Re-Identification of DNA through an Automated Linkage Process

Bradley Malin and Latanya Sweeney

Laboratory for International Data Privacy
School of Computer Science and Heinz School of Public Policy
Carnegie Mellon University
Pittsburgh, Pennsylvania

This work demonstrates how seemingly anonymous DNA database entries can be related to publicly available health information to uniquely and specifically identify the persons who are the subjects of the information even though the DNA information contains no accompanying explicit identifiers such as name, address, or Social Security number and contains no additional fields of personal information. The software program, REID (Re-Identification of DNA), iteratively uncovers unique occurrences in visit-disease patterns across data collections that reveal inferences about the identities of the patients who are the subject of the DNA. Using real-world data, REID established identifiable linkages in 33-100% of the 10,886 cases explicitly surveyed over 8 gene-based diseases.

INTRODUCTION

DNA is understood to be as, or more, personal than a fingerprint. But having a database of only DNA entries is often believed to be anonymous because the data look anonymous. After all, if a DNA entry is not accompanied by any explicit *demographics*, how could the person who is the subject of the DNA be identified? Associating only DNA information to named persons seems impossible in this situation, yet this work demonstrates how the release of autonomous collections of DNA by hospitals, for example, can re-identify patients to their DNA.

DNA sequences are increasingly becoming a part of the patient medical record.[1] This trend is the result of several factors. First, the cost of sequencing has been declining for over a decade due to automated sequencing while the storage capacity of computers has grown tremendously yet declined in price.

Second, many diseases are increasingly being found to have a DNA component, which can be used for diagnostic confirmation of the presence or absence of a disease. In some situations it is a deterministic component to disease, such as in Huntington's disease and cystic fibrosis.[2,3] In other situations, it acts as a probabilistic component that helps to establish the chances of being afflicted with a certain disease.[4]

Third, DNA is a valuable commodity for institutions that release the information for research purposes. Many fields from population genetics, basic science, and statistics are interested in such datasets. Recently, DNA information has been of great interest to the biopharmaceutical industry, for example, where single nucleotide polymorphisms (SNPs) and allelic variants of genes have shown promise for tailoring drugs to specific genotypes.[5]

Importantly, DNA is unlike typical family history or the results of a patient's longitudinal medical record. DNA has an undetermined amount of latent information that corresponds to undiscovered genes or relationships between the genotype (DNA sequence) and phenotype (clinical observation).

The collection of DNA into these population-based databases occurs at many different kinds of institutions. Collection can be found at government research sites, such as the National Cancer Institute, which are the result of clinical trials and basic research. Other collections of DNA may be found at hospitals like Massachusetts General Hospital or Rush Presbyterian of Chicago, as the result of diagnostic testing. Databases of DNA sequences are harbored at commercial companies, such as decode Genetics, Celera Genomics, and Incyte Genomics, where the gene discovery is of high commercial value.[6,7] These DNA collections are autonomously controlled, so decisions about sharing DNA data are made locally and independently.

BACKGROUND

There have been several computational systems presented that help render data anonymous. These include Scrub[8], which locates personally identifying information in unrestricted textual documents, and the Datafly[9] and Mu-Argus[10] systems, which attempt to render field-structured person-specific databases sufficiently anonymous. Last year we introduced the CleanGene System, which addresses linear DNA information within genetic databases[11].

CleanGene computes the likelihood that a DNA database entry can be re-identified to the

1067—5027/01/$5.00

particular person from which the DNA originated. It takes genotype-phenotype relationships into account, which allow for inferences to be discerned about the expected clinical or DNA information, depending on which dataset is used as the basis for inference. When inferring clinical information from DNA, CleanGene utilizes knowledge about the type of mutation that the DNA harbors and discerns between different types of mutation to recognize a specific gene-based disease whose diagnosis code can appear in the clinical information. For example, it is well known that Huntington's disease has a strong inverse relationship between the size of the CAG triplet repeat expansion and the age of onset of the disease.[2] Thus, the repeat size estimates the age at which the diagnosis code will appear in the clinical information.

In comparison to CleanGene, this work addresses the situation more generally and does not involve any specific knowledge of genotype-phenotype relationships. Instead, this work uses the mere existence of the DNA entry in multiple data sets to draw inferences about where the person has been. The person's visit pattern is then linked to other information to explicitly identify the person. As a result, this approach is simpler than CleanGene and requires virtually no specialized knowledge.

Before we describe how this new system works and report real-world results, we will take a moment to talk about publicly available hospital data and its identifiability. The National Association of Health Data Organizations (NAHDO) reported that 44 of the 50 states (or 88%) have legislative mandates to gather hospital-level data on each patient vist.[12] Many states have subsequently distributed copies to researchers, sold copies to industry and made versions publicly available. These data collections are expected to remain available because they are not regulated by HIPAA. While the publicly available versions do not include any explicit identifiers such as name or address, they do include demographic fields such as {*5-digit ZIP*, *gender*, *date of birth*}.

Experiments were conducted to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently.[13] It was found 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*}. Matching these values against a population register, like a voter list or local census data, re-identifies the result to particularly named individuals.[14]

Therefore, the approach taken in this work is to link the visit pattern found in publicly available hospital discharge data to the pattern of entries found in multiple hospital DNA databases, thereby relating hospital discharge data to DNA entries and revealing demographics such as {*5-digit ZIP*, *gender*, *date of birth*} specific to DNA entries. The expanded results are then linked to particularly named individuals.

METHODS

This work concerns the development of a new software program named REID (Re-Identification of DNA). The methodology behind REID relies on the facts that DNA are unique to each person, has minimal change over time, and is becoming routinely collected and subsequently shared. Consider the following hypothetical scenario in which REID would operate.

In 1994, Alice visits the University of Chicago Medical Center, where her is DNA sequenced as a diagnostic test for a particular disease. Two years later, Alice receives treatment for a disorder at Rush Presbyterian Hospital (in Chicago). Once again, Alice's DNA is sequenced. At both hospitals, the linear sequence of Alice's DNA is stored in a DNA database. There may be some variation between the two sets of sequences, due to random mutation during cell division over time, as well as difference in tissue type that the DNA was procured from. However, the difference between Alice's two samples of DNA would still be more similar to each other than Alice's DNA would be to the sequences of some random individual, Bob. REID uses these patterns of where Alice's DNA appears, along with publicly and semi-publicly available hospital discharge data to relate her DNA to her by name.

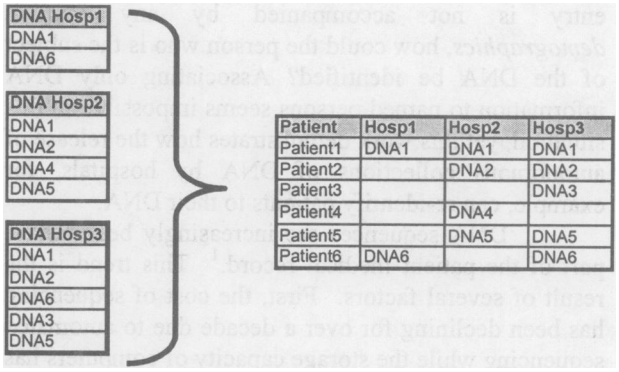

Figure 1. DNA data to Patient-Hospital Matrix

Materials

This study uses publicly available hospital discharge data from the state of Illinois. The databases cover the years 1990 through 1997, with approximately 1.3 million hospital discharges per year (each database). Collection information has compliance with greater than 99% of discharges occurring in hospitals in the state of Illinois.[15] Patient demographics, hospital

identity, diagnosis codes, and procedure codes are among the attributes stored with each database entry.

The REID system is written in Java and uses Java Database Connectivity (JDBC) to connect to a relational database, consisting of profiles for individuals with diseases that have a known DNA basis. These profiles are longitudinal, information over time, datasets, which are constructed based on the uniqueness of combinations of demographics of individuals in the hospital discharge data. Each profile consists of all inpatient visits during the eight-year time period of this study. Figure 1 provides an example.

Computer Approach

Figure 2 shows the basic operation of the REID algorithm. The actual algorithm includes some attention to assumptions made in this basic operation, but Figure 2 does provide a description of the basic approach.

The basic approach begins in Step 1 by constructing a matrix that itemizes which DNA is found at which hospitals, thereby mapping a specific patient to hospital visits based on DNA incidence. The table on the right in Figure 1 provides an example.

In Step 3, each row in the matrix is visited and compared, in step 3.2, to every other row to see if the pattern of visits is unique. If there were no other patients exhibiting the same visit pattern, then in Step 3.3, the information is linked to the identical demographic pattern found in the hospital discharge data to identify the {*Date of birth, Gender, ZIP*} specific to the DNA's incidence pattern.

The basic REID algorithm assumes that DNA and DISCHARGE are specific to the same disease gene over the same population. Today it is not the case that each hospital maintains its own DNA database for each disease gene. However, the algorithm remains the same if the hospital collects DNA for other diseases but for which the sequence includes the disease gene that is the subject of the re-identification. It is also not the case that there exists a central collection of DNA in the United States but some are underway.

Assumptions.

The basic REID algorithm also assumes that each patient has a unique {*DOB, Gender, ZIP*}. As noted earlier, this is only the case for 87% of the population of the United States. However, a ZIP chart is available that reports the identifiabilty of each ZIP, so that likelihood measures could be assigned. This is done in the full verision of REID, but not in the basic version shown in Figure 2. In the ZIP codes that are found in the real-world data on which the program was executed, the identifiability of {*DOB, Gender, ZIP*} was 98-100% unique.[11]

Input:

Table **DNA**(*HID, Sequence*), which is the union of all DNA available from hospitals specific to the disease gene. *HID* is the hospital identification number and *Sequence* is the DNA from *HID*. Table **DISCHARGE**(*HID, DOB, Sex, ZIP,...*), which is the union of all hospital discharge available for visits from the hospitals that include a diagnosis specific to the disease gene.

Output:

ID(*Sequence, DOB, Sex, ZIP*) which relates DNA sequences to identifiable demographics specific to the persons who are the subjects of the DNA sequences.

Method:

1. Construct table PATIENT(*PID*, HID_1,...,HID_n) where *PID* is a sequential number starting at 1, assigned at the construction of the table; and, each PATIENT(HID_i) is a *Sequence* from HID_i in DNA and all PATIENT(*PID=i*,HID_i) is the same sequence.
2. Let ID be empty
3. **for** $p \leftarrow 1$ **to** |PATIENT| **do**:
   - 3.1. *count* $\leftarrow 0$
   - 3.2. **for** $p2 \leftarrow p + 1$ **to** |PATIENT| **do**:
     - 3.3.1. **if** PATIENT(*PID=p*, HID_1,...,HID_n) ≡ PATIENT(*PID=p2*,HID_1,...,HID_n) **then do**:
       - 3.2.1.1 *count* $\leftarrow$ *count* + 1
   - 3.3. **if** *count* ≡ 0 **then do**:
     - 3.3.1. ID $\leftarrow$ ID $\cup$ { {*seq, dob, gender, zip*} } where for each *HID* that has a *Sequence* for *PID=p* in PATIENT, there exists exclusively $HID_i \in$ DISCHARGE having same *dob, gender, zip* associated with *DOB, GENDER, ZIP*, respectively.
4. **return** ID

Figure 2. Basic version of REID algorithm

Complexity.

The computational speed of the basic REID algorithm provided in Figure 2 is as follows. Step 2 executes each |PATIENT| times. Within each iteration, step 3.2 executes |PATIENT| times though some efficiency is realized. If the DNA incidence pattern is unique, as determined by the value of *count* being 0, then the matching pattern is sought in DISCHARGE, which requires a linear traversal through a matrix that associates patients to hospital visits; it is constructed in the same way as the DNA incidence matrix described in step 1 except rather than using DNA information, DISCHARGE is used. So, the overall computation time is $O(|\text{PATIENT}|^2)$. Therefore, on today's computers, the algorithm operates in real-time.

Upper limit.

The maximum number of patients that can be identified by REID is limited because as the number of patients increase, there can be more patients that possible combinations of hospital visits. One way this limit is avoided is to prune the DNA data and the hospital discharge data to only examine a specific disease gene, as has been referred throughout. Even still, the computational limit on the maximum number of patients able to be re-identified, assuming optimal distribution of DNA in hospital visits is: $MaxPatients = 2^{|HOSPITALS|} - 1$.

Disease Gender	# of Unique Individuals	# of Hospitals	Average # People per hospital	Percent of Cohort Identified
Huntington's disease	426	172	2.47	50.00%
Cystic Fibrosis	1146	174	6.60	32.90%
PKU	772	57	1.35	75.32%
Hereditary Hemmor. Telang.	429	159	2.70	52.21%
Friedreich's Ataxia	129	105	1.22	68.99%
Sickle Cell Anemia	7730	207	37/34	37.34%
Refsum's Syndrome	4	8	0.50	100.00%
Tuberous Sclerosis	250	119	2.10	51.60%

Figure 3. Selection of classes used for re-identification.

RESULTS

Figure 3 demonstrates the identifiability of different DNA database entries based on the REID system. Results are from 33-100% identified, with the success rate decreasing as the number of patients increase. The common fields used for this study were {*hospital visited, diagnosed disease* }. The distribution of hospital visits skew toward more visits at hospitals that specialize in the treatments of certain types of disorders, as well as the size of the hospital. Despite the coalescence of hospital visits to several hospitals, there are many hospitals with a smaller number of hospital visits. An example of the number of hospital visits for a specific disease is shown in Figure 4.

The relationship between identifiability and the number of hospitals and individuals in the discharge dataset is depicted in Figure 5. There is an inverse power relationship between the average number of patients per hospital (which is different than the average number of hospital visits per hospital) and the fraction of the individuals in the discharge database that could be linked to their respective DNA database entries.

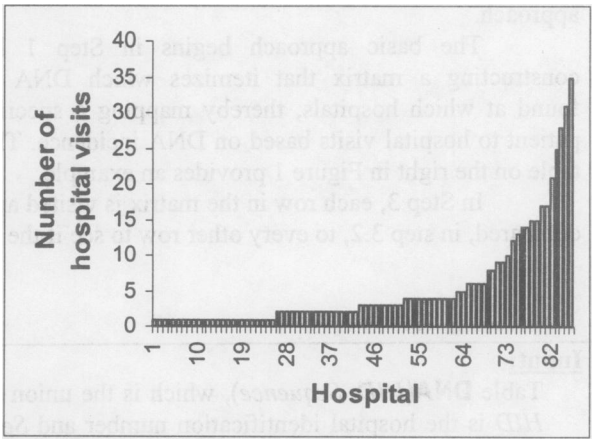

Figure 4. Distribution of hospital visits for the DNA-based disease tuberous slcerosis. The gender class is male. The visits span 1990-1997 for all hospitals in the state of Illinois.

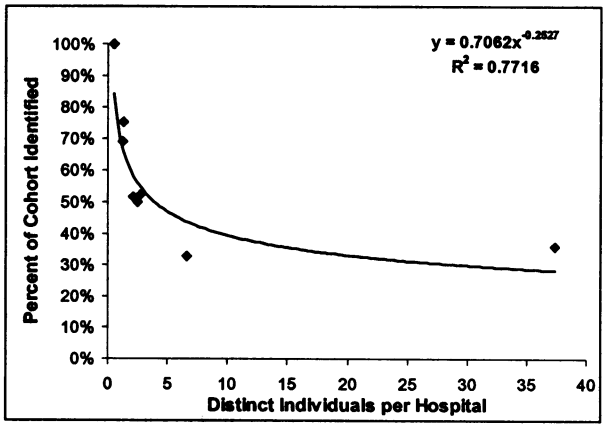

Figure 5. Demonstrates the inverse relationship between number of individuals from a re-identified dataset that can be linked to a de-identified DNA dataset.

The power relationship does demonstrate that as the ratio of hospital visits to number of hospitals visited increases, the number of hospitals with a small number of hospital visits declines.

DISCUSSION

These DNA re-identification experiments demonstrate the effectiveness of REID at finding inferences that uniquely identify DNA to the person who is the subject of the DNA even when the DNA data itself contains no additional fields of data.

The results are further alarming because the number of common features in DNA are 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[16,17,18], so these problems must be addressed. 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.

The REID system architecture is not limited to hospital discharge and DNA databases, or even medical information in general. The system is generalizes to other forms of data beyond DNA. Further, the common approaches of generalization to prevent linking[9,10] may prove to be solutions to this approach provided the DNA information remains practically useful and not all data holders make the same generalizations. Other possible solutions include random removal or addition of DNA from the data by each data holder. Finally, it is important to note that REID did not "link" values but exploited a generally observed and inferable relation, making it different than the classic privacy problem found when sharing medical data.

Acknowledgements

The authors 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 CMU and the U.S. Bureau of the Census.

References

1. Altman RB. Bioinformatics in support of molecular medicine. *Proc AMIA Symp.* Nov 1998; 53-61.
2. 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.
3. Knowles MR, Friedman KJ, Silverman LM. Genetics, diagnosis, and clinical phenotype. In *Cystic Fibrosis in Adults.* Yankaskas JR and Knowles MR, ed. Lippincott-Raven Publishers. Philadelphia. 1999. 27-42.
4. Siegmund K and McKnight B. Modeling hazard functions in families. *Genetic Epidemiology.* 1998; 15: 147-171.
5. Sherry ST, Ward M, Sirotkin K. Use of molecular variation in the NCBI dbSNP database. *Human Mutation.* 2000; 15: 68-75.
6. Hess P and Cooper D. Impact of pharmacogenomics on the clinical laboratory. *Mol Diagn.* 1999 Dec; 4 (4): 289-298.
7. Lemonick MD. Brave new pharmacy. *Time.* 2001 Jan 15; 157 (2): 58-67.
8. L. Sweeney. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In: Cimino, JJ, ed. Proceedings, *JAMIA.* Washington, DC: Hanley & Belfus, Inc., 1996:333-337.
9. L. Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, *JAMIA.* Washington, DC: Hanley & Belfus, Inc., 1997.
10. A. Hundepool and L. Willenborg. μ- and τ-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality.* Bled: 1996.
11. Malin BA and Sweeney LS. Determining the Identifiability of DNA Database Entries. *Proc AMIA Symp.* Nov 2000; 537-541.
12. National Association of Health Data Organizations, *NAHDO Inventory of State-wide Hospital Discharge Data Activities* (Falls Church: National Association of Health Data Organizations, May 2000).
13. Sweeney, L. *The Identifiability of Data.* (book publication forthcoming May 2001)
14. Sweeney LS. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine, and Ethics.* 1997; 5: 98-11.
15. "Data release overview," *State of Illinois Health Care Cost Containment.* Springfield: 1998.
16. 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.
17. Greely HT. Iceland's plan for genomics research: Facts and implications. *Jurimetrics.* 2000; 40: 153-191.
18. Rothenberg KH. Genetic information and health insurance: State legislation approaches. *Journal of Law, Med, Ethics.* 1995; 23 (312): 312-319.