MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press
© 2004 IMIA. All rights reserved

Privacy-Preserving Data Releases for Health Report Generation

Claus Boyens[a], Ramayya Krishnan[b], Rema Padman[b]

[a] *Insitute of Information Systems,Humboldt University Berlin, Germany*
[b] *The Heinz School of Public Policy & Management, Carnegie-Mellon University, Pittsburgh, PA, USA*

Abstract

Regional healthcare initiatives seek to improve the quality of healthcare by collecting, analyzing, and disseminating information about chronic diseases such as diabetes. The data required to support such initiatives comes from several organizations such as insurers, physicians, hospitals, pharmacies and labs each of which gather and maintain data for the purpose of healthcare delivery. In this paper, we focus on mediator-based architectures and the privacy problems that arise in the healthcare context owing to the linkage of information about patients, physicians, and diseases enabled by the mediator. In particular, we examine privacy issues for the two separate steps of the actual data release. First, raw data is released to the (not necessarily trustworthy) mediator and second, the mediator creates and releases the health report. For both steps, we present a technical solution that permits the final report to be useful to the user while respecting the data owners' privacy.

Keywords:
Privacy, untrusted mediator, interval inference.

Introduction: Creating Regional Health Reports

Regional healthcare initiatives have recently been created to improve the quality of healthcare in their communities. Among other reasons, this development is driven by high numbers of hospital-acquired infections and by increasing hospitalization rates for people with chronic diseases such as diabetes.

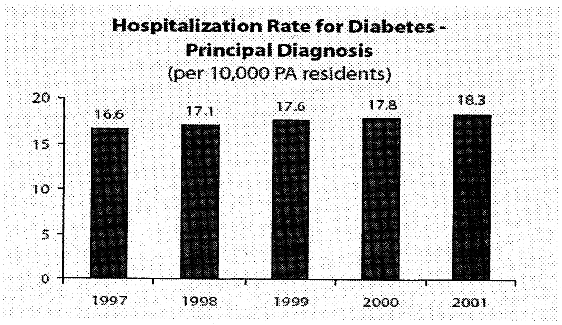


Figure 1 - A driver underlying the creation of healthcare initiatives

Figure 1 shows this development for the state of Pennsylvania that has alarmed the healthcare community and prompted the creation of these community-wide healthcare initiatives [1].

With regard to diabetes, it is widely believed that adequate diagnostic and preventive measures help reduce short-term complications. Hence, an important indicator for adequate care is the participation of affected patients in preventive screenings for Hemoglobin A1c (HbA1c), LDL cholesterol levels, foot and eye exams.

Each screening is required to be undertaken on a recommended schedule (e.g. eye exams once a year). The aim of a regional healthcare initiative is to increase the use of these preventive screenings among all affected patients by compiling and releasing information about compliance rates. Measuring these compliance rates is a difficult task because generally, the delivery of healthcare services involves many different parties such as physicians, pharmacies, laboratories, and insurers such as health maintenance organizations (*HMOs*). Hence information about patients, disease diagnosis, medications, prevention, and treatment methods is often distributed among heterogeneous databases.

The integration of these heterogeneous data sources with the objective of supporting community-wide data access is an important problem and has been addressed by a number of researchers [2, 3, 4, 5, 6]. Approaches range from the creation of data warehouses (see work on CATCH, a data warehouse in support public health in [2]) to the use of mediator-based architectures. In this paper, we focus on mediator-based approaches. Whereas cost and security considerations have usually been taken into account in prior work on mediators, we are more concerned with the privacy implications that can be an outcome of the (desired) data linkage and data fusion enabled by the mediator. Referring to our diabetes case, the involved parties may have different concerns with possible outcomes of the analysis.

The *patient* may principally be afraid of a central pooling of her data because the disclosure of a formerly unknown disease might adversely affect life insurance premiums.

The *physician* may be confronted with the fact that his test compliance rates differ significantly among patients of different age, race, income, gender, and insurance plan.

The *HMO* may fear that detailed internal data may be inferred by competitors and used in marketing campaigns.

A *laboratory* may be uncomfortable with the fact that its test analysis times differ significantly among HMOs (although the same fee is charged).

As depicted in Figure 2, these concerns have to be considered for two subsequent data releases. First, the data holders have to provide raw data to the mediator, such as the one run by a regional healthcare initiative, by allowing it, for example, to query their databases. This does not constitute a privacy issue as long as the mediator itself is trusted. But often data owners do not want to give away their most confidential data at all [7], and adequate technical and legal measures have to be instituted.

Analyzing the raw data, the mediator now creates the health report. It has to ensure that the generated report in support of community health does not permit inferential disclosure of information that is private and confidential. The publication of this report is the second data release

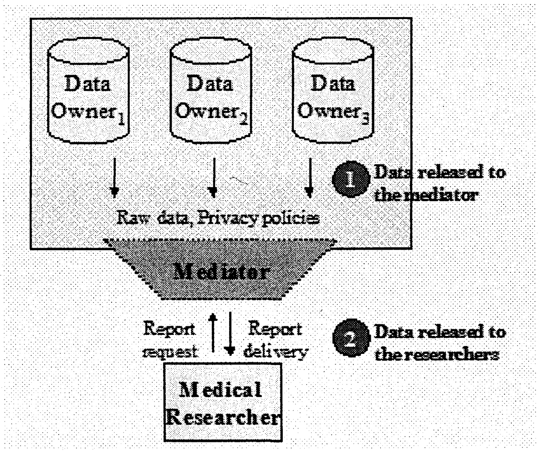


Figure 2 - Two types of health data releases

Recent work in the data mining literature considers the related problem of privacy-preserving data mining [8]. In this literature, the problem addressed is one where in a central server (in our setting, the mediator) needs to learn the statistical properties of data supplied by clients (in our setting, the data owners) in a privacy preserving manner. Of particular interest to us are the ways in which statistical properties such as measures of central tendency (mean, median etc.) and measures of dispersion (standard deviation, variance, range) are computed in a privacy preserving manner.

While we focus in this paper on issues related to data release 2, there are direct connections between data release 1 and data release 2. The objective in data release 1 is to gather the statistical properties of the data from *each* data owner without obtaining access to their raw data. A recent approach of direct relevance to our work is [9]. The focus in data release 2 is to release aggregated statistical properties of all of the data from all the data owners in such a manner as to prevent interval inference. Thus, if a technical solution were to be found to both data release 1 and data release 2, the architecture would support privacy preserving acquisition of statistical properties of the data from each data owner and privacy preserving dissemination of aggregate statistical properties across all data owners.

In the next section, we discuss privacy problems in the context of the first data release and how they can be accounted for. The succeeding section is dedicated to the second data release, the creation of the health report by the mediator, which is the main focus of this paper. We propose an "audit & aggregate" methodology to detect and limit interval inference, an important case of privacy violation. We conclude with a discussion of our results and with an outlook on further research.

Releasing Data to the Mediator (Data release 1)

The inference of information that is private and confidential based on the final health report is just one aspect of data release that has to be considered in the design of the mediator architecture shown in Figure 2. Some physicians, pharmacies, laboratories, or HMOs may already be hesitant or even reluctant to pass on their confidential data to the mediator. In particular, when this data is stored by an online provider of the mediating service, threats to the raw data are many.

- External attacks directed at the service provider's database are still possible, and the risk is hard to estimate.
- Malicious staff on the provider's side (bribed or disgruntled employees etc.) may want to cause harm to their company and its customers.
- Incompetent staff on the provider's side may unintentionally grant data access to unauthorized parties.
- The potential consequences of bankruptcy or change of ownership of the provider may be serious. In the worst case, a direct competitor of one of the provider's customers might end up owning all the outsourced business data.

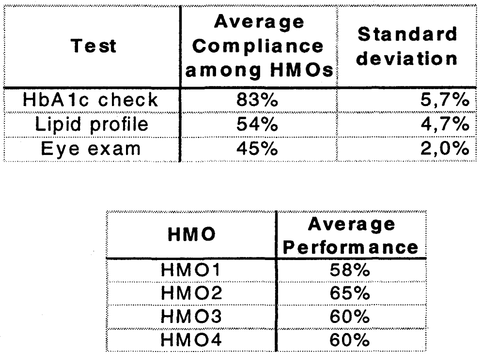
In this case, only encrypting and/or anonymizing data may alleviate these threats.

Encryption means that the transferred data is not readable to the mediator anymore, hence processing opportunities are very limited [7]. *Anonymization* would mean de-identifying a data record from its owner. For instance, HMOs would send test compliance rates while suppressing their identity. This would still allow the mediator to calculate aggregates (like averages) but would make it impossible for the patient to distinguish between HMOs and to pick the one that would suit their needs best.

The only remaining alternative is to certify the mediator as a *trusted third party* that does not store the data persistently and only uses it for report generation. In the next section, we will assume the existence of such a trusted mediator.

Releasing data to researchers via a mediator (Data release 2)

In particular, we focus on the interval inference problem for sensitive HMO data. To illustrate the relevance of this problem, consider the following information about test compliance rates in 2001 as shown in Figure 3. It is in part based on real-world data taken from [1].



Test	Average Compliance among HMOs	Standard deviation
HbA1c check	83%	5,7%
Lipid profile	54%	4,7%
Eye exam	45%	2,0%

HMO	Average Performance
HMO1	58%
HMO2	65%
HMO3	60%
HMO4	60%

Figure 3 - Two example tables to measure test compliance

The upper table contains the mean test compliance rates in the entire community (e.g., a county) and its associated standard deviation. The lower table indicates the general performance of each HMO. Since each HMO considers its own compliance rates for each of these tests (e.g. the HbA1c check) as sensitive data, this information is not displayed. However, given the aggregate data published by the mediator in both tables, bounds can be inferred about the sensitive values. For example, HMO_1 can use its knowledge of its own compliance rates and the published data to infer that HMO_2's compliance rate for the HbA1c check is between 87.2% and 88.5% which corresponds to an inferred interval of [0.872; 0.885].

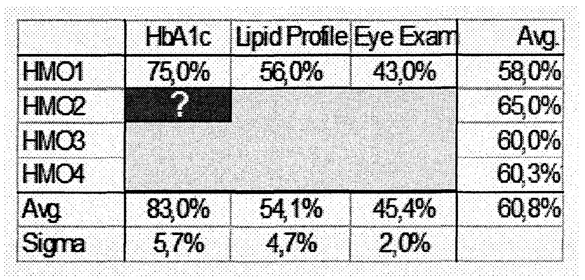
Mediators should detect and limit this type of privacy breach. Our objective is to develop new models and methods for the prevention of interval inference that can be incorporated into the mediator.

Returning to our initial example of diabetes care, the healthcare initiative is interested in determining the nature of its data publication strategy. Consider an example where the mediator has to choose the kind of information it should publish/make available about compliance rates for the different preventive tests shown in Figure 3. It could publish the mean rates and perhaps a measure of dispersion (e.g., standard deviation). Furthermore, to let healthcare consumers make informed decisions about which HMO provides the best diabetes care, a measure that orders the HMO's overall average test compliance rate could also be published. What are the implications of publishing this data from a privacy standpoint?

In the first place, the generation of these tables requires a comprehensive query over HMO, patient, and laboratory databases. Although HMOs may agree to publish an aggregate performance measure, they may consider rates for specific single tests as internal data and may be concerned that they could be used for marketing campaigns by their competitors. As shown for eye exams in [1], HMOs often do not provide this data. Although these individual test data are not contained in the tables of Figure 3, we will now show that an HMO can compute tight bounds for this confidential data of its competitors based on the data published by the mediator.

In the following, we will assume that HMO_1 is the health plan that wants to acquire detailed information about specific test compliance rates of its competitors. First, HMO_1 can sum up all the data it has, as depicted in Table 1.

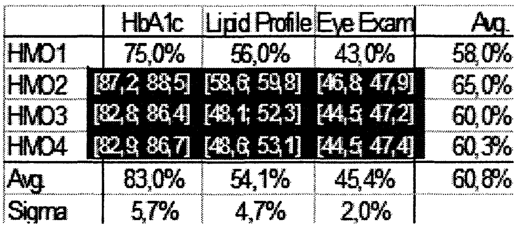
Table 1: Information known to HMO_1



	HbA1c	Lipid Profile	Eye Exam	Avg.
HMO1	75,0%	56,0%	43,0%	58,0%
HMO2	?			65,0%
HMO3				60,0%
HMO4				60,3%
Avg.	83,0%	54,1%	45,4%	60,8%
Sigma	5,7%	4,7%	2,0%	

In a second step, it can solve a Non-Linear Programming (NLP) problem for each unknown cell. The minimization NLP determines the lower bound, and the maximization NLP determines the upper bound for the cell in question. Thus, solving the NLP for each cell gives HMO_1 interval bounds for all sensitive cells. Table 2 shows that they are surprisingly narrow.

Table 2: Intervals inferred by the snooping HMO_1



	HbA1c	Lipid Profile	Eye Exam	Avg.
HMO1	75,0%	56,0%	43,0%	58,0%
HMO2	[87,2; 88,5]	[58,6; 59,8]	[46,8; 47,9]	65,0%
HMO3	[82,8; 86,4]	[48,1; 52,3]	[44,5; 47,2]	60,0%
HMO4	[82,9; 86,7]	[48,6; 53,1]	[44,5; 47,4]	60,3%
Avg.	83,0%	54,1%	45,4%	60,8%
Sigma	5,7%	4,7%	2,0%	

Given these intervals that can be inferred, every HMO has to decide whether the proposed publication is acceptable or not. Usually, HMOs specify the maximum disclosure risk for each of their sensitive cells *before* their data is given to the mediator. These criteria for disclosure risk can be simple measures such as minimum interval widths or more complex measures such as minimum information entropy, which we will discuss in the next section. If for at least one cell the risk criteria set by the HMOs are not met, a disclosure is detected and the data publication that was proposed by the mediator cannot take place.

The simulation of HMO_1's investigative behaviour by the mediator to imitate a snooper is called *disclosure detection* or *disclosure audit*. This principle has been applied, for example, by the U.S. Bureau of Census to ensure that after the publication of statistical tables, no conclusions can be inferred about individuals [1992]. Recently, Li et al. [11,12] proposed an integer programming approach for disclosure detection and the addition of random noise for disclosure limitation in sum queries. The mediator can limit disclosure by systematic aggregation and, in most of the cases, still publish useful data for the legitimate data user [13]. By making the marginal information more and more fuzzy until, in the worst case, it is suppressed, we can generate "audit and aggregrate" policies that satisfy all the stakeholders.

Trading off Privacy and Data Utility

Aggregating data reduces disclosure risk but at the same time, reduces data utility for the legitimate user. The tradeoff is captured well in the R-U map (Risk vs. Utility) introduced by Duncan et al. [14, 15]. Applying it to our context, Table 3 displays the three general dimensions of marginal information about table rows and columns. A measure of central tendency is most precisely described with the arithmetic mean. Dispersion can best be captured via the standard deviation, and the best order relation among the rows and columns is a total order. All these measures indicate high data utility (3), whereas their suppression would mean lowest data utility (1).

Table 3: Aggregating marginal information

Measure Utility	Central tendency (MC)	Dispersion (MD)	Order (MO)
High [3] ⇓ aggregate	Arithmetic mean $\mu_i = \frac{1}{n}\sum_j a_{ij}$	Standard deviation $\sigma_i = (\frac{1}{n}\sum_j (a_{ij} - \mu_i)^2)^{\frac{1}{2}}$	Total order of means $\mu_{i1} \le \mu_{i2}$ or $\mu_{i2} \le \mu_{i1}$ $\forall$ $i_1 \neq i_2$
Medium [2] ⇓ aggregate	Rounded mean (interval of 0.05 width) $[(\mu_i \, \mathrm{DIV} \, 0.05)*0.05; (\mu_i \, \mathrm{DIV} \, 0.05+1)*0.05]$	Spread $S_i = max_j\{a_{ij}\} - min_j\{a_{ij}\}$	Partial order of means (e.g against median) For a given $C \in R$ $\mu_i \le C$ or $C \le \mu_i$ $\forall i$
Low [1]	Suppress - % -	Suppress - % -	Suppress - % -

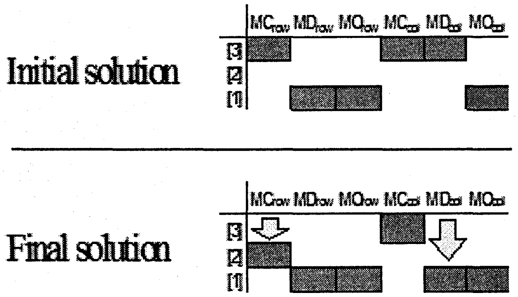

Initial solution
MC$_{row}$ MD$_{row}$ MO$_{row}$ MC$_{col}$ MD$_{col}$ MO$_{col}$
[3]
[2]
[1]
Final solution
MC$_{row}$ MD$_{row}$ MO$_{row}$ MC$_{col}$ MD$_{col}$ MO$_{col}$
[3]
[2]
[1]

Figure 4 - Example of reduction of data utility

The marginal information published in Table 1 would be of data utility (3, 1, 1) for the rows (only means are published) and (3, 3, 1) for the columns (means and sigmas are published). If the intervals inferred in Table 2 are too tight and thus violate one of the HMO's privacy policy, the data utility has to be reduced in a useful manner (disclosure limitation). "Useful" in this context means limiting the disclosure with least data utility loss possible. Figure 4 shows an example. The gray boxes represent the data utility assigned to a specific metric. In Table 1, the 1st (row average), the 4th (column mean) and the 5th (column sigma) metrics are published with highest data utility while the rest are suppressed (data utility 1). Increasing privacy protection now means "pushing the gray boxes down" as displayed in the final solution. We have developed an algorithm that minimizes this loss of data utility and is explained in detail in [13].

More elaborate versions of the algorithm take into account the preferences of the healthcare initiative and assign extra weight to especially useful metrics, thus allowing explicit analysis of tradeoffs between privacy and data utility.

Conclusions

For mediator-based health report creation, we raised privacy concerns at two different stages during the creation process and we presented technical solutions to address them. Clearly HMOs are not the only target for snoopers. All the stakeholders participating in the healthcare initiatives have an interest in "sanitizing" the outcomes of such extensive data collection and analysis efforts. A well-balanced mediator will have to satisfy the requirements of all the parties. Future work will address these issues and develop strategies to implement and test the systems and solution methods on real-world data.

References

[1] PHC4 2002: Pittsburgh Healthcare Cost Containment Council (2002). Diabetes Hospitalization Report, 2001 Data, Pittsburgh, November 2002. [http://www.phc4.org/adobe/Diab01.pdf]

[2] Berndt D, Fisher J, Hevner A, Studenicki J. Healthcare Data and Quality Assurance. IEEE Computer 2001: 34 (12): 56-65.

[3] Kossmann D. The state of the art in distributed query processing. *ACM Computing Surveys* 2000.

[4] R. Krishnan, X. Li, D. Steier, L. Zhao (2001). On Heterogeneous Database Retrieval: A Cognitively Guided Approach. In *Information Systems Research*, Volume: 12. September 2001, Number: 3. Pgs: 0286-0301

[5] G. Wiederhold (1993). Intelligent integration of information. In *Proceedings of the ACM SIGMOD conference*, Washington, DC.

[6] G. Wiederhold, M. Bilello, V. Sarathy, X. Qian (1996). A security mediator for health care information. In *Proceedings of the 1996 AMIA (formerly SCAMC) Conference.*

[7] C. Boyens, O. Günther (2003). Using Online Services in Untrusted Environments - A Privacy-Preserving Architecture. In *Proc. 11th European Conference on Information Systems*, Naples, Italy.

[8] R. Agrawal, R. Srikant (2000). Privacy-preserving Data Mining. In *Proceedings of the ACM SIGMOD conference on the Management of Data*, 2000.

[9] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke (2002). Privacy Preserving Mining of Association Rules. In *Proc. of 8th Intl. Conf. on Knowledge Discovery and Data Mining* (KDD), July 2002.

[10]L. V. Zayatz (1992). Linear programming methodology used for disclosure avoidance purposes at the census bureau. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*

[11]Y. Li, L. Wang, and S. Jajodia (2002a). Preventing Interval-based Inference by Random Perturbation. In *Proceedings of the Workshop on Privacy Enhancing Technologies (PET 2002)*, San Francisco.

[12]Y. Li, L. Wang, X. Wang, and S. Jajodia (2002b). Auditing interval-based inference. In *Proceedings of the 14th Conference on Advanced Information Systems Engineering (CAiSE'02)*, Toronto, Canada, May 27-31 2002.

[13]C. Boyens, R. Krishnan, R. Padman (2004). On Privacy-Preserving Access to Distributed Heterogeneous Healthcare Information. In *Proc. 37th Haiwai'i International Conference on System Sciences*, to appear.

[14]G. Duncan, S. Keller-McNulty (2001). Disclosure risk vs. data utility: The R-U confidentiality map. *Technical Report. Statistical Sciences Group. Los Alamos National Laboratory.*

[15]George T. Duncan, Stephen E. Fienberg, Ramayya Krishnan, Rema Padman, and Stephen F. Roehrig, "Disclosure Limitation Methods and Information Loss for Tabular Data," in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by Pat Doyle, Julia I. Lane, J.M. Theeuwes and Laura V. Zayatz, North-Holland, 2001.

Address for correspondence

Rema Padman, PhD.
Associate Professor
The Heinz School of Public Policy & Management
Carnegie-Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
rpadman@andrew.cmu.edu