Using Binning to Maintain Confidentiality of Medical Data

Zhen Lin, R.N., M.S., Micheal Hewett, Ph.D., Russ B. Altman, M.D., Ph.D.
Department of Genetics, Stanford Medical Informatics, Stanford University School of Medicine, Stanford, CA

Biomedical informatics in general and pharmacogenomics in particular require a research platform that simultaneously enables discovery while protecting research subjects' privacy and information confidentiality. The development of inexpensive DNA sequencing and analysis technologies promises unprecedented database access to very specific information about individuals. To allow analysis of this data without compromising the research subjects' privacy, we must develop methods for removing identifying information from medical and genomic data. In this paper, we build upon the idea that binned database records are more difficult to trace back to individuals. We represent symbolic and numeric data hierarchically, and bin them by generalizing the records. We measure the information loss due to binning using an information theoretic measure called mutual information. The results show that we can bin the data to different levels of precision and use the bin size to control the tradeoff between privacy and data resolution.

INTRODUCTION

Healthcare practitioners, insurance companies, and researchers all need access to clinical data. In order to prevent potential misuse of this information, such as discrimination by employers or insurance carriers, access must be tightly controlled. The administrative simplification portion of the Health Insurance Portability and Accountability Act of 1996 provides numerous new regulations that affect how health information is managed [1]. It permits only de-identified health information to be used or disclosed for research purposes, and requires research subjects to be informed of the use of their health information. At the same time, National Institutes of Health (NIH) is developing a policy on sharing research data that will expect researchers to share final research data from NIH-supported studies with other researchers [2]. Therefore, there is an increasing tension being created between patient privacy rights and research data availability. To resolve this tension, we need to ensure confidentiality of sensitive patient information while supporting relevant and important research. We describe here a method to make data available in a form that is more difficult to trace back to a specific person.

One area particularly affected by privacy issues is research into clinical applications of DNA data. The recent explosion in the availability of DNA sequences and the rapid drop in sequencing costs suggest that DNA may soon be an important source of individual data for clinical medicine. For example, the pharmacogenomics community studies correlations between genomic variations, in the form of single nucleotide polymorphisms (SNPs) or other variants, with the side effects and other adverse reactions to drugs. We are currently building a knowledge base, the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB; http://www.pharmgkb.org/) correlating genomic data and clinical data to drug response [3, 4]. Since the wide availability of data is likely to accelerate research, we would like to make this data publicly available for general research without compromising the privacy of individual patients.

Typically, researchers have access to medical data without obvious identifiers such as name or social security number. However, Sweeney and coworkers have shown that this does not ensure anonymity, since other aspects of the data may be enough to identify a person when linked with publicly available information, such as regional voting records, hospital discharge records, and motor vehicle registrations [5]. Similarly, they demonstrated that information from a locally maintained genomic database could be used to identify individuals [6].

There have been previous approaches to maintain confidentiality of medical data. SCRUB alters identifying information from clinical narrative documents by replacing it with constant pseudo information [7]. Instead of changing data, TIHI mediates access to queries and their responses based on roles [8]. For example, a cardiovascular researcher may be forbidden to access patients' HIV information. Data generalization programs such as DataFly [9] and μ-Argus [10] investigated aggregating related data into groups, called bins, to prevent access to individual data records. Others have applied statistical methods such as cell suppression, data swapping, and sampling to generalize numerical and categorical data, for instance lab results [10-12]. With the variety of types

of data in medical research, a general-purpose algorithm for anonymization is still needed.

Biomedical data are commonly hierarchical. Some well-known hierarchies are ICD9 [13], MeSH [14], Gene Ontology [15], and Enzyme Classification [16]. Our approach is to map medical data into hierarchies and anonymize them by representing specific data as more general nodes in the hierarchy. For example, we might map a diagnosis of "Ectopic Junctional Tachycardia" to "Supraventricular Tachycardia", which may be shared by more people. Similarly, we generalize "Thiopurine S-methyltransferase" to "Transferase". We generalize numerical data such as "15 years old" by constructing ranges such as "13-20 years old", which the algorithm might discover in the data or a user can specify with meaningful delineations.

Our algorithm ensures that no unique records are available to the users. We generalize data upwards in hierarchies until the values of records are shared by a user-specified number of records, called the *bin size*. This parameter controls the tradeoff between privacy and data integrity. The larger the bin size, the less chance there is for any record to be traced to a single person, yielding more privacy. Conversely, lower bin sizes result in more specific and detailed data. To achieve absolute anonymity for a particular data set may be difficult or even impossible, but we can use bin size as an indicator of different levels of anonymity.

To evaluate the information lost in the process of binning, we use an information theoretic measure called *mutual information* [17]. In this case, mutual information is the reduction in uncertainty about binning results due to knowledge of the original data set. With this metric, we can compare the performance of our algorithm with different parameters on the same data set.

METHODS

We decompose the binning problem into two steps: (1) single-attribute binning, then (2) multiple-attribute binning. An attribute is a type of data, e.g. *age*. The bin size is the number of records with the same value. The first step bins each attribute individually so that each one fulfills its bin size requirement. The second step bins each combination of attributes so that each combination fulfills its bin size requirement. The bin size for single attributes may be different than that for multiple attributes. Also, the user may specify the relative importance of the attributes to predispose the algorithm to preserve some attributes over others.

Single-attribute Binning

We bin each attribute individually so that no value in the attribute is represented by fewer records than the bin size. We first map each attribute into a hierarchical tree where the most general description of the data is at the root of the hierarchy and the most specific descriptions are the leaves. If the data are symbolic, we build this hierarchy using domain knowledge. If the data are numeric, we build a hierarchy where the leaves are individual values and higher levels in the hierarchy are broader ranges. The ranges can be specified using meaningful values (e.g. child: 0-12 years, adolescent: 13-19 years, young adult: 20-29 years, etc.) or assigned by the algorithm. Table 1 illustrates the algorithm of binning a numeric data set {3,1,2,4,5,1,6,1}, using a bin size of 2.

Steps	Example of Numeric Data
1. Sort:	{1,1,1, 2,3, 4,5,6}
2. Define range:	[1,1.5) [1.5,3.5) [3.5,7)
3. Generalize data:	{[1,1.5) [1,1.5) [1,1.5) [1.5,3.5) [1.5,3.5) [3.5,7) [3.5,7)[3.5,7)[3.5,7)}

Table 1. Binning numeric data.

Bin size: 2

1. Sort the original data points and group them so that at least as many records as the specified bin size.
2. Define ranges that contain the data values in the groups.
3. Replace each value with its range.

After the procedure, each category will hold at least as many records as the specified bin size.

Binning involves promoting records to increasing generality until each node or leaf has at least as many records as the specified bin size (or no record at all). To achieve this, for each node with too few records, we can perform one of two operations: promote one or more records from a child node, or promote all records from the current node to its parent. When there are multiple solutions that all satisfy the bin size requirement, we favor the one in which records remain at more specific nodes, since that solution preserves more information (see Figure 1).

Multiple-attribute Binning

After binning attributes individually, combinations of attributes may be unique. For example, in a data set, there may be 15 people with ages between 20 and 25, and 7 people with myocardial infarction (MI), each of which satisfies an individual attribute bin size of 5. However, there might be only one person aged 20-25 with MI. Thus, to preserve the bin size requirement, we must also consider combinations of attributes.

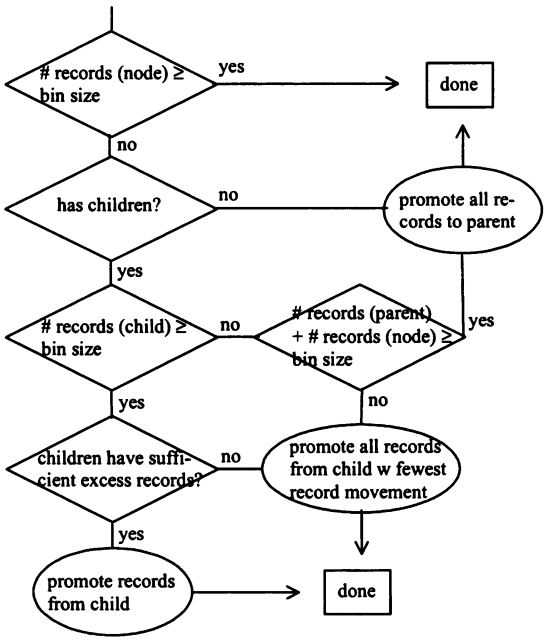

Figure 1. Algorithm to bin one node.

We adopted a few heuristics that determine the sequence of binning. Among all multiple-attribute combinations, we bin the one which is least represented in the data set first. Within a combination of attributes, the order of attributes to bin is based on predefined user preference, or the attribute with the fewest records if no preference is given. If binning an attribute does not satisfy the combination bin size, we select the next attribute instead. In any reasonably-sized database, at least one attribute will be acceptable.

Modeling Single Nucleotide Polymorphisms (SNPs)

SNPs are the most common type of genetic variation in the human genome. SNPs are chromosomal positions where two or more variant bases exist, each with at least 1% to 5% prevalence within a population [18]. SNPs can serve as genetic markers for diseases and also be used in individual identification [19]. We bin SNPs by modeling them as a type of symbolic data. We represent a SNP based on its identity and location (see Figure 2). The most detailed SNP information can be described as a specific nucleotide change on a specific location on the genome. Less detailed SNP information can be described based on indicating the type of nucleotides change, or the specific location on the genome. We can classify SNPs as transitional (a substitution between purines), or transversional (a substitution between a purine and a pyrimidine). We can generalize

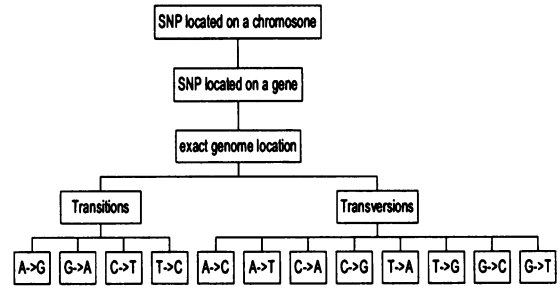

Figure 2. A SNP-type taxonomy.

SNP information even more by describing where they roughly occur only.

Experiments on a Simulated Medical Data Set

We simulated numeric data, symbolic data, and SNP data. We randomly drew numeric data from a uniform distribution. For symbolic data, we chose randomly from all the possible data values. SNP data is a special case of symbolic data, where we chose values only from the leaves of the hierarchy, since those are the observable values of SNPs.

To evaluate the performance of the algorithm, we measure the mutual information between the original and binned data. We model the hierarchical data as a probability distribution over the values at the leaves of the tree. Records in nodes are uniformly distributed across all leaves that are descendents. If p and q are the probability distributions of the original and binned data, the mutual information (I) of an attribute in the data can be obtained using the formula [17]:

$$I(p;q) = H(p) - H(p \mid q) = \sum_{x,y} r(x,y) \log_2 \frac{r(x,y)}{p(x)q(y)}$$

where H is the entropy, x and y are all possible values of the attribute, and r is the joint probability distribution. The total mutual information for a data set is the sum of the mutual information for each attribute.

RESULTS

Our first data set simulates a small set of 515 medical records containing one symbolic attribute and one numeric attribute. The numeric data contains real numbers ranging from 0 to 20. The symbolic data contains cardiovascular diagnoses drawn from a subset of the MeSH hierarchy. Our second data set simulates a SNP data set with 515 records. Each record contains a type of SNP (a leaf in the hierarchy), selected randomly.

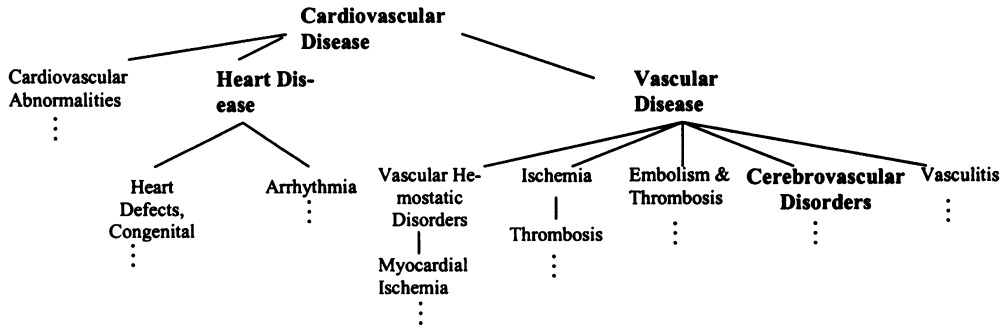

Figure 3. Example of single-attribute binning and multiple-attribute binning results on partial cardiovascular diagnosis hierarchy.

There are originally 253 nodes. These 13 nodes are those with records after single attribute binning with a bin size of 20. The 4 nodes in bold are those that still contain records after multiple-attribute binning with a bin size of 20.

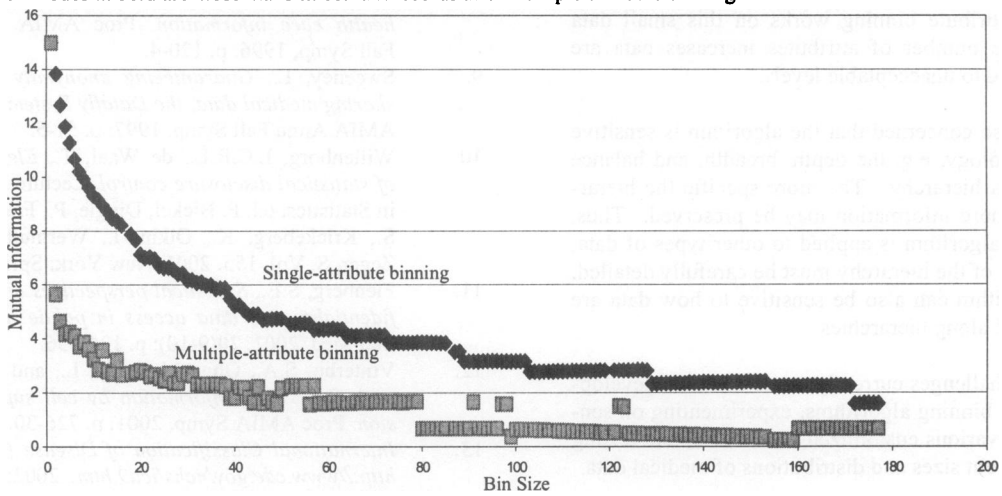

Figure 4. Mutual information decreases as bin size increases.

Single-attribute and multiple-attribute binning results on the symbolic attribute are shown in Figure 3. Single-attribute bin size and multiple-attribute bin size are both 20. Before binning, 253 nodes of the hierarchical tree have records. After single-attribute binning, 13 of them have at least 20 records. However, after multiple-attribute binning only 4 of them remain.

On our simulated cardiovascular diagnoses data set, we also applied the binning algorithm with different bin sizes. As the bin size increased, the mutual information decreased (see Figure 4). For this dataset, when the bin size is increased beyond 80, the binning algorithm provides minimal additional reductions in mutual information.

A subset of the binned SNP data is shown in Figure 5. Records are promoted along the SNP hierarchy to satisfy a bin size requirement of 10.

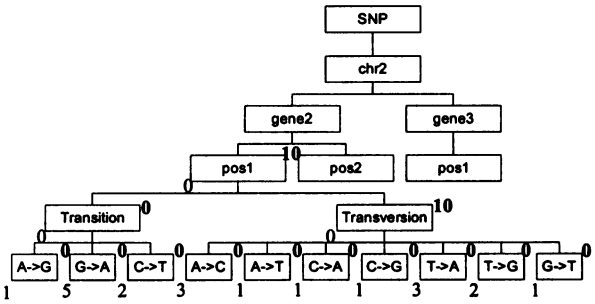

Figure 5. A subset of original SNP data and binning result.

Bin size: 10

The number at the lower left of each node is the count of records with that value in the original data set. The bold number at the upper right of each node is the count of records with that value after binning. After binning, each node has either at least as many records as the required bin size (10) or no records.

DISCUSSION

We have developed a method for disclosing numeric and symbolic medical data at various levels of precision. We can vary the bin size parameters to control the tradeoff between disclosure and privacy. To illustrate the method, we have applied it to small simulated medical data sets and have included numeric, diagnostic, and SNP data.

As the bin size requirement increases, the records may all be promoted into the roots, thus rendering the data set virtually useless for analysis. We are exploring ways to determine the appropriate bin size considering the tradeoff of information loss. Although multiple-attribute binning works on this small data set, as the number of attributes increases data are generalized to unacceptable levels.

We are also concerned that the algorithm is sensitive to the topology, e.g. the depth, breadth, and balance of the data hierarchy. The more specific the hierarchy, the more information may be preserved. Thus, when the algorithm is applied to other types of data, the design of the hierarchy must be carefully detailed. The algorithm can also be sensitive to how data are distributed along hierarchies.

The key challenges currently are to continue developing robust binning algorithms, experimenting on sensitivity to various customized hierarchies, and testing with different sizes and distributions of medical data.

ACKNOWLEDGEMENTS

ZL, MH and RBA are supported by the NIH/NIGMS Pharmacogenetics Research Network and Database (U01-GM61374). We thank Teri Klein and Mildred Cho for their support. We also thank Jeffrey Chang, Mike Liang, and Brian Naughton for useful conversations, and the reviewers for their helpful comments.

REFERENCES

1. *Standards for privacy of individually identifiable health information. Office of the Assistant Secretary for Planning and Evaluation, DHHS. Final rule.* Fed Regist, 2000. **65**(250): p. 82462-829.
2. *NIH Draft Statement on Sharing Research Data, http://grants.nih.gov/grants/policy/data_sharing/index.htm.* 2002: National Institutes of Health Office of Extramural Research.
3. Altman, R.B. and Klein, T.E., *Challenges for biomedical informatics and pharmacogenomics.* Annu Rev Pharmacol Toxicol, 2002. **42**: p. 113-33.
4. Hewett, M., et al., *PharmGKB: the Pharmacogenetics Knowledge Base.* Nucleic Acids Res, 2002. **30**(1): p. 163-5.
5. Sweeney, L., *Weaving technology and policy together to maintain confidentiality.* J Law Med Ethics, 1997. **25**(2-3): p. 98-110.
6. Malin, B. and Sweeney, L., *Determining the identifiability of DNA database entries.* Proc AMIA Symp, 2000: p. 537-41.
7. Sweeney, L., *Replacing personally-identifying information in medical records, the Scrub system.* Proc AMIA Annu Fall Symp, 1996: p. 333-7.
8. Wiederhold, G., et al., *A security mediator for health care information.* Proc AMIA Annu Fall Symp, 1996: p. 120-4.
9. Sweeney, L., *Guaranteeing anonymity when sharing medical data, the Datafly System.* Proc AMIA Annu Fall Symp, 1997: p. 51-5.
10. Willenborg, L.C.R.L., de Waal, T., *Elements of statistical disclosure control.* Lecture Notes in Statistics, ed. P. Bickel, Diggle, P., Fienberg, S., Krickeberg, K., Olkin, I., Wermuth, N., Zeger, S. Vol. 155. 2001, New York: Springer.
11. Fienberg, S.E., *Statistical perspectives on confidentiality and data access in public health.* Stat Med, 2001. **20**(9-10): p. 1347-56.
12. Vinterbo, S.A., Ohno-Machado, L., and Dreiseitl, S., *Hiding information by cell suppression.* Proc AMIA Symp, 2001: p. 726-30.
13. *International Classification of Disease (ICD), http://www.cdc.gov/nchs/icd9.htm.* 2002: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
14. *Medical Subject Headings (MeSH), http://www.nlm.nih.gov/mesh/.* 2002: U.S. Department of Health and Human Services, National Library of Medicine.
15. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
16. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.
17. Cover, T.M., Thomas, J.A., *Elements of information theory.* 1991, New York: Wiley.
18. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.
19. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nat Genet, 1999. **22**(3): p. 231-8.