Compromising Privacy in Distributed Population-Based Databases with Trail Matching: A DNA Example

Bradley Malin Latanya Sweeney

December 2002

CMU-CS-02-189

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

This paper is concerned with the privacy of person-specific data collected over multiple institutions. In particular, we focus on an example of person-specific DNA sequences collected and stored at various hospitals in a defined geographic region. The applications of human genetics and genomic analysis have generated much discussion with respect to privacy and confidentiality in ethical, legal, and social issues. For the most part, the previous analysis has concentrated on direct application and disclosure of the genetic information of an individual, however, there has been much less attention devoted to the question of computational challenges to privacy in the secondary sharing of de-identified databases (i.e. released in a format devoid of directly identifying information, such as name, address, or phone number). We introduce methods for determining the re-identifiability of such DNA data and, in the process of doing so, prove that the removal of identifying information from DNA does not sufficiently protect the privacy of the entities to which the data was derived from. We demonstrate, through several novel re-identification algorithms, that despite a lack of personal demographic information, such database entries can be re-identified through linkage to other publicly available databases, such as hospital discharge information through the use of hospital visit and data collection patterns, which we refer to as data trails, which are iteratively discovered from released data collections. Using real-world data, we are able to determine when identifiable linkages can occur for a substantial number of individuals with particular gene-based disorders. Furthermore, we provide empirical analysis of the re-identification algorithms with respect to population-institution visit distributions and data trails.

This research was supported by the Laboratory for International Data Privacy at Carnegie Mellon University.

Keywords: *data privacy, anonymity, security, re-identification algorithms, databases*

1 Introduction

The dramatic increase in the quantity of knowledge corresponding to the relationships between genome sequence and an entity's phenotype or potential phenotype has useful applications in genetic and molecular biology basic research, clinical medical research, biopharmaceutical research and development [1], public health analysis [2], and occupational safety [3]. Current statistics compiled by the National Center for Biotechnology Information (NCBI) demonstrate, that as of the end of 2002, almost 14000[1] human genetic loci have been documented and established, over a thousand of which have been characterized as influencing genetic disease [4]. The discovery and physical mapping of human genetic components have greatly benefited by recent developments in bioinformatics, automated sequencing, and digital storage technologies, thus allowing for an exponential increase in the discovery and analysis of genetic loci [5]. Yet despite the considerable research benefit that will be produced from collections of such data, society must consider the consequences to privacy that can emerge when large quantities of person-specific data are stockpiled.

The genetic information of an individual is understood to be as, or more, personally revealing than a fingerprint [6]. In recognition of this fact, the privacy of an individual's genetic information has been discussed at length in several communities, including those pertaining to law, public policy, molecular medicine, the biopharmaceutical industry, and public health [7]. Discussions within and between such communities have, for the most part, focused on issues of 1) ownership of information, 2) the ethical duty of physicians and counselors to protect their patients' rights, and 3) genetic discrimination. The previous arguments address the direct release and application of genetic information, however, genetic data is useful in realms beyond the locale of an initial collection. Many groups harboring collections of genetic information share, or hope to do so in the future, entity-specific data for various endeavors, such as licenses to private or academic research groups, public use datasets, and public health research. The issue of secondary use and sharing of collected genetic data has been discussed by several of the information collecting communities, but the question "How do you ensure the privacy of an individual in a released dataset?" has led to insufficient response and solutions. It is recognized that anonymity is necessary for patients, but initial attempts to protect privacy were based on the removal of direct identifying attributes, such as name, address, and phone number. This technique is known as de-identification and subsequent released datasets are referred to as de-identified. With respect to DNA sequence data, a de-identified database of mere DNA entries, with no additional explicit demographic information or identifiers included, can appear sufficiently anonymous, since associating collected DNA information to named persons seems impossible. One might make argue that there exists no master registry against which the DNA data could be directly compared to reveal the associated person's identity and, therefore, the identity of the DNA is protected. So how then, can individuals be re-identified, or in other words, how can an adversary learn the explicit identity (i.e. name, address, phone number) of an individual?

This work demonstrates that inferences drawn from de-identified DNA information, and other publicly available sources of medical information, can be used to divulge the exact identities of the persons from whom the DNA originated. Our methods are based on determining unique features from the set of locations (e.g. hospitals) that both data for an identified entity and the corresponding de-identified data are stored at. The constructed unique features allow for seemingly anonymous data to be re-identified, or have a known identity label a previously de-identified piece of data. The basic outline of the attack proceeds as follows. We consider a set of institutions, such as hospitals, that collect data on a patient population. Each institution collects two types of data. The first is identifiable heath information, such as demographics of age, gender, local zip code and a representation of the diagnoses made. The second consists of DNA sequence data on particular individuals. Each hospital keeps a record of which DNA relates to which identified patient, yet when the data is released all identifiers are removed from the DNA. The attack method is based on the fact that an adversary can collect releases from multiple institutions and reconstruct the locations that a particular entity visited, and likewise for the DNA data.

[1] See http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html for current statistics

We refer to such reconstructions as trails, and it is through unique features in such trails that a linkage is established between identities and their DNA data.

The work is divided into five sections, including the current introduction. In the second section, relevant background of DNA information systems, genotype-phenotype relationships, and previous attempts to protect data are presented. In this section, we also map the growth of DNA information and its change from single representative sequences to population-based, or many distinct individuals, collections. The third section provides the details to our methods for re-identification. Results of re-identifiability are demonstrated in the fourth section with publicly available hospital discharge information. In the fifth section, we attempt to provide an intuitive statistical model for why re-identification occurs. In the final section, we discuss the implications of this research and explore limitations of and extensions to the methodology.

2 Background

We begin with an introduction to the current relationships between human genetics and digitally stored information. This issue is addressed to demonstrate the increasing simplicity of access and scope of information that are available for a novice to learn the relationships between genes and disease. Thus, any detailed descriptions of the history of the rise of computers and molecular biology are withheld, so that we may concentrate on several key features of the current status of information storage and dissemination. For a historical account of the relationships between online information and molecular biology, we refer the reader to [8].

To cope with the ever-increasing quantities of genetic research data and facilitate scientific discovery, the NCBI established the Online Mendelian Inheritance in Man (OMIM) database in 1987 [9]. This database has been subdivided into several search mechanisms, one for specific searches corresponding to the keywords about genetic traits and the other corresponding to the location of known disease genes. The OMIM database is a catalogue of human genes and disorders containing textual information on data pertinent to a certain gene, as well as cytogenetic maps and reference information. Moreover, the OMIM database lists the current resolution level of the chromosomal mapping of each gene or genetic locus entry, as well as important allelic variants that are known causes of clinical phenotypic abnormalities.

Other major central databases are specific to published mutations by gene and mutation type, such as the Human Gene Mutation Database [10] or by annotated sequence, such as the Database of Single Nucleotide Polymorphisms [11]. In addition to such databases, online websites, such as GeneClinics at the University of Washington, have been implemented to facilitate the flow of information between the public and the medical genetics community. GeneClinics goals are to provide "disease-specific information on molecular genetic testing and its role in diagnosis, genetic counseling, and when appropriate, surveillance of at-risk relatives" [12]. Access to such information databases is critical to the progress of human mutation research [13].

2.1. First Generation DNA Databases

A recent survey of the World Wide Web [14] finds that there are currently more than 500 molecular biology and genomic databases online. While OMIM and other NCBI-curated databases are organized as collective references, they do not exhaust all non-disease causing allelic variants of every known gene. Yet, many additional genomic information databases do exist that compile all known allelic variants for particular genes and are publicly available. Many of the databases are online and provide gene specific information corresponding to known mutations or polymorphisms. Some WWW based sites construct a physical map displaying the mutations. While most online genetic databases are publicly available, the data that such systems include generally fall into one of two groups. The first type of site lists standard genomic information consisting of the raw consensus sequences (the most common non-mutated DNA sequence) and polymorphic sequences. The other type of site is more specific and detailed for a particular genetic locus. Such sites catalogue the known characterized mutations in the genetic locus. They also record specific mutation's and polymorphism's respective relationships to molecular and clinical phenotypes. Information in these databases appears to be sufficiently anonymous, for they aim to provide the first examples of particular mutations and their dissemination for continuing research. The individual supplying the mutation may be geographically located anywhere in the world. We term such databases as first generation publicly available genetic databases.

First generation databases, including locus specific and central mutation databases, convey the impression of being harmless collections with a research orientation. In fact, such are the intentions of the providers, however, the same sites may directly compromise the privacy of the data contributing subjects. One instance of this type of database is exemplified in the Cystic Fibrosis Mutation Database maintained by the Cystic Fibrosis Genetic Analysis Consortium [15]. The database provides a mutation table of all published mutations in the CFTR gene, polymorphisms in the coding and non-coding regions of the gene, and references from which the mutations were submitted. The listed mutations have origins located in the United States, France, Italy, Soviet Union, and other places around the world, and totaling over 900 mutations. The value of such databases has been paramount in helping researchers determine hotspots for mutation and understand clinical phenotypes associated with specific mutations. Nonetheless, it should be noted that first-generation genetic collection databases are not necessarily anonymous. While they do not harbor explicit identifying data about individuals, such databases may be discredited in the realm of anonymity. If it was the objective of an unwanted intruder to compromise anonymity in the collection, one could possibly perform at least one re-identification based on the publication and references associated with each reported piece of genetic information. The research we present below does not elaborate on this hypothesis; rather, the previous is introduced to demonstrate the general false belief of privacy that de-identification, or systematic removal of explicit identifying features of information offers for DNA data.

2.2. Second Generation DNA Databases

The ability to conduct diagnostic DNA sequence analysis has fuelled the collection of population-based DNA, which we term second generation DNA databases. In such cases, submissions of well-characterized genomic data corresponding to multiple individuals appear in the database. The submissions, which are partitioned into entries analogous to each individual, include sequences from individuals within a specific population, such as a particular hospital, biopharmaceutical company, or government clinical trial. As DNA diagnostic sequencing speed continues to increase and the economic cost of such tests decrease, so too will the quantity of entries and completeness of records in genomic databases grow. We have already seen the rise of the aforementioned databases in certain university hospitals and private companies, such as Incyte Genomics, deCode Genetics [16], and the PharmGKB project [17].

Collecting and sharing such information is useful to researchers and clinicians, yet, due to geographical specification and additional inferences that can be extracted from the sequences, which we present a general algorithm for below, the privacy concerns for second generation databases are far more grievous than those touched upon above for the first generation. The techniques in which an attack could be launched on second generation DNA databases are more extensive and would have more effect on the personal life of a larger number of patients. There are a multitude of reasons why one may desire to keep their genetic information private, such as to prevent discrimination in employment, or insurance, or avoid social stigma [18]. In addition, the heredity of genetic features shifts the scope of this problem from the individual to the family. Thus, it is the records of second generation DNA databases that our research is concerned with.

2.3. Computational Methods of Privacy Protection

There have been several computational systems introduced that help render data anonymous. These include Scrub [19], which locates personally identifying information in unrestricted textual documents, and the Datafly [20] and Mu-Argus [21] systems, which attempt to render field-structured person-specific databases sufficiently anonymous. A recent attempt to protect DNA sequences has been introduced in the work of Lin et al [22]. In this work, the technique of "binning" is applied to protect single nucleotide polymorphism (SNP) data. This technique attempts to release data such that for each set of characteristics in the data there are at least as many released sequences equal to the size of a specific bin. However, the technique addresses variant regions in DNA with a length of one nucleotide, and as such, it does not address the issue of many other common types of mutations, including insertions, deletions, inversions, or repeat mutations. Furthermore, the method protects regions that have variation of a maximum of two different nucleotides, but mutations may occur through any of the four possible nucleotides. Currently, there are no known generalized techniques to protect DNA sequence beyond that of single nucleotide variations or for nucleotide regions with a distribution of nucleotides greater than two. As such, none of the proposed systems can stymie the attack strategy described below.

3 Re-identification Methods

We propose a system for determining the re-identifiability of sequenced DNA, independent of any explicit demographics or identifiers maintained with the information. The method reveals DNA database entries that can be re-identified to the subjects of the data. Figure 1 provides a graphical overview of the re-identification procedure, and the steps of the procedure are elaborated upon below.

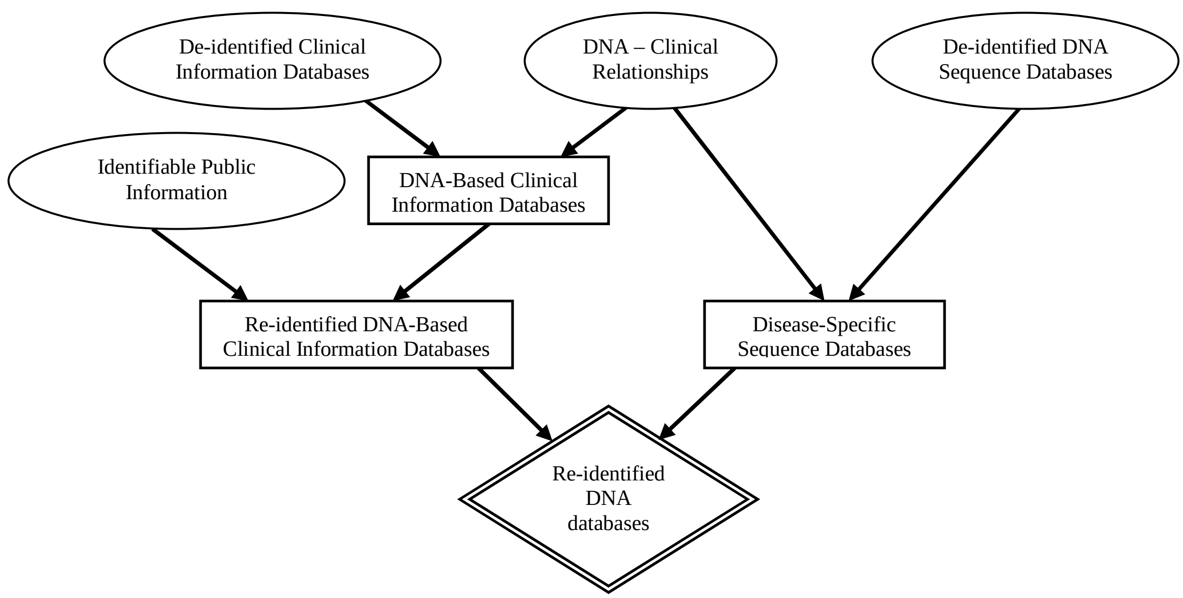

Figure 1. Overview of the DNA re-identification procedure.

First, clinical information databases are partitioned into patient profiles with DNA-based diseases and those without such diseases. Next, this information is re-identified with publicly available information. Third, DNA databases from various institutions are partitioned into entries with a known disease gene mutation and entries without known mutation. Finally, the re-identified clinical information partition and known DNA mutation partition are crossed for linkage. Re-identification occurs when a positive link is established. The following in sections 3.1 through 3.4 are expanded upon from their initial publication in [23].

3.1. Step 1: Relating Clinical Codes and Genetic Disease

Many diseases are known to have genetic influence, and of such diseases, a growing number are being found to depend on interactions of multiple gene products and environmental influence, such as certain cancers. Mutations in specific genes may not necessarily cause a certain disease, but instead can raise the risk of developing that disease, such as mutations in the BRCT domain of BRCA1 and breast cancer [24], or variants of the APOE4 gene and late onset Alzheimer's disease [25]. Still, there exists an expanding group of genetic influence diseases that are caused by mutation in a single gene. These diseases span a variety of biological processes involving cancer, immunity, metabolism, the nervous system, signaling, and molecular transporters [26]. As a result of the growing number of characterized disease genes, the first step is to determine and organize the relationships between single gene detectable disease and publicly available hospital discharge data. Information and references on the relationships between genes and diseases can be found at NCBI's website. OMIM and Genes and Disease (GD) [27] pages were used to determine clinically tractable diseases with a DNA basis. Most of the genetic disorders discussed on these websites are the direct result of mutations in a single gene. To determine diseases with a deterministic genetic basis that would be useful for this study, we searched for diseases on the GD pages in the diagnoses codes of publicly available discharge databases. The codes used in this study were International Classification of Diseases, Ninth revision (ICD-9) [28].

3.2. Step 2: Re-Identifiability of Individuals with Genetic Diseases

Population-based health record profiles were constructed from state collected hospital discharge databases, which we refer to as health data profiles, via the method depicted in Figure 2. For each hospital visit in the health data that contains a diagnosis corresponding to that of a single disease gene, a profile is constructed consisting of the attributes {*date of birth*, *gender*, *ZIP*, *disease*, *hospital visit info*}, where *ZIP* is the patient's residential postal code. Profiles are then merged based on census demographics for {*age*, *gender*, *ZIP*} so that values for *hospital visit info* from profiles that are likely to relate to the same person were combined. The set of resulting profiles contain the demographics for persons diagnosed with targeted diseases and information from the hospital collecting data for a second-generation DNA database. It has been shown using data linkage algorithms that 80-100% of discharge database entries can be accurately re-identified using publicly available population registers [29]. Therefore, if we can match re-identified patients from a discharge database, who have been diagnosed with a simple genetic disease, to corresponding genetic sequences stored in a second-generation DNA database (maintained by a hospital), we can reveal the identity of the DNA contributors in almost all cases.

The combination of attributes from a database that can uniquely indicate an entity are termed a quasi-identifier (QI) [30]. The identifiability of the quasi-identifier was evaluated by determining the uniqueness of the quasi-identifier value combinations in the population. For example, consider the quasi-identifying value {4/6/75, M, 61010}. If there exists only one living person with this combination during the period the data was collected, then this individual could be positively re-identified. However, if there existed more than 1, then attempts to determine the identification of the individual, using solely the quasi-identifier, would yield ambiguous results.

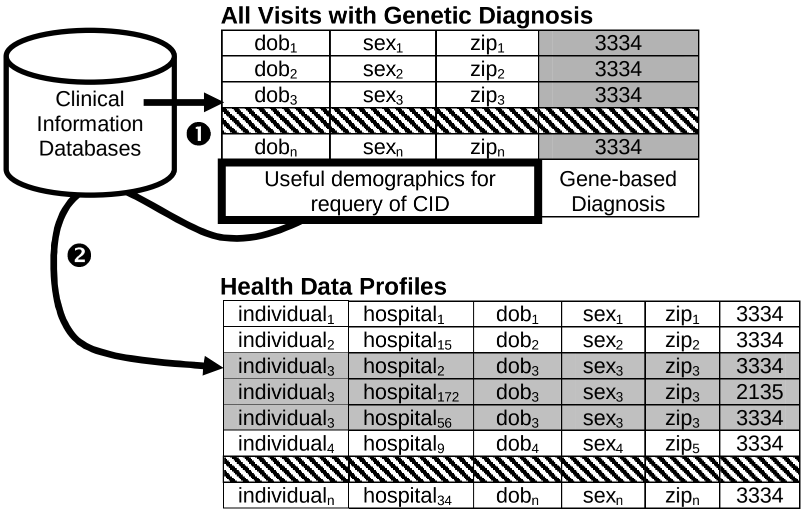

Figure 2. Longitudinal health record profile construction, for Huntington's disease (i.e. ICD-9 3334) from clinical information databases. The first step queries for all distinct demographics with the specified diagnosis. The shaded column corresponds to the specific diagnosis code used for the query. Next, the databases are requeried with the captured demographic combinations to append additional clinical information from other hospital visits. Individual_i is an appended identifier for each unique entity, determined from demographic information. The shaded set corresponds to a profile for a single individual.

3.4. Step 3: Direct Information Inference from DNA Sequence

In previous work, we introduced the CleanGene [23] methodology, which offered several methods for inference from sequence. This tool can be useful in the inference of 1) deterministic disease gene mutations, 2) the hospital releasing a sequence, as well as in particular cases, 3) gender of the subject may be learned.

3.5. Step 4: Linking DNA and Health Database Entries

Here we introduce several methods for linkage of health and DNA data. Both of the presented methods employ techniques to link data with minimal information relative to the genotype-phenotype relationship. There are several assumptions invoked to assist in this analysis. First, the only genetic knowledge considered is as follows: no prior knowledge is used to discern between different types of mutations for a specific disease. For example, it is well known that Huntington's disease is characterized by a strong inverse relationship between the size of the CAG triplet repeat expansion and the age of onset of the disease [31, 32]. Thus, there exist additional inferences that can be extracted beyond correspondence of gene and ICD-9 code. Yet, in this study we are concerned with the general situation only; an individual has an abnormal number of CAG repeats in the HD gene. In doing so, we prevent any inaccurate or ambiguous linkages that might occur due to the overlapping of age of onset variance associated with each repeat size. Furthermore, if a mutation is found in a disease gene sequence released by a data collecting institution, then there exists the presence of an observable clinical phenotype as specified by the associated ICD-9 code.

The final assumption states that the DNA of an individual has minimal changes the times from collection from one institution to another. For example, consider the following hypothetical scenario. In 1996, Mr. Jones went to the University of Chicago Medical Center and had his DNA sequenced for some reason, such as diagnostic testing of a particular disease. Two years later, Mr. Jones has treatment for a disorder at another hospital in Illinois, such as Rush Presbyterian Hospital (in Chicago). Once again, Mr.

Jones has his DNA sequenced or his DNA is sent from the first collecting hospital. At both hospitals, the Mr. Jones' DNA sequence is stored in a DNA database. There may be some variation between the two sets of sequences, due to some random occurrence, such as mutation during cell division over time, sequence analysis glitch, or difference in tissue type that the DNA was procured from. However, the difference between Mr. Jones two samples of DNA would still be more similar than his DNA and the sequences of that of some random individual, Mr. Smith.

3.5.1. Intersect-Purge

The Intersect-Purge (IP) system utilizes longitudinal datasets, which are constructed via the uniqueness of combinations of demographics of individuals in discharge databases as described above. The formal algorithm is presented in Figure 3.

Algorithm: Intersect-Purge	
Input	1) Patient profiles of clinical data (basic hospital visit information) 2) DNA database information from collecting institutions
Output	List of re-identified DNA database entries
Assumes	DNA entries of different individuals can be resolved
Step 1	Construct binary matrix of patient health information and collecting institutions
Step 2	Construct binary matrix of DNA information and collecting institutions
while	Unique pair (DNA, hospital) exists
Step 3	Attempt re-identification of pair
Step 4	Remove outlier column (patient) and row (hospital)

Figure 3. Pseudocode for the Intersect-Purge algorithm

The following assumptions are made for convenience.

1) Let *HOSPITALS* be the set of hospital identification numbers (HID) for which DNA and hospital discharge data are available specific to the disease gene. In Figure 4, *HOSPITALS* = $\{H_1, H_2, H_3\}$.
2) Let *DNA* be the union of all DNA available from hospitals specific to the disease gene. In fact, *DNA* is a table over (*HID*, *Sequence*).
3) Let *DISCHARGE* be the union of all hospital discharge available for visits from the hospitals that include a diagnosis specific to the disease gene.

Step 1: Establish Uniqueness. The goal of IP is to determine how data from a re-identified database, the hospital discharge profiles, can be linked to de-identified database, the DNA entries. To establish this link, we have generalized the genotype-phenotype relationship and considered only the use of audit trails in the re-identification process. Therefore, the only factor indicative of uniqueness is the number of individuals visiting a particular hospital. When there is only one individual at a hospital, then a re-idenitification must occur between this individual and the lone DNA sequence at this hospital. However, if there are additional features that permit a relationship to be established between DNA and health data the number of possible re-identifications can be increased. If gender of the individual was explicitly stated or could have been inferred from *DNA* [33], then the sets *HOSPITALS* and *DNA* could have been partitioned into mutually exclusive smaller sets based on the gender of the individual and the gender of the DNA. Therefore, for any attribute with greater than one value, we can increase the maximum number of possible unique entries. In actuality, the number of combinations is equal to $\prod_i |a_i|$, where $|a|$ is the number of distinct values for an attribute that is common between the two datasets being linked. For example, consider the use of the following fields {*hospital id*, *sex*, *gp*[2]}, the number of distinct classes for re-identification is |*hospital id*|*|*sex*|*|*gp*|. When the DNA and health datasets are crosses several key facts are

[2] The set of genotype-phenotype relationship for a particular genetic loci.

employed; namely that 1) it is known that the hospitals visited in the two datasets are the same, 2) there is only one type of mutation considered that can cause the observed clinical diagnosis, and 3) we can distinguish between the genders in the database entries. Thus, if it is impossible to distinguish between the genders, then the number of classes that the datasets would reveal would be the number of hospitals.

Step 2: Remove Duplicate Information. Once a DNA entry is re-identified, the entry and its associated identied information no longer need to be considered for re-identification. Removal of this data is possible, since each DNA entry is traceable from one institution to another. The removal of duplicate data, can allow for new DNA sequences and identified data to become unique outliers.

Step 3: Iterate Until Fail Re-identification. After removal of the re-identified individuals from all other hospitals that they visited, the search process restarts. The program continues in an iterative manner until the search step fails to re-identify an individual.

3.5.1.1. Intersect-Purge Example

A graphical overview of a simple situation is provided in Figure 4 in which three individuals deposit clinical and genetic information over 3 hospitals. Patient P_1 visits hospital H_1. Patient P_2 visits hospitals H_1 and H_2. And finally, patient P_3 visits hospitals H_1, H_2 and H_3. Following a patient's visit, a hospital is reported to maintain a copy of the DNA sequence information for the patient, denoted $ACTG_i$ and hospital discharge information for the patient's visit that includes {*Birthdate*, *Gender*, *ZIP*}. The hospital discharge information is denoted DGZ_i.

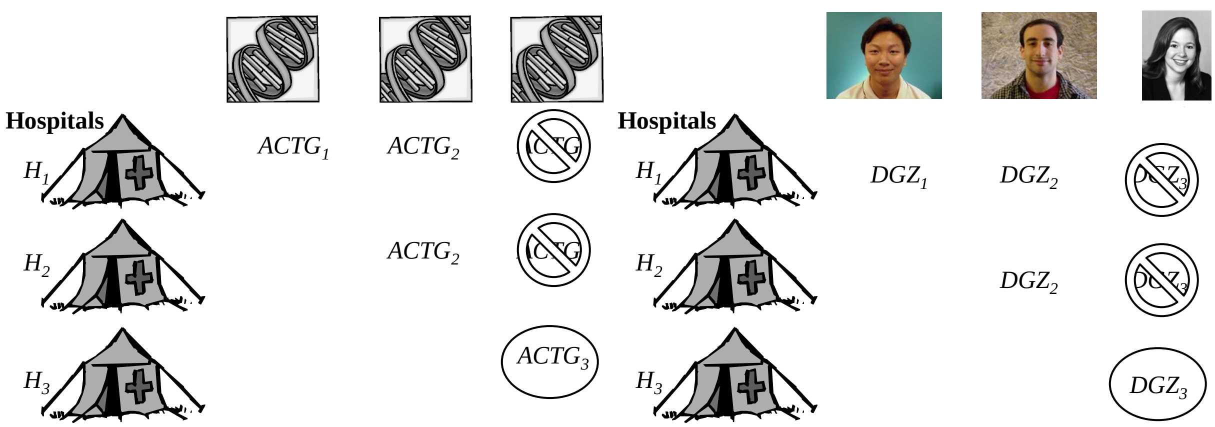

Figure 4. Graphical representation of patient information in hospital databases and Intersect-Purge after first round of outlier detection. The open circle represents a linkage of DNA to re-identifiable health information, the prohibit symbol represents removal of the information from the consideration range in the next pass of the algorithm.

If H_3 has only one patient in *DNA* presenting the disease gene and the only patient reported in *DISCHARGE* with a diagnosis of the disease, then the DNA sequence must originate from DGZ_3, which identifies P_3. Upon knowing the identity of P_3, we can remove all instances of P_3 from *HOSPITALS* and *DNA* by matching the P_3 Sequence and P_3 demographic information. As a result, we can reiterate the search for a unique occurrence of patient data and notice that, as a result of the removal of $P_{3,}$, H_2 now has an outlier of one patient in DNA presenting the disease gene and the patient reported in *DISCHARGE* with a diagnosis of the disease, so the remaining DNA sequence harbored at H_2 must originate from DGZ_2, which identifies P_2. This patient was not unique at H_2 during the first round of outlier detection with IP, due to the presence of P_3's information. Similarly, with the knowledge of P_2 identity, we can then search *HOSPITALS* and *DNA* to remove all instances of P_2. As a result, after two iterations, H_1 has only one patient in DNA presenting the disease gene and the patient reported in *DISCHARGE* with a diagnosis of the disease, so the DNA sequence must originate from DGZ_1, which identifies P_1. Following this simple process of identifying a unique occurrence, identifying it, and then purging it from the others, made it possible to render the DNA sequences identifiable.

3.5.1.2. Complexity

The problem is simplified by considering that under our model, DNA sequences are distinct to individuals, as such, the number of sequences is equal to the number of identified entities. This analysis proceeds, in terms of the matrix representation of IP. First, the binary data-location matrix *M*, sized |*PATIENTS*| × |*HOSPITALS*|, is filled for all health data profiles. This process completes in *O*(|*DISCHARGE*|) time. Concurrently, we construct and maintain an additional row *R* for the column sums, where *R*(*i*) is the column sum for row *i*. First, the row sum is parsed for an instance of a 1 in this row and, when found, we link this patient to the DNA sequence in the subtract 1 from the column sum. This step is of *O*(|*HOSPITALS*|) for the scan and *O*(|*PATIENTS*|) for the column scan. The re-identified row and column are removed and the search repeats with a table of size (|*PATIENTS*|-1) × (|*HOSPITALS*|-1). We iterate the search, re-identification, and removal process until no more re-identifications can be made. The complexity of this search strategy is approximately quadratic and is $O([max(|PATIENTS|, |HOSPITALS|]^2)$.

3.5.1.3. Re-identification Upper limit

The maximum number of patients that can be identified by IP is bounded and directly dependent on the product of the set sizes of observed values for a common attribute as shown above. The actual limit is a linear function of this product. The maximum number of re-identified patients can only be achieved when the number of distinct patients is less than or equal to the number of partitioning classes $\prod_i |a_i|$-1. In the examples hospitals above, the maximum is simply the number of hospitals considered.

3.6. Re-identification of DNA in Trails

The IP algorithm is able to re-identify through iterative outlier detection of isolated occurrences of patients in hospitals (*i.e*. P_3 in H_3). Thus, IP considers the set of individuals that visit each hospital. When a hospital is found with a single occurrence of an individual from their respective class (i.e. male or female), a linkage is made. However, the hospital trail for an individual would be more distinguishing than searching for hospital visit sizes of 1. Thus, it seems that it would be more useful to search the patterns of hospitals visited for each patient. The hospital visits made by each patient should be utilized as the set, rather than the patients at each hospital. This is exactly how the workings of the Re-identification in DNA (REID) are based. Figure 9 provides pseudocode for the basic operation of the REID algorithm, which utilizes JDBC as did IP. The actual algorithm includes some attention to assumptions made in this basic operation, but Figure 5 does provide a description of the basic approach. Valid assumptions will be clarified below.

Steps 1 and 2: Construct Trails. The hospital trails for each patient are a binary representation of the set of all hospitals considered for all individuals. This construction proceeds in the same manner as was performed Intersect-Purge.

Step 3: D**etermine Uniquess of Trail.** For each row in the matrix, if it is unique, or in other words, if there are no other patients exhibiting the same visit pattern, then the data is linked to the identical demographic pattern found in the other dataset for re-idenitification.

Algorithm: Re-idenitification of DNA	
Input	1) Patient profiles of clinical data (basic hospital visit information) 2) DNA database information from collecting institutions
Output	List of re-identified DNA database entries
Assumes	DNA entries of different individuals can be resolved
Step 1	Construct audit trail for each patient
Step 2	Construct audit trail for each DNA sequence
Step 3 **for**	Each audit trail
If	Audit trail is unique
	Attempt re-identification of audit trail

Figure 5. Pseudocode for the REID algorithm.

3.6.1. Re-identification of DNA Example

To demonstrate the power of REID, consider the more complex situation depicted in Figure 6. By applying Iterative-Purge to the dataset, there would not be any individuals that could be identified. The algorithm would not be able to initiate. There are no outliers in the first round. The algorithm simply stops after one iteration and returns no individuals.

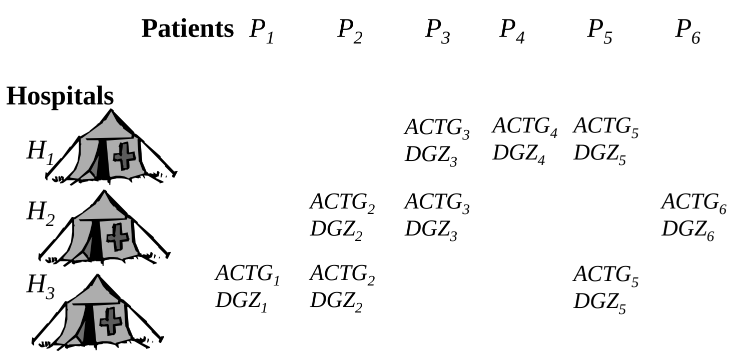

Figure 6. DNA data collection scenario. IP fails to re-identify any individual, while REID re-identifies all individuals.

Let us consider the algorithm in terms of sets of clinical information. In the first pass of the algorithm the patient sets at each hospital would be: $\{P_3, P_4, P_5\}$ for H_1, $\{P_2, P_3, P_6\}$ for H_2, and $\{P_1, P_2, P_5\}$ for H_3. Notice that the size of each set is greater than the necessary size of 1 for IP to determine the identity of an individual. However, it is useful to consider an alternate representation of the matrix and use the sets of hospitals that patients visit for each patient instead of each hospital. We are now interested in determining if the hospital trail of a patient is unique. To determine the uniqueness of the hospital trail, we can consider the occurrence of each patient at a hospital. Either the patient made a visit or the patient did not make a visit. The REID representation is expanded to incorporate all hospitals in the known set of hospitals for each patient as in Figure 7.

H1: $\{P_3, P_4, P_5\}$
H2: $\{P_2, P_3, P_6\}$
H3: $\{P_1, P_2, P_5\}$

$\longrightarrow$

P1: $\{H_3\}$
P2: $\{H_2, H_3\}$
P3: $\{H_1, H_2\}$
P4: $\{H_1\}$
P5: $\{H_1, H_3\}$
P6: $\{H_2\}$

$\longrightarrow$

P1: $\{\neg H_1, \neg H_2, H_3\}$
P2: $\{\neg H_1, H_2, H_3\}$
P3: $\{H_1, H_2, \neg H_3\}$
P4: $\{H_1, \neg H_2, \neg H_3\}$
P5: $\{H_1, \neg H_2, H_3\}$
P6: $\{\neg H_1, H_2, \neg H_3\}$

$\longrightarrow$

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Figure 7. Resolving IP ambiguities by conversion of set to matrix representation.

Direct comparison of rows in the resulting matrix of Figure 7 reveals that each individual has a unique trail that allows for a linkage to occur between identity and DNA. The example presented considers the set of hospitals that the six patients visited. Yet this set is a subset of the total number of hospitals in a geographic location, such as a state in the United States. Thus, if there exists a larger number of visited hospitals for all considered patients, the representation of the above patients would still be identifiable provided that no other patients visited the same hospitals.

3.6.2. Re-identification Upper limit

The maximum number of patients that can be identified by REID is bounded by the number of binary strings for the number of attribute combinations. Therefore, the maximum number of reidentifications is equal to $2^{\Pi_i |a_i|}-1$.

3.6.2. Complexity

The computational speed of the basic REID algorithm provided in Figure 5 is as follows. Step 1 and 2 each have a one time execution/construction of $O(|DISCHARGE|)$ and $O(|DNA|)$. We analyze the complexity of trail comparison of health data. The DNA trail analysis is a direct corollary. Step 3, executes each |PATIENT| times. For each iteration, the trail is compared to the number of patients remaining to be checked for re-identification. Thus, each trail is compared to at most |PATIENT| trails, but almost always less than this maximum. Therefore, the overall computation time is quadratic and is $O(|PATIENT|^2)$.

4 Results

The re-identification algorithms are assessed with hospital discharge data (called health data) for the state of Illinois spanning the years 1990 through 1997. In many states, such datasets are publicly available.[34] There are approximately 1.3 million hospital discharges per year in the health data, which reportedly corresponds to hospital compliance of 99+% for all discharges occurring in Illinois hospitals.[35] Diagnosis codes, procedure codes, patient demographics, and hospital identity are among the information documented for each visit. For this analysis we assume the second-generation DNA database under question results from patients that are in the hospital.

4.1. Code-Gene Relationships

Over 30 diseases caused by a mutation in a single gene were found to have distinct annotations in the discharge data at a first order search of specific names of disease and genetic characterization. A partial listing of the diagnoses that can be extracted from the hospital discharge data is presented in Table 1. The ICD-9 codes represent disease that manifest as the result of mutations in single genes. The preliminary search has not been exhausted, since the names of some diseases are classified differently in the clinical information than its genetic counterpart. Examples of such well-defined diseases with different names in the database include diastrophic dysplasia, spinal muscular atrophy (SMA), and Angelman syndrome.

#	Disease in Medical Release Data	ICD-9	Known Gene(s)
1	Adrenoleukodystrophy	3300	ALD
2	Amyotrophic Lateral Sclerosis (ALS)	33520	SOD1, ALS2, ALS4, ALS5
3	Burkitt's Lymphoma	2002	MYC
4	Chronic Myeloid Leukemia	2051 20510 20511	BCR, ABL
5	Cystic Fibrosis	27700 27701 V181 V776	CFTR, CFM1
6	Duchenne's Muscular Dystrophy (paralysis)	33522	DMD
7	Ellis-van Creveld (chondroectodermal dysplasia)	75655	EVD
8	Essential Tremor (idiopathic) (autosomal dominant account for ½ of the cases)	3331	ETM1 (FET1), ETM2
9	Familial Mediterranean Fever (amyloidosis)	2773	FMF
10	Fragile X	75983	FMR1
11	Friedrich's Ataxia	3340	FRDA
12	Galactosemia	2711	GALT
13	Gaucher's disease (cerebroside lipidosis)	2727 3302	GBA
14	Hemophilia Type A	2860	HEMA
15	Hereditary Hemmorhagic Telangiectasia	4480	H
16	Huntington's Chorea	3334	HD
17	Hyperphenylalaninemia (Phenylketonuria)	2701	PAH
18	Immunodeficiency with hyper-Igm (HIM)	27905	TNFSF5
19	Machado-Joseph Disease (Spinocerebellar Ataxia 3)	3348	MJD
20	Marfan Syndrome	75982	FBN1
21	Menkes Syndrome	75989	ATP7A
22	Methemoglobinemia	2897	HBB, HBA1, DIA1
23	Myotonic dystrophy	3592	DM
24	Pendred's syndrome (familial goiter with deaf-mutism)	243	PDS
25	Prader-Willi Syndrome	75981	SNRPN
26			
27	Refsum's Disease	3563	PAHX
28	Sickle Cell Anemia	28260	HBB
29	Spinocerebellar ataxias – or atrophy	3349	SCA1
30	Tangier disease (familial high-density lipoprotein deficiency)	2725	ABC1
31	Tay-Sachs	3301	HEXA
32	Tuberous Sclerosis (Pringle's disease)	7595	TSC1, TSC2
33	Vitelliform Macular Dystrophy (Best Disease)	36276	VMD2
34	von Hippel-Lindau (angiomatosis retinocerebellosa)	7596	VHL
35	Werner's disease or syndrome	2598	WRN
36	a) Werdnig-Hoffmann disease b) Kugelberg-Welander	3350 33511	SMA1 SMN/NAIP region
37	Wilson's Disease	2751	ATP7B

Table 1. Sample of ICD-9 code descriptions with known gene counterparts.

4.3. Re-identification of DNA with REID and IP

This section of the results utilizes discharge databases spanning the years 1990-1997. Eight different single gene disorders are analyzed including cystic fibrosis (CF), Friedrich's Ataxia (FA), hereditary hemorrhagic teleganictasia (HHT), Huntington's Disease (HD), phenylketonuria (pku), refsum's disease (RD), sickle cell anemia (SC), and tuberous sclerosis (TS). A summary of the DNA based disease cohorts is provided in Table 2.

Disease	Patients	Hospitals	Min	Max	Mean	Median	St. Dev
CF	1149	174	1	8	1.155098	1	1.805918
FA	129	105	1	5	1.126099	1	1.711538
HD	419	172	1	7	1.221283	1	1.77327
HHT	429	159	1	8	1.065105	1	1.697674
PKU	77	57	1	10	1.528985	2	2.068182
RD	4	8	2	2	2	2	0
SC	7730	207	1	34	1.822576	2	2.380466
TS	220	119	1	8	1.107661	1	1.79021

Table 2. Summary statistics of the genetic datasets studied with the re-identification algorithm.

4.3.1. Identifiability of DNA with REID and IP

Figure 8 demonstrates the identifiability of different DNA database entries based on the IP and REID systems. Results range from 0% to 100% identifiability with IP and from 33-100% identifiability with REID. As can be observed in Tables 3 and 4 and is graphically showing in figure 9, the risk of re-identifiability decreases as the number of patients per hospital increases. This is an expected result and will be elaborated upon below. Two different sets of common fields were used for this study. The first of re-identification attributes, shown in figure 14, consists of {*hospital visited*, *diagnosed disease*}. The second set includes the additional attribute gender. As expected, the addition of the sex attribute increases the identifiability of the data.

Disease	Average # Disease Patients per hospital	Percent of Cohort Re-identified	
		IP	***REID***
CF	6.60	3.22%	32.90%
FA	1.23	38.76%	68.99%
HD	2.48	11.03%	50.00%
HHT	2.70	10.49%	52.21%
PKU	1.35	33.77%	75.32%
RD	0.50	100.00%	100.00%
SC	37.34	0.32%	37.34%
TS	2.10	16.00%	51.60%

Table 3. Selection of classes used for re-identification with the IP and REID algorithms.

Disease-Gender	Average # Disease Patients per hospital	Percent of Cohort Re-identified	
		IP	*REID*
CF-F	3.92	7.18%	43.09%
CF-M	3.95	8.11%	39.36%
FA-F	0.88	80.00%	80.00%
FA-M	0.96	57.97%	78.26%
HD-F	1.26	47.59%	79.14%
HD-M	1.87	32.49%	50.63%
HHT-F	1.74	22.95%	64.34%
HHT-M	1.62	25.41%	63.24%
PKU-F	1.08	59.62%	80.77%
PKU-M	1.00	64.00%	80.00%
RD-F	0.50	100.00%	100.00%
RD-M	0.50	100.00%	100.00%
SC-F	22.09	1.34%	43.76%
SC-M	18.61	1.21%	36.51%
TS-F	1.10	58.76%	78.35%
TS-M	1.41	39.02%	61.79%

Table 4. Selection of classes blocked by gender used for re-identification with the IP and REID algorithms. (M=male, F=female)

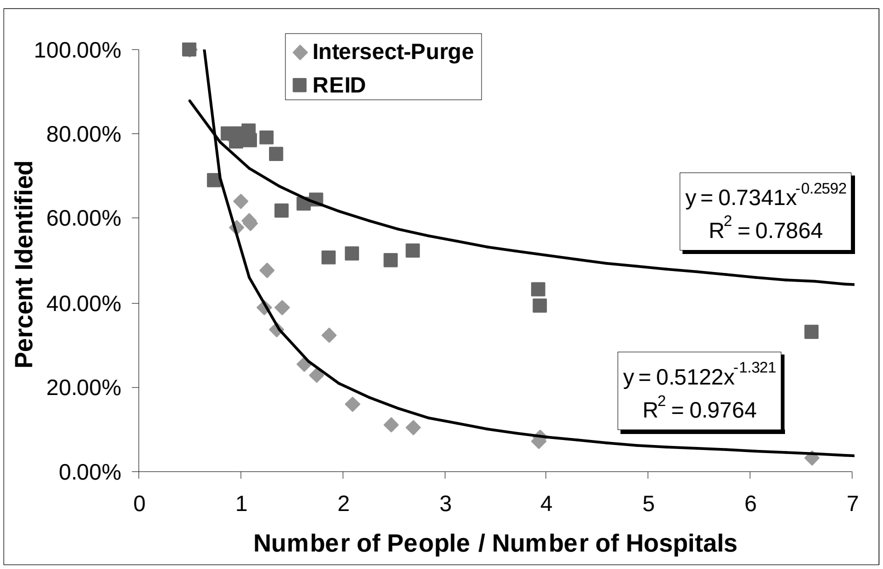

Figure 8. Relationship between re-identification capabilities of IP and REID. Complete cohorts are used with both types of classes (male, female, and no gender consideed).

The relationship between re-identifiability, the number of hospitals, and the number of distinct individuals in the discharge dataset is depicted in Figure 9. We demonstrate the previous with the IP algorithm. Notice that there is an inverse power relationship between the average number of patients per hospital and the fraction of the individuals in the discharge database that could be linked to their respective DNA database entries.

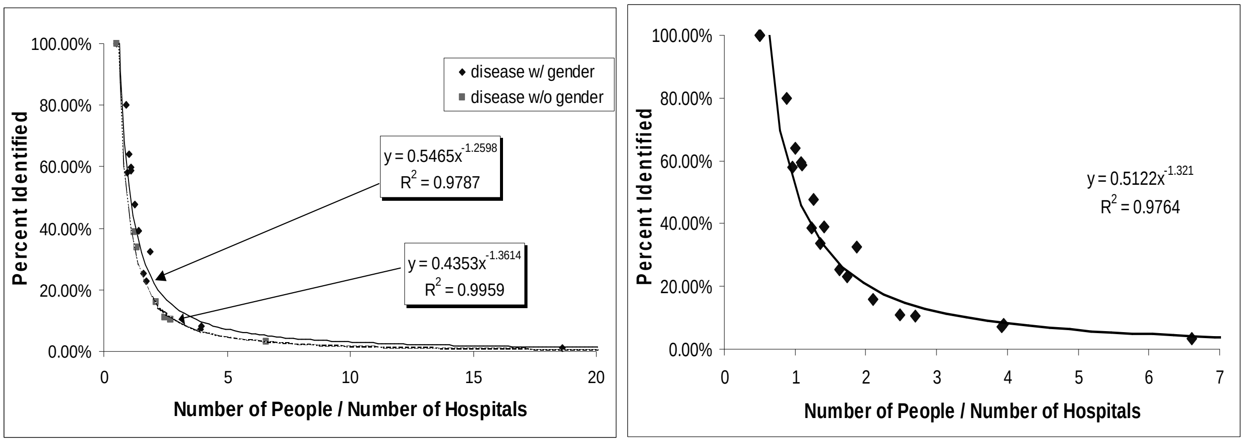

Figure 9. Identification using Intersect-Purge. left) Comparison of different class types: 1)Split on disease and gender. 2) Split on disease alone. right) Both class types 1 and 2 (from a) combined. Regression line has minimal change.

The identifiability of both class types (hospitals with and without gender) for IP and REID are plotted in Figure 8. The raw data used for these plots is directly from Tables 3 and 4. Notice, that identifiability for REID depreciates at a much slower rate than IP. By the time that IP is within 0.5% of zero identifiability, REID continues to identify over 30% of the disease cohort, in this case sickle cell anemia.

The probability of an audit trail being found to be unique is not dependent on the distribution of patients and hospitals alone. There is also a dependence on the actual numbers of people and hospitals in the system being analyzed. Intuitively, one would expect that a system with 50 patients and 50 hospitals would have more re-identifications occur than a system with 50 patients and 20 hospitals. A similar result would be expected if we varied the number of patients while holding the number of hospitals constant. This was analyzed and found to be true. The number of patients visiting a certain number of hospitals was analyzed with respect percent of the subgroup of the cohort that was re-identified. We considered the number of patients with respect to the number of available hospital audit trails. The number of audit trails was taken as the number of possible trails for the number of hospitals visited. For example, consider a cohort that visited 150 hospitals. If one hospital was visited, then 150 choose 1, or 150, possible audit trails exist. If two hospitals were were visited, then 150 choose 2, or 11175, trails possible trails exist. This is analysis is shown for the REID algorithm in Figure 10, where both the number of patients and the number of audit trails are considered in the log scale. However, while all possible audit trails are available for each patient, only a small fraction are actually observed in the dataset. This is due to the fact that as the number of hospitals visited increases, less patients actually visit this increased number. So, while the number of available audit trails increases, a lower ratio audit trails are actually used than when there were a lesser number available. For a more real world representation of re-identifiability with respect to the number of audit trails we analyzed this concept with respect to the populations from the above re-identification experiments. The results are provided in Figure 11, and follow the expected trend of increased re-identifiability.

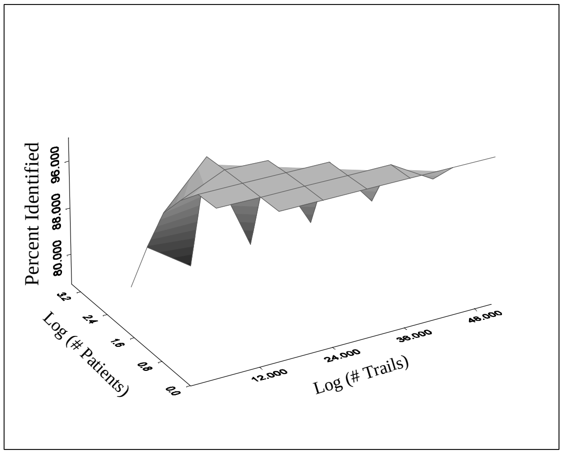

Figure 10. Re-identifiability as a function of number of patients and theoretical number of audit trails. The two plots are the same, merely in different orientations of the vertices. The data used for this plot is from all available genetic datasets. The number of patients and available number of audit trails are plotted in the log scale. As the shade of the surface plot becomes lighter, re-identifications increase.

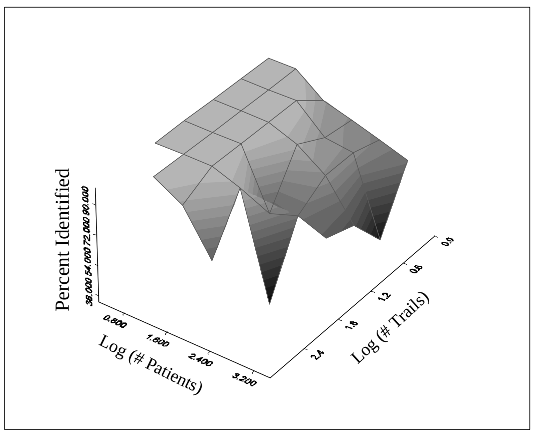

Figure 11. Re-identifiability as a function of number of patients and actual number of audit trails. The two plots are the same, merely in different orientations of the vertices. The data used for this plot is from all available genetic datasets. The number of patients and available number of audit trails are plotted in the log scale. As the shade of the surface plot becomes lighter, re-identifications increase.

5 Discussion

5.1. Hospital Audit Trails versus Single Hospital Outlier Detection

It is evident from the empirical results that the REID algorithm provides a stronger re-identification tool than the IP algorithm. However, it may not be readily apparent why REID outperforms IP. The theoretical maximum re-identifications for each algorithm are depicted in Figure 12. IP is basically an iterative outlier detection method. As such, it is necessary for an outlier to exist for IP to begin outlier detection. The outliers are a single patient with a particular disease. The number of patients that can be identified, as discussed in the upper limit analysis of section 3.5.7 is linearly dependent on the number of hospitals visited. One can never re-identify more patients than there are hospitals. However, the more common a disease is, it is obvious that the average number of patients per hospital increases. With this increase, the chances of finding a small number of patients in any particular hospital are small.

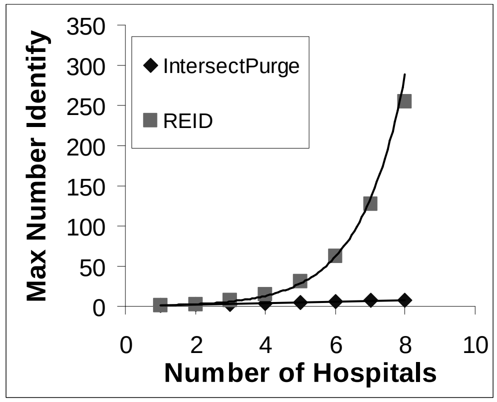

Figure 12. Theoretical maximum number of identifiable individuals for REID and IP. The number of classes considered is the number of hospitals in the dataset. REID has an exponential relationship, while IP has a linear.

REID circumvents this linear relationship by invoking a combinatorial aspect of how individual in a real world population visit hospitals. Hospitals with more than one patient presenting a particular disease should not be ruled out, since the same patients may visit other hospitals. Not all of these patients visit the same hospital. Thus, as the number of hospitals increases, the number of audit trails exponentially increases, permitting more opportunity for re-identification.

5.2. Probability of Re-identification

While the theoretical maximum re-identifiability growth is a feature that explains the difference between IP and REID, it is rarely achieved in most real world systems. One reason for the inability to reach maximum re-identification relates to the distribution of patients' hospital visits. The distribution is not uniform. Therefore, we can consider each hospital with a distinct probability of a patient visiting it. Therefore, let us consider the following representation. Let the set of patients with a particular disease be $P = \{p_1, p_2, p_3,\ldots,p_M\}$, and let the complete set of hospitals visited by such patients be $H = \{h_1, h_2, h_3,\ldots,h_N\}$. Furthermore, let the number of distinct patients visiting each particular hospital be represented by the set $X = \{x_1, x_2, x_3,\ldots,x_N\}$. If we consider hospitals independently of each other, then the probability of a patient visiting the i^{th} hospital is a simple Bernoulli probability x_i/M. Therefore, the probability of any particular audit trail a being observed is the multinomial:

$$P(a) = \prod_{i=1}^{N} \left[\frac{x_i}{M}\theta(h_i) + \left(1 - \frac{x_i}{M}\right)(1 - \theta(h_i)) \right],$$

where $\theta(x)$ is an indicator function defining the binary status of the audit trail for the institution under consideration. While this model accounts for the non-uniform distribution of individuals in hospitals, there may still exist correlation between the hospitals that are visited at the individual level.

For a more accurate representation of the probability of any particular audit trail being observed, we map the set X to a new set X'. This mapping occurs by the following measure of dependence in hospital visits. Consider two hospitals h_1 and h_2. If all of the individuals that visited hospital 1, visited hospital 2 as well, then there is complete containment of the hospital visits of hospital 2 by set of hospital 1 visits. As such, by including the hospital visits made to hospital 2, when an individual has visited hospital 1, the probability of an audit trail with visits to hospital 1 and 2 is erroneously inflated, since the same amount of information can be coded by the hospital 1 visit. Therefore, in the new set X', x'_1 is set equal to x_1, while x'_2 is set equal to x_2-x_1 to account for the double counting. We define independence to be the case when there is incomplete containment of neither for hospitals visits when compared with a second patient. Figure 13 provides a visual representation of the concept.

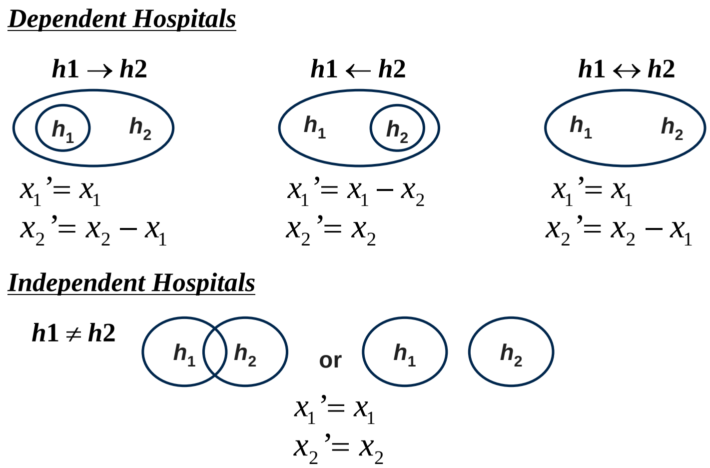

Figure 13. Graphical set representation of hospital visit dependence and independence

So, how do the probabilities change with removal of complete dependent hospital visits? This we compute as [$P(a)$-$P(a')$] / $P(a)$, where $P(a)$ is the log probability of an audit trail before removal of dependent hospital visits and $P(a')$ is the log probability of an audit trail after removal. The results of this analysis with cystic fibrosis cohort are provided Figure 14. For each audit trail the change in the probability of observing each particular audit trail from the cystic fibrosis dataset was estimated. The change in probability is normalized by the magnitude of the original probability. The magnitude was used to prevent a change in sign of the measured feature, since probabilities measured in the log scale have a negative value. One could simply have chosen to normalize by the negation of the original audit trail probability.

It is clear that for most of the observed audit trails in this dataset, the removal of dependent hospital visits reduced the probability that an audit trail would be found to be unique. To orient the reader, since the probabilities were measured in the log scale, the more negative a value, the improbable it is to observe that audit trail. Therefore, when the measured value of *Pr*(*original audit trail*)-*Pr*(*revised audit trail*) is negative, then we can say that we would expect the revised audit trail to have less of a chance of being re-identified. Similarly, if the sign of the difference was positive, then it is expected that the audit trail has an increased probability of being found unique.

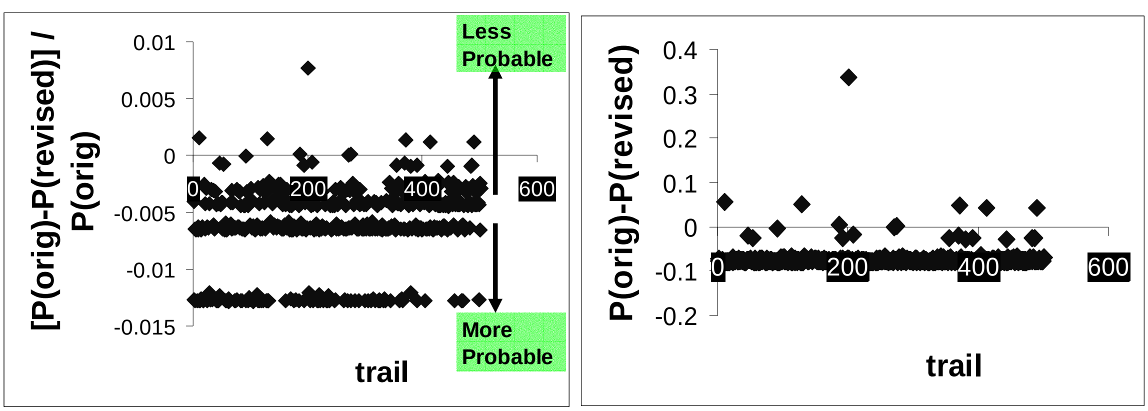

Figure 14. (left) Representation of how the probability of observing an audit trail changes with removal of dependent hospital visits. The probability changes were normalized by the magnitude of the original probability. The cystic fibrosis dataset was used. (right) Same data unnormalized.

The resulting probabilities for real data audit trail is shown in figure 15 for the disease cystic fibrosis. We find that for as the frequency of an audit trail decreases, there is an increase in the probability of that audit trail being unique. However, it must be noted, that there are substantial differences in probability for each audit trail when compared with another. As a result, the removal of the dependent hospital visits does not have a substantial change on the probability that an audit trail will be observed. We validate (not shown) that removal of the dependent hospital visits has minimal effect on the probability of an audit trail being observed. The probabilities of different audit trails have a large enough difference, such that when the dependent hospital visits are removed from the set, the revised probabilities arising from the multinomial model do not change the relative ordering of the datapoints. This is graphically depicted in Figure 15.

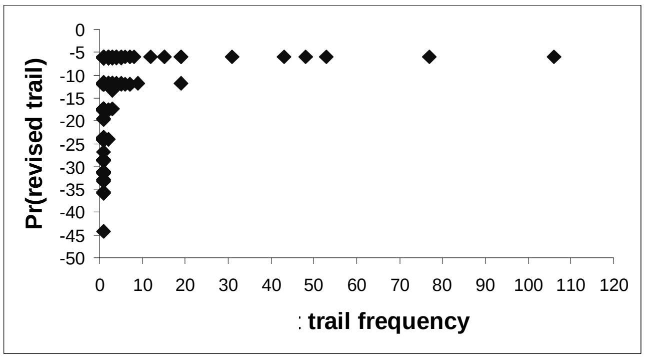

Figure 15. Probability of uniqueness for the reduced set of hospital visits and audit trail frequency (i.e. the number of times a particular audit trail was actually observed).

5.3. Other Means of Re-Identification

It should be noted that population registers and summaries of populations are not only way to establish the identity of an individual. There are many other publicly available utilities to assist in re-identification. One such utility that we explore is the Social Security Death Index Database. Through the use of this database and knowledge of the approximate date of death from health information, we can establish the identity of an individual. The date of death from health information may be directly stated in the health information or it may be predicted from the observed trajectory of the disease in question. Thus, while a longitudinal health profile may seem ambiguous when crossed with a population register, in fact there may be other information uniquely characterize the individual of interest.

5.4. Concerns for Genetic Privacy

The DNA re-identification experiments of this work demonstrate the effectiveness of constructing trails to infer additional information about the individuals in databases. Such inferences can be used for uniquely creating linkages and re-identifying the DNA information to the persons who are the subject. These re-identifications can be performed even when the DNA data itself contains no additional fields of data, such as gender. The results are further alarming because the number of common features in DNA is expected to increase with time, thereby providing more inferences to other fields of publicly and semi-publicly available data. This underscores privacy concerns that impact on the ability to conduct research [36, 37, 38], and as such, the biomedical must address such problems. Furthermore, we underscore the realization that DNA includes latent information that may be useful at a later time of study, but is not known at a particular time. Such types of information may consist of SNPs and allelic gene variants that can be used for specific treatments or additional genes that have to be discovered that play a role in susceptibility to disease.

Acknowledgements

The authors thank the insightful advice received through discussions with Robert Murphy, Rema Padman, and Victor Weedn. The authors also wish to thank the State of Illinois for the use of their data. This work was funded in part by the Laboratory for International Data Privacy at Carnegie Mellon University and the United States Bureau of the Census.

References

[1] Altman RB. Bioinformatics in support of molecular medicine. *Proc AMIA Symp*. Nov 1998; 53-61.
[2] Mendelsohn ML, Peeters JP, and Normandy MJ, eds. Biomarkers and Occupational Health – Progress and Perspectives. Washington, DC: Joseph Henry Press, 1995.
[3] Shulte PA, et. al., eds. Molecular Epidemiology: Principles and Practices. New York: Academic Press, 1993.
[4] Beroud C, Collod-Beroud G, Boileau C, Soussi T, and Junien C. UMD (Universal mutation database); a generic software to build and analyze locus-specific databases. *Hum Mutat* 2000; 15(1): 86-94.
[5] McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. Online information available at: http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html
[6] Leonard C, Chase G, and Childs G. "Genetic counseling: A consumer's view". New England Journal of Medicine 1972; 287: 433-439.
[7] Rothenberg KH. Genetic information and health insurance: State legislation approaches. *Journal of Law, Med, Ethics* 1995; 23 (312): 312-319.
[8] Baxevanis AD and Ouellette BF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. New York: Wiley and Sons, Inc, 1998.
[9] Hamosh A, et. al. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* 2000; 15: 57-61.
[10] Krawczak M and Cooper DM. The human gene mutation database. *Trends Genet* 1997; 13: 121-122.
[11] Sherry ST. Use of molecular variation in the NCBI dbSNP database. *Hum Mutat*. 2000; 15: 105-113.
[12] Tarczy-Hornoch, P, Covington ML, Edwards J, Shannon P, Fuller S, and Pagon RA. Creation and maintenance of Helix, a web based database of medical genetics laboratories, to serve the needs of the genetics community. *Proc AMIA Symp* 1998; 341-345.
[13] Kazazian HH. Overview: progress toward a new millennium of medical genetics. *Hum Mutat* 2000; 15: 2-3.
[15] Cystic Fibrosis Genetic Analysis Consortium. Cystic Fibrosis Mutation Data Base. Available at: http://www.genet.sickkids.on.ca/cftr.
[16] Anderson B and Arnason E. Iceland's database is ethically questionable. *BMJ* 1999;318: 1565.
[17] Altman RB and Klein TE. Challenges for biomedical informatics and pharmacogenomics. *Annu Rev Pharmacol Toxicol* 2002; 42: 113-33.
[18] Pear R. Clinton bans use of genetic makeup in federal employment. *New York Times* Feb 9 2000; Section A: 16.
[19] Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. *Proc AMIA Symp* 1999; 333-337.
[20] Sweeney L. Three computational systems for disclosing medical data in the year 1999. *Medinfo* 1998; 9 Pt 2: 1124-1129.

[21] Hundepool A and Willenborg L. m- and t-argus: software for statistical disclosure control. Third International Seminar on Statistical Confidentiality. Bled: 1996.
[22] Lin Z, Hewett M, Altman R. Using Binning to Maintain Confidentiality of Medical Data. *Proc AMIA Symp* 2002; 454-459.
[23] Malin BA and Sweeney LA. Determining the identifiability of DNA database entries. *Proc AMIA Symp* 2000; 547-551.
[24] Cortesi L, et. al. Comparison between genotype and phenotype identifies a high-risk population carrying BRCA1 Mutations. *Genes Chromosomes Cancer* Feb 2000; 27 (2): 130-5.
[14] Discala C, Benigni B, Barillot E, and Vaysseix G. DBCat: a catalog of 500 biological databases. *Nucleic Acids Research*. 2000; 28(1): 8-9.
[25] Selkoe DJ. Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* Apr 2001; 81(2): 741-66.
[26] National Center for Biotechnology Information. Genes and Disease webpages. Available at: http://www.ncbi.nlm.nih.gov/disease
[27] Genes and Disease. National Center for Biotechnology Information. Available at: http://www.ncbi.nlm.nih.gov/disease/
[28] International Classificiation of Diseases, 9th revision. Available at the Centers for Disease Control and Prevention: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/1998/
[29] Sweeney L. The Identifiability of Data. (*book publication forthcoming 2003*)
[30] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness, and Knowledge-based Systems*. 2002; 10 (7): 571-588.
[31] Andrew SE, et. al. "The relationship between trinucleotide (cag) repeat length and clinical features of Huntington's disease" *Nature* 1993; 4: 398-403.
[32] Brinkman RR, Mezei MM, Theilmann J, Almqvist E, and Hayden MR. The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. *Am J Hum Genet* 1997; 60: 1202-1210.
[33] Caenazzo L, Ponzano E, Greggio NA, and Cortivo P. Prenatal sexing and sex determination in infants with ambiguous genitalia by polymerase chain reaction. *Genet Test*. 1997-1998; 1(4): 289-291.
[34] Sweeney L. Weaving Technology and Policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics* 1997; 25: 98-11.
[35] "Data release overview," State of Illinois Health Care Cost Containment. Springfield: 1998.
[36] Greely HT. Iceland's plan for genomics research: Facts and implications. *Jurimetrics*. 2000; 40: 153-191.
[37] Hall MA and Rich SA. Laws restricting health insurers' use of genetic information: Impact on genetic discrimination. Am J Hum Genet 2000; 66: 293-307.
[38] Rothstein MA, ed. Genetic Secrets: Protecting Privacy and Confidentiality in the Genetic Era. New Haven: Yale University Press, 1997.