Towards the Optimal Suppression of Details when Disclosing Medical Data, the Use of Sub-combination Analysis

Latanya Sweeney

Laboratory for Computer Science, Massachusetts Institute of Technology, USA

Abstract

Sharing medical data with researchers, economists, policy makers, administrators and other secondary viewers, immediately summons for consideration the dichotomy between the recipient's needs and disclosure risk. Finding the optimal balance between the suppression of details within the data needed to maintain confidentiality on the one hand, and the specificity required by the recipient in order for the data to remain useful on the other hand, is quite difficult. We present a new computational technique based on stepwise consideration of all sub-combinations of sensitive fields. This technique can be used within the Datafly or m-Argus architectures to help achieve optimal disclosure and we show that doing so provides more specific data than Datafly would normally release and improves the confidentiality of results from m-Argus.

Keywords:

Computational disclosure control; Confidentiality; Medical record linkage; Computer security; Databases

Introduction

Analysis of the detailed information contained within electronic medical records promises many advantages to society, including improvements in medical care, reduced institution costs, the development of predictive and diagnosis support systems, and the integration of applicable data from multiple sources into a unified display for clinicians; but these benefits require sharing the contents of medical records with secondary viewers, such as researchers, economists, statisticians, administrators, consultants, and computer scientists, to name a few. The public would probably agree these secondary parties should know some of the information buried in the record, but such disclosure should not risk identifying patients.

We can certainly remove all explicit identifiers such as name, address and phone number. The de-identified result however, is often far from anonymous [1,2] because anonymous implies that the data cannot be manipulated or linked to identify any individual. Consider Table 1 for example. If the contents of this table are a subset of an extremely large and diverse database then the three records listed in Table 1 may appear anonymous. Suppose the US postal code (called a ZIP code) 33171 primarily consists of a retirement community; then there are very few people of such a young age living there. Likewise, 02657 is the ZIP code for Provincetown, Massachusetts USA, in which we found about 5 black women living there year-round. The ZIP code 20612 may have only one Asian family. In these cases, information outside the data identifies the individuals. Parallel examples are found in other countries including the Netherlands, Italy and the United Kingdom [2].

Further, de-identified information can be linked and matched to publicly-available population registers such as city directories, local census data and voter registration lists that include birth date and postal code along with the accompanying name and address of each person. Population registers can be used to re-identify de-identified data since other personal characteristics, such as gender, date of birth, occupation and postal code, often combine uniquely to identify individuals.

Table 1. De-identified data that is not anonymous.

ZIP Code	Birthdate	Gender	Ethnicity
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Background

In 1996, The European Union, working with statistical offices and universities from the Netherlands, Italy and the United Kingdom, began a project to develop specialized software for disclosing data such that the identity of any individual contained in the released data cannot be recognized. Statistics Netherlands produced a first version of the resulting program which is named m-Argus [2]; however, we must note that Statistics Netherlands considers this first version of m-Argus a rough draft. A presentation of the concepts on which m-Argus is based can be found in Willenborg and De Waal [3].

In 1997, Sweeney presented the Datafly System which evolved from the practical needs of hospitals, agencies and other medical data collectors in the United States to disclose data to researchers while still maintaining patient

confidentiality [1]. Unlike the m-Argus system which is designed for the release of public use data from national repositories, the Datafly System can also work with small, specialized databases that are collected and controlled autonomously. Since Datafly can control all access to the underlying medical database, it can be used in role-based security within an institution, as well as, in batch mode for exporting data from an institution.

Both m-Argus and Datafly make decisions based on the notion of a minimal bin size. Any subset of the data that can be defined in terms of combinations of characteristics must contain at least *n* individuals. So, a minimal bin size reflects the smallest number of individuals matching the characteristics and is quite useful in providing a degree of anonymity within data. The larger the bin size, the more anonymous the data. As the bin size increases, the number of people to whom a record may refer also increases, thereby masking the identity of the actual person.

The user requests a minimal bin size. To achieve it, both systems generalize values within fields as needed and remove extreme outlier information from the released data, where outliers are extreme values not typical of the rest of the data. Unsafe combinations of sensitive fields are eliminated by generalizing fields within the combination and by removing or suppressing data. The Datafly System removes entire records when one or more fields contain outlier information. The m-Argus System simply suppresses or blanks out the outlier values at the cell-level; this process is called cell suppression. The resulting data from m-Argus typically contain all the rows and columns of the original data though values may be missing in some cell locations. Results from Datafly can have fewer records or fields than was originally requested.

Preliminary comparisons between results from the m-Argus and Datafly systems tend to show that Datafly provides more confidentially secure data while m-Argus includes more specificity within the data [4]. The goal of this work is to provide a technique that can be used within these two architectures such that Datafly results would retain more detail and m-Argus results would become more secure.

Methods

Datafly is a program that interfaces a user with an Oracle server, which in turn, accesses a medical database. Datafly was written using Symantec C and Oracle's Pro*C Precompiler. It processes all queries to the database. We replaced Datafly's control algorithm with one we implemented based on a technique we term sub-combination analysis. Results are reported using both the original version of Datafly as well as this modified version.

Sub-combination analysis examines all combinations of sensitive fields within a set of such fields to achieve a minimal bin size. Given a set containing *f* number of fields, all *r*-combinations are examined, where *r* starts at 2 and proceeds sequentially to *f* ensuring at each iteration stage that the minimal bin size is achieved for all records in all *r*-combinations of fields. Otherwise, cell suppression of outliers or generalization of all values within a field occurs until the result is achieved before proceeding to the *(r+1)* combinations.

Once a cell is suppressed, it can match any value, and so suppressed cells are often counted more than once to ensure each combination adheres to the minimal bin size. The choice of which cell to suppress is determined optimally by giving precedence to values which occur most often in multiple outliers. In a final step, we suppress another value in each field that contains an isolated suppression to further mask the identity of the outlier. Cells selected for complementary suppression come from the most popular combinations.

The end result is guaranteed to achieve the minimal bin size across all combinations and sub-combinations of fields within the set of sensitive fields. Of course, a database may have many sets of sensitive fields. For example, there may be a set of sensitive demographic fields, a set of sensitive diagnosis fields, and so forth, where sensitivity is determined by the likelihood the field can be used to link to other known data. Sub-combination analysis is applied to each of these sets. In the next sections, we will examine how sub-combination analysis can be used within the Datafly system. Following that, we will discuss some overall results and shortcomings with this technique.

Table 2. There is only one Caucasian female even though there are many females and many Caucasians. There is also only one Caucasian male born in 1964 that resides in postal code 02138.

SSN	Ethnicity	Birth	Sex	ZIP
819181496	Black	9/2/65	m	02141
195925972	Black	2/1/65	m	02141
902750852	Black	1/8/65	f	02138
985820581	Black	8/4/65	f	02138
209559459	Black	1/7/64	f	02138
679392975	Black	2/4/64	f	02138
819491049	Caucasian	1/5/64	m	02138
749201844	Caucasian	3/1/65	f	02139
985302952	Caucasian	8/3/64	m	02139
874593560	Caucasian	5/5/64	m	02139
703872052	Caucasian	2/6/67	m	02138
963963603	Caucasian	3/9/67	m	02138

Table 2 provides an example wherein there is only one occurrence of a female Caucasian even though there are many females and many Caucasians. This unique record can be identified using the 2-way combination, Caucasian female. Table 2 also contains a unique record involving a 4-way combination of a Caucasian male born in 1964 that lives in the 02138 ZIP code. In the next subsections we will show the original results from the Datafly and m-Argus systems using this data and how the results are changed

when sub-combination analysis is employed within these architectures.

The Datafly System

In the Datafly System, results are based on a profile of the recipient. Clearly, the profiles of a doctor caring for a patient, a clinical researcher studying risk factors for heart disease and a health economist assessing the admitting patterns of physicians are all different, so it is not surprising that Datafly quantifies associated "trust" with sensitive data. For the purpose of this analysis, we will profile the data as being made available for public use which to Datafly represents full distrust of the recipient and maximum concern over the sensitivity of the fields contents.

As a result, the fields Ethnicity, Birth, Sex and ZIP will be considered one concatenated field. The US Social Security number (SSN) is a unique identifier assigned to individuals in the US. The SSN field is not included in Datafly's concatenated field since generalizing the field simply involves replacing all values with made-up alternatives. We will assume that the required minimal bin size is 2 and Datafly can drop no more than 10% of the total number of records (*N*) to achieve this minimal bin size.

Table 3a. Fields from Table 2 that can be generalized are considered one large concatenated field. In this case the required bin size is 2 with a maximum loss of 10%, so further generalization must be done even though the Birth field has been generalized to the year.

Ethnicity	Birth	Sex	ZIP	BinSize	% of N
Caucasian	1965	f	02139	1	8%
Caucasian	1964	m	02138	1	17%
Black	1965	m	02141	2	33%
Black	1965	f	02138	2	50%
Black	1964	f	02138	2	67%
Caucasian	1964	m	02139	2	83%
Caucasian	1967	m	02138	2	100%

Datafly reviews the values in the concatenated fields and discovers that the minimal bin size is not attained, even though values in the fields Ethnicity, Sex and ZIP do adhere to the minimal bin size if evaluated separately. The Birth field has the largest number of bins, so values in the Birth field are generalized to the month which still fails, and then to the birth year. Table 3a shows the analysis at this intermediate stage. From there, ZIP has the most number of bins, so ZIP is generalized. The final analysis appears in Table 3b and the overall result is Table 3c.

Table 3b. The ZIP field from Table 3a has the largest number of bins, so it is generalized in order for the entire concatenation of fields to achieve the minimal bin size. The record containing the Caucasian female remains an outlier; it is not released.

Ethnicity	Birth	Sex	ZIP	BinSize	% of N
Caucasian	1965	f	02130	1	8%
Black	1965	m	02140	2	33%
Black	1965	f	02130	2	50%
Black	1964	f	02130	2	67%
Caucasian	1967	m	02130	2	100%
Caucasian	1964	m	02130	3	17%

Table 3c. Results from applying the Datafly System to the data in Table 2. The minimum bin size is 2. The profile identifies only the demographic fields as being likely for linking. The Caucasian female record was dropped as an outlier.

SSN	Ethnicity	Birth	Sex	ZIP
902387250	Black	1965	m	02140
197150725	Black	1965	m	02140
486062381	Black	1965	f	02130
235978021	Black	1965	f	02130
214684616	Black	1964	f	02130
135434342	Black	1964	f	02130
458762056	Caucasian	1964	m	02130
860424429	Caucasian	1964	m	02130
259003630	Caucasian	1964	m	02130
410968224	Caucasian	1967	m	02130
664545451	Caucasian	1967	m	02130

Using Sub-Combination Analysis

Incorporating sub-combination analysis into the Datafly System is straightforward. Instead of concatenating the set of related sensitive fields, the sub-combination algorithm does the following. First, it examines all pairs of fields in the original data. Since there are 5 fields, there are 10 combinations. The 6 combinations that do not include the SSN field are shown in Table 4a. The value of each pair is basically a bin, and the bins with occurrences less than the minimum required bin size are considered unique and termed outliers. Clearly for all combinations that include the SSN, all such pairs are unique. One value of each outlier pair must be suppressed. For optimal results, we suppress values which occur in multiple outliers where precedence is given to the value occurring most often. Processing continues by looking at all 3-combinations to see if they adhere to the minimal bin size. This leads to the suppression of 02138 in the unique record of the Caucasian male born in 1964. Then all 4-combinations are examined with no further suppressions required. As a final step, we suppress another value in each field that contains a single suppression to further mask the identity of the outlier. The final result is shown in Table 4b.

Table 4a. The first stage of processing using sub-combination analysis in the Datafly system where all 2-combinations of data are analyzed for outliers, which are highlighted below.

Ethnicity	Birth
Black	1965
Black	1965
Black	1965

Ethnicity	Sex
Black	m
Black	m
Black	f

Ethnicity	ZIP
Black	02141
Black	02141
Black	02138

Black	1965	Black	f	Black	02138
Black	1964	Black	f	Black	02138
Black	1964	Black	f	Black	02138
Caucasian	1964	Caucasian	m	Caucasian	02138
Caucasian	1965	Caucasian	f	Caucasian	02139
Caucasian	1964	Caucasian	m	Caucasian	02139
Caucasian	1964	Caucasian	m	Caucasian	02139
Caucasian	1967	Caucasian	m	Caucasian	02138
Caucasian	1967	Caucasian	m	Caucasian	02138

Birth	Sex
1965	m
1965	m
1965	f
1965	f
1964	f
1964	f
1964	m
1965	f
1964	m
1964	m
1967	m
1967	m

Birth	ZIP
1965	02141
1965	02141
1965	02138
1965	02138
1964	02138
1964	02138
1964	02138
1965	02139
1964	02139
1964	02139
1967	02138
1967	02138

Sex	ZIP
m	02141
m	02141
f	02138
f	02138
f	02138
f	02138
m	02138
f	02139
m	02139
m	02139
m	02138
m	02138

Table 4b. Final results from using sub-combination analysis within the Datafly architecture to the data in Table 2.

SSN	Ethnicity	Birth	Sex	ZIP
902387250	Black	1965	m	02141
197150725	Black	1965	m	02141
486062381	Black	1965	f	02138
235978021		1965	f	02138
214684616	Black	1964	f	02138
135434342	Black	1964	f	02138
458762056	Caucasian	1964	m	
976245951		1965	f	
860424429	Caucasian	1964	m	02139
259003630	Caucasian	1964	m	02139
410968224	Caucasian	1967	m	02138
664545451	Caucasian	1967	m	02138

While both versions of the Datafly system generalized values in the Birth field to the birth year, much more detail remained in the ZIP field of the version that used sub-combination analysis. Though each record in the original Datafly results, as shown in Table 3c, probably maps to many more people than those in Table 4b, both releases adhere to the user's specified minimal bin size.

The m-Argus System

Using the m-Argus System, the user provides an overall bin size and specifies which fields are sensitive by assigning a value between 0 and 3 to each field. The first step is to check that each identifying field adheres to the minimum bin size. Then, pairwise combinations are examined for each pair that contains the "most identifying" field and those that contain the "more identifying" fields. Finally, 3-combinations are examined that include the "most" and "more" identifying fields. The result is shown in Table 5. Obviously, there are many possible ways to rate these identifying fields, and different identification ratings yield different results. The ratings presented in this example produced the most secure result using the m-Argus program though admittedly one may argue that too many specifics remain in the data for it to be released for public use. For example, the unique record for the Caucasian male born in 1964 that resides in 102138 remains in the results. This could be as identifying as a Black female in Provincetown, Massachusetts whose uniqueness was discussed in Table 1.

Table 5. Results from applying the m-Argus system to the data in Table 2. The minimum bin size is 2. The profile for fields was: SSN, "most identifying;" birth, sex and ZIP, "more identifying;" and, ethnicity, "identifying." The uniqueness of the Caucasian female is suppressed; but, there remains a unique record for the Caucasian male born in 1964 in 02138.

SSN	Ethnicity	Birth	Sex	ZIP
	Black	1965	m	02141
	Black	1965	m	02141
	Black	1965	f	02138
	Black	1965	f	02138
	Black	1964	f	02138
	Black	1964	f	02138
	Caucasian	1964	m	02138
		1965	f	
	Caucasian	1964	m	02139
	Caucasian	1964	m	02139
	Caucasian	1967	m	02138
	Caucasian	1967	m	02138

Incorporating sub-combination analysis into the m-Argus System involves replacing the sensitivity measures and the control structure that dictates which combinations are examined. Instead, fields are grouped into sets of sensitive fields as in the Datafly System. Then, all combinations are examined based on sub-combination analysis. The final result is the same as with the Datafly System; refer to Table 4c. In comparison, Table 5 from the original m-Argus System still contains a unique record for a Caucasian male born in 1964 that lives in the 02138 ZIP code since there are 4 characteristics that combine to make this record unique, not 2. In Table 4c, sub-combination analysis properly suppressed cells that identified both uniquely-identifying records.

The biggest difference using sub-combination analysis in the Datafly System versus in the m-Argus System concerns the SSN field and user control over suppression and generalization. The responsibility of when to generalize and when to suppress lies with the user in m-Argus and is automatically determined in Datafly. For this reason, the m-Argus program operates in an interactive mode so the user can see the effect of generalizing and can then select to undo the step.

Suppressing SSN values makes little difference in this example. However, when working across multiple tables, the ability to link data across tables within the database to the same person is lost without consistent replacement of identifiers which provide such links. In fairness to m-Argus, the current version does not work across multiple tables and as a result it does not take into account these issues. Datafly on the other hand, provides consistent one-way hashing of unique identifiers.

Results

The database we used was a de-identified subset of a pediatric medical record system [5]. It consisted of 300 patient records; we were primarily concerned with demographic, provider, diagnosis, and procedure fields commonly used for linking released data to other known data. We measured results from both the regular Datafly System and the version of Datafly that incorporated sub-combination analysis in terms of entropy. This measure provides a means to express data quality.

Table 6a. Data quality measures for the original Datafly System.

bin size	3	12	27
Gender	420	420	420
VisitDate	2097	2089	2067
Ethnicity	899	560	546
Diagnosis	2740	1481	1490
Birthdate	1692	1385	1120

Table 6b. Data quality measures for the Datafly System that incorporated sub-combination analysis.

bin size	3	12	27
Gender	600	600	600
VisitDate	2100	2100	2100
Ethnicity	900	900	600
Diagnosis	3000	1500	1500
Birthdate	1800	1500	1200

For each field in the original database, we counted the number of different bins in each field. The quality is simply the total number of bits required to account for all bins in all fields in all records. When generalization of a field occurs, the number of bins decreases and likewise the number of bits required to represent those bins decrease. When outliers are dropped, the values of those bins are no longer included in the total count as well. So the higher the resulting value, the better the data quality. The results over a range of minimal bin sizes are shown in Tables 6a and 6b. Clearly, sub-combination analysis produced more detailed data at each setting. The values in Table 6b are round numbers since there were always 300 records released, but the regular version of Datafly could drop records containing outlier values.

Discussion

In concluding, using sub-combination analysis offers many advantages but there are a few caveats. First, Datafly produces results in real-time O(N log N), but using sub-combination analysis, processing time grew exponentially. Second, producing the most optimal suppressions in cases where most if not all of the data is being released can reveal the values of previously suppressed cells since suppressions may change with subsequent releases.

Acknowledgments

The author is grateful to God for the opportunity to present this work. The author thanks Beverly Woodward, Ph.D., Professor Peter Szolovits, Isaac Kohane, M.D. Ph.D., and Sylvia Barrett. This work was supported by National Library of Medicine Grant 1-T15-LM07092.

References

[1] Sweeney, L. Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, *American Medical Informatics Association.* Nashville: Hanley & Belfus, Inc, 1997.

[2] Hundepool, A. and Willenborg, L. m- and t-argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality.* Bled: 1996.

[3] Willenborg, L. and De Waal, T. *Statistical disclosure control in practice.* New York: Springer-Verlag, 1996.

[4] Sweeney, L. Computational disclosure control for medical microdata, the datafly system. *Record Linkage Workshop.* Washington: Bureau of the Census, 1997.

[5] Kohane, I. Getting the data in: three-year experience with a pediatric electronic medical record system. In: Ozbolt J., ed. Proceedings, *Symposium on Computer Applications in Medical Care.* Washington, DC: Hanley & Belfus, Inc, 1994:457-461.

Address for correspondence

Email: sweeney@medg.lcs.mit.edu
URL: http://www.medg.lcs.mit.edu/people/sweeney/