can link the genotypes of genomic Table 4. with phenotypes of

[1]Genome wide association study (GWAS) is a generic term, which National Institute of Health (NIH) (http://nih.gov) defines as including "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition."

[2]The Cancer Genome Atlas (TCGA). A proposed project to chart the inherited and acquired mutations that relate to the onset, diagnosis, progression, and treatment of cancers, by genotyping bio-specimens and examining the genomic data in light of clinical data on the patients. Piloting is starting. (http://cancergenome.nih.gov)

Table 5: Anonymized tables after genotype-phenotype attack

		Quasi-identifier			Sensitive attributes (Phenotype)		
	Record #	**Age**	**Gender**	**Zip**	**Disease**	**Disease State**	**Treatment**
(T_{Match1})	1	50	Male	50**	Refsum	4	Medicine – A
	2	50 – 55	Male	50**	Huntington	3	Medicine – A
(T_{Match2})	3	40 – 42	Female	500*	Galactosemia	3	Medicine – C
	4	40 – 45	Female	500*	Fragile X	1	Medicine – F

(a) Patient Medical table after genotype-phenotype attack

		Quasi identifier		Sensitive attribute	Extracted Attributes (Genotype)		
	Record #	**Gender**	**Age**	**DNA**	**Disease**	**Disease State**	**Medicine**
(T_{Match1})	1	Male	50	catg....	Refsum	4	Medicine – A
	2	Male	45 – 50	actg....	Sickle Cell	1	Medicine – B
(T_{Match2})	3	Female	40 – 42	ttag....	Galactosemia	3	Medicine – C
	4	Female	37 – 42	aacc....	Refsum	3	Medicine – F

(b) Patient Genomic table after genotype-phenotype attack

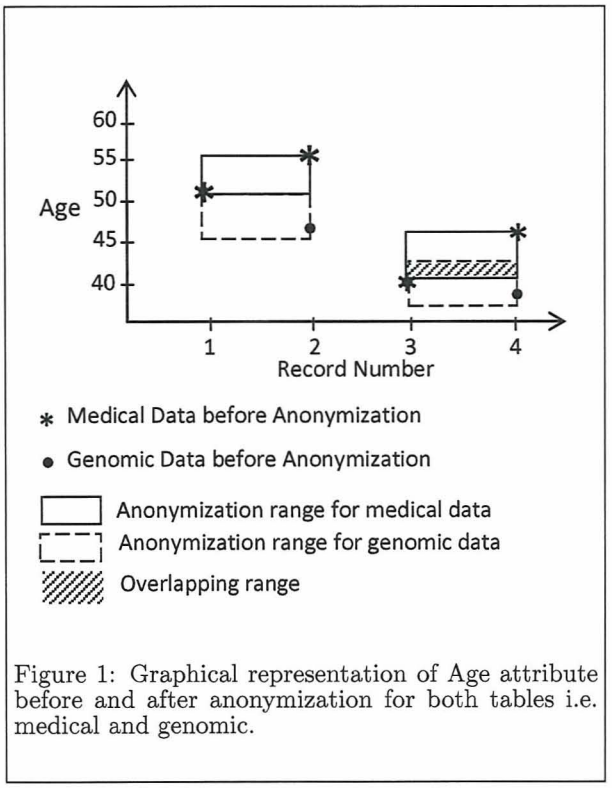


Figure 1: Graphical representation of Age attribute before and after anonymization for both tables i.e. medical and genomic.

medical Table 3(a). Record 1 and 3 of the genomic Table 4. are linked with record 1 and 3 of the k-anonymised medical Table 3(a); because the *genotype* (extracted attributes) of these genomic records match with the *phenotype* (sensitive attributes) of the corresponding k-anonymised medical records. As the anonymization range of *age* attribute in both tables is overlapping, so for linked records (1 and 3) the anonymization range of *age* attribute is also reduced to overlap part. Both anonymised tables (medical and genomic) are shown in Table 5(a),(b) respectively after genotype-phenotype attack.

It is possible that a genotype of genomic table may match with more than one phenotype in medical table or vice versa. Even in this case the privacy of patients is at risk because the k-anonymity property of both tables (medical and genomic) is compromised. Both medical and genomic Table 5(a),(b) are no more 2-anonymous. After genotype-phenotype attack, size of equivalence classes for both tables (medical and genomic) is reduced to 1, which is less than the original value of k i.e. 2. As per the definition of k-anonymity property the size of equivalence class should be more than or equal to k. Both tables (medical and genomic) after genotype-phenotype are now also open to other re-identification risks (Sweeny 2002). Its clear that k-anonymity is prone to genotype-phenotype attack. Sharing the patient data is helpful in many ways but it is seriously considered that such sharing is only prospective if the privacy concerns of the data owners (like genotype-phenotype attack) are properly addressed.

4 Genotype-Phenotype Attack Experiments

In this section, we explain our genotype-phenotype attack experiments to illustrate the risk of this attack. First we define some of the definitions used in these experiments.

Definition 3: (True Match) A pair of match records (T_{Match}), is said to be True Match ($T_{TrueMatch}$) when both records belongs to same person.

Definition 4: (False Match) A pair of match records (T_{Match}), is said to be False Match ($T_{FalseMatch}$) when it do not belongs to same person.

True match is actual risk to privacy of patients. Situation of False Match is possible because one genotype/phenotype may match one-to-many records. For example, there can be two or more different patients having same age, gender and disease.

Definition 5: (True Match Precision) Ratio between true match records ($T_{TrueMatch}$) and Match (T_{Match}) records is the *True Match Precision* $P_{True}(GP)$ of genotype-phenotype attack. The True Match Precision $P_{True}(GP)$ can be defined as:

$$P_{True}(GP) = \frac{\sum T_{(TrueMatch)}}{\sum T_{(Match)}}$$

The maximum value of true match precision is 1, when all match records are the true match. Similarly when none of the match record is true match record the true match precision is 0. More the value of is closer to 1, higher the risk of privacy compromise.

In our experiments, we sample data from a survey data set (Asuncion & Newman 2007) to simulate medical and genome data sets. The survey data set has total 8 attributes. We use *Age* and *sex* as quasi-identifier for both data sets, and rest information is assumed as phenotype and genotype in simulated medical and genome data sets. Table 6. depicts the division of test data. There is a match if values apart from *Age* and *sex* in two records from medical

Table 6: Attributes of Test Data set

Attribute	A_{QI}/A_s
Age	A_{QI}
Work Class	A_s
Years of Education	A_s
Marital Status	A_s
Occupation	A_s
Race	A_s
Sex	A_{QI}
Native Country	A_s
Salary	A_s

Table 7: Division of Test Data set into two sub-data sets

	Simulated Medical	Simulated Genome
Initial Records	10000	10000
Overlap Records	6912	6912
Data set Size	16912	16912

and genome data sets are identical. We control the overlapping records in the both data sets.

Both quasi-identifiers (*age,sex*) present in medical and genomic data. Rest of (sensitive) attributes depict phenotype in Simulated Medical data set and genotype in Simulated Genomic data set. We used Mondrian (LeFevre et al. 2006) k-anonymization in these experiments. In all iterations of experiments both quasi-identifiers (age, sex) use the same generalization hierarchy. The initial characteristics of both data sets (simulated medical and genomic) are described in Table 7.

Adult dataset has 33824 distinct records. We divided distinct records into two datasets each of 16912 records, out of that 6912 records were marked as overlap records. We repeat five set of experiments, varying the ratio of true overlap records (from 0% to 100%) within overlap records. Table 8. depicts the ratio between true and false overlap records before each set of experiment.

In each set of experiment both data sets k-anonymised to different values of k then we calculated true and false match records for each k-anonymization. Table 9. depicts the results of our all set of experiments.

It is clear that in each set of experiments on average 90% true overlap records are correctly identified. All correctly identified true match records depict the success of genotype-phenotype attack and can breach the patients privacy.

Individual privacy breach is not the only risk of genotype-phenotype attack. As per the definition of privacy compromise, the privacy of the patients is on risk if anonymised table no more satisfies k-anonymity property. In our experiments, one or both of the simulated data sets (Medical and Genomic) do not hold the k-anonymity property after genotype-phenotype attack. Figure 2. depicts the degree of k-anonymity before and after genotype-phenotype attack in simulated medical data set. To illustrate the actual privacy risk, fig 2. includes only those equivalence classes that contain the true match records.

Table 8: Division of overlap records between True and False overlap records for different experiments

Experiment #	Total True Match Records	Total False Match Records	%True Match Records
1	0	6912	0%
2	1728	5184	25%
3	3456	3456	50%
4	5184	1728	75%
5	6912	0	100%

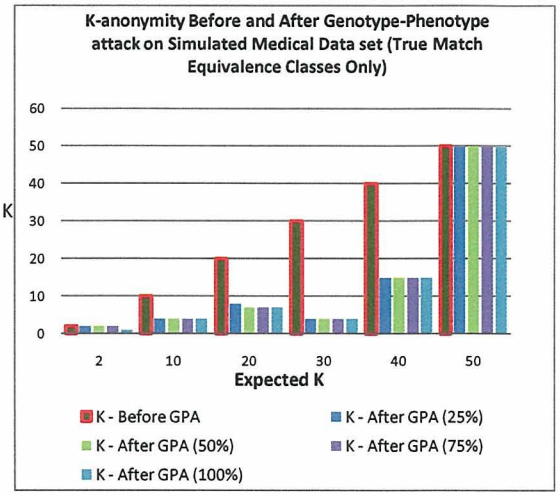


Figure 2: K-anonymity before and after genotype-phenotype attack on Simulated Medical data set (True Match Records Equivalence Classes only)

It is clear, that after genotype-phenotype attack either one or both data sets (Simulated Medical/Genomic) do not hold the k-anonymity property. In most cases when k-anonymity property is compromised, the degree of k-anonymity decreased to 50%. Interestingly, when k=50 the degree of k-anonymity remains unchanged after genotype-phenotype attack. For higher value of k, anonymised data set loses significant information and is of less beneficial. Any re-identification attack is only a risk when anonymised data is still useful.

5 Conclusion and Future Work

We have shown that genotype-phenotype attack is a risk to privacy preserving patient data publishing. We show through experiments that k-anonymity property can be compromised due to this attack. Data set compromised is subject to other re-identification. With the increased medical information available in medical records and better understandings of genetic links of diseases with gene defects, the threat of genotype-phenotype attack is increasing. Even though both medical and genomic data sets have been properly anonymised individually by different organizations, the change of genotype-phenotype attack still exists.

It is an important future work for privacy preserving data publishing to guard against genotype-phenotype attack.

References

Altman, R. & Klein, T. (2002), 'Challenges for biomedical informatics and pharmacogenomics', *Annual review of pharmacology and toxicology* **42**, 113–133.

Table 9: Test results of Simulated Medical and Simulated Genomic data sets

Total *True Match* Records	Total *False Match* Records	K	'K' after GP Attack (Simulated Medical)	'K' after GP Attack (Simulated Genomic)	Identified *True Match* Records	Identified *False Match* Records	True Match Precision
0	6912	2	1	1	0	5877	0
		10	4	1	0	5844	0
		20	7	20	0	5783	0
		30	4	30	0	5690	0
		40	12	40	0	5647	0
		50	50	50	0	5288	0
1728	5184	2	1	1	1502	4375	0.255
		10	4	1	1492	4352	0.255
		20	7	20	1474	4309	0.254
		30	4	30	1450	4240	0.253
		40	12	40	1435	4212	0.252
		50	50	50	1337	3951	0.253
3456	3456	2	1	1	2988	2889	0.508
		10	4	1	2968	2876	0.507
		20	7	20	2936	2847	0.506
		30	4	30	2891	2799	0.505
		40	12	40	2867	2780	0.504
		50	50	50	2665	2623	0.504
5184	1728	2	1	1	4457	1420	0.758
		10	4	1	4427	1417	0.756
		20	7	20	4377	1406	0.755
		30	4	30	4308	1382	0.752
		40	12	40	4275	1372	0.751
		50	50	50	3984	1304	0.753
6912	0	2	1	1	5877	0	0.999
		10	4	1	5844	0	0.998
		20	7	20	5783	0	0.997
		30	4	30	5690	0	0.993
		40	12	40	5647	0	0.992
		50	50	50	5288	0	1

Altschul, S. (1990), 'Basic local alignment search tool', *Journal of Molecular Biology* **215(3)**, 403–410.

Asuncion & Newman (2007), 'Uci machine learning repository'.
URL: *http://archive.ics.uci.edu/ml/datasets.html*

Burnett, L., Barlow-Stewart, K., Pros, A. & H.Aizenberg (2003), 'The gene trustee: a universal identification system that ensures privacy and confidentiality for human genetic databases', *Journal of Law and Medicine* **10(4)**, 506–513.

Caenazzo, L. (1998), 'Prenatal sexing and sex determination in infants with ambiguous genitalia by olymerase chain reaction', *Genetic Test* **1(4)**, 289–291.

Clayton, E. (2003), 'Ethical, legal, and social implications of genomic medicine', *New England Journal of Medicine* **349(6)**, 562–569.

DHHS (2002), 'Standards for privacy of individually identifiable health information, final rule', *Department of Health and Human Services, Federal Register* **67(157)**, 53182–53273.

Gaudet, Arsnault & Belanger (1999), 'Procedure to protect confidentiality of familial data in community genetics and genomics research', *Clinical Genetics* **55(4)**, 259–264.

GINA (2008), 'Genetic information nondiscrimination act (gina)'.
URL: *http://www.genome.gov/24519851*

Hall, M. & Rich, S. (2000), 'Patients' fear of genetic discrimination by health insurers: the impact of legal protections', *Genet Med* **2**, 214–221.

LeFevre, K., DeWitt, D. J. & Ramakrishnan, R. (2006), 'Mondrian multidimensional k-anonymity', *ICDE'06* p. 25.

Machanavajjhala, A., Kifer, D. & Gehrke, J. (2007), 'L-diversity: Privacy beyond k-anonymity', *ACM Transactions on Knowledge Discovery from Data* **1(1)**.

Malin, B. A. (2000), 'Determining the identifiability of dna database entries', *Journal of the American Medical Informatics Association* pp. 537–541.

Malin, B. A. (2001), 'Re-identification of dna through an automated linkage process', *Journal of the American Medical Informatics Association* pp. 423–427.

Malin, B. A. (2005*a*), 'An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future', *Journal of the American Medical Informatics Association* **12**, 28–34.

Malin, B. A. (2005*b*), 'Protecting genomic sequence anonymity with generalization lattices', *Methods of Information in Medicine* **44(5)**, 687–692.

Malin, B. A. & Sweeney, L. (2004), 'How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems', *Journal of the American Medical Informatics Association* **37**, 179–192.

Malin, B. A. & Sweeny, L. (2002), 'Inferring genotype from clinical phenotype through a knowledge based algorithm', *Proceedings of the Pacific Symposium of Biocomputing* pp. 41–52.

Michael, Y. & Galperin (2007), 'The molecular biology database collection', *Nucleic Acids Research* **35**(D3-D4).

M.West, Ginsburg, G., Huang, A. & Nevins., J. (2006), 'Embracing the complexity of genomic data for personalized medicine', *Genome Research* **16**, 559–566.

NHS (2000), 'Data protection act 1998 - protection and use of patient information', *NHS Executive, (HSC 2000/009)* .

Roses, A. (2000), 'Pharmacogenetics and pharmacogenomics in the discovery and development of medicines', *Nature* **38**, 815–818.

Rothstein, M. (1997), *Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era*, New Haven: Yale University Press.

Sweeney, L. (2000), Uniqueness of simple demographics in the u.s. population, Technical Report LIDAP–WP4, Data Privacy Laboratory, Carnegie Mellon University, Pittsburgh, PA.

Sweeny, L. (2002), 'K-anonymity: a model for protecting privacy', *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems* **10(5)**, 571–588.

Vaszar, L., Cho, M. & Raffin, T. (2003), 'Privacy issues in personalized medicine', *Pharmacogenomics* **4(2)**, 107–112.

Wong, R. C.-W., Li, J., Fu, A. W.-C. & Wang, K. (2006), '(alpha, k) anonymity: An enhanced kanonymity model for privacypreserving data publishing', *KDD06* pp. 754–759.

Xiao, X. & Tao, Y. (2007), 'm-invariance: Towards privacy preserving re-publication of dynamic data sets', *SIGMOD* pp. 698–700.