Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

SHARE: system design and case studies for statistical health information release

James Gardner,[1] Li Xiong,[2,4] Yonghui Xiao,[2] Jingjing Gao,[3] Andrew R Post,[3,4] Xiaoqian Jiang,[5] Lucila Ohno-Machado[5]

[1]Digital Reasoning Systems Inc, Franklin, Tennessee, USA
[2]Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, USA
[3]Center for Comprehensive Informatics, Emory University, Atlanta, Georgia, USA
[4]Department of Biomedical Informatics, Emory University, Atlanta, California, USA
[5]Division of Biomedical Informatics, University of California San Diego, San Diego, California, USA

Correspondence to
Dr Li Xiong, Department of Mathematics and Computer Science, Emory University, 400 Dowman DR, Atlanta, GA 30322, USA, lxiong@emory.edu

Received 14 April 2012
Accepted 11 September 2012

ABSTRACT

Objectives We present SHARE, a new system for statistical health information release with differential privacy. We present two case studies that evaluate the software on real medical datasets and demonstrate the feasibility and utility of applying the differential privacy framework on biomedical data.

Materials and Methods SHARE releases statistical information in electronic health records with differential privacy, a strong privacy framework for statistical data release. It includes a number of state-of-the-art methods for releasing multidimensional histograms and longitudinal patterns. We performed a variety of experiments on two real datasets, the surveillance, epidemiology and end results (SEER) breast cancer dataset and the Emory electronic medical record (EeMR) dataset, to demonstrate the feasibility and utility of SHARE.

Results Experimental results indicate that SHARE can deal with heterogeneous data present in medical data, and that the released statistics are useful. The Kullback–Leibler divergence between the released multidimensional histograms and the original data distribution is below 0.5 and 0.01 for seven-dimensional and three-dimensional data cubes generated from the SEER dataset, respectively. The relative error for longitudinal pattern queries on the EeMR dataset varies between 0 and 0.3. While the results are promising, they also suggest that challenges remain in applying statistical data release using the differential privacy framework for higher dimensional data.

Conclusions SHARE is one of the first systems to provide a mechanism for custodians to release differentially private aggregate statistics for a variety of use cases in the medical domain. This proof-of-concept system is intended to be applied to large-scale medical data warehouses.

OBJECTIVES

Recent studies and advisory reports to the government[1–3] have pointed out that information sharing with appropriate privacy protection is one of the most critical challenges of our time, which has the potential to help revolutionize healthcare. In particular, the Institute of Medicine's committee on health research and the privacy of health information concludes[3] that the current Health Insurance Portability and Accountability Act (1996) (HIPAA) privacy rule (http://www.hhs.gov/ocr/privacy/) does not protect privacy well and calls for an entirely new approach to protecting privacy in health research.

We present and describe a new software framework, statistical health information release (SHARE), for releasing statistical health information with differential privacy, a strong privacy framework for statistical data release. Through studies with real medical datasets, we get insight into the feasibility and utility of applying differentially private statistical data release to medical data.

BACKGROUND AND SIGNIFICANCE

The problem of preserving patient privacy in disseminated biomedical datasets has attracted increasing attention by both the biomedical informatics and computer science communities.[3–7] The goal is to share a 'sanitized' version of the individual records (microdata) that simultaneously provides utility for data users and privacy protection for the individuals represented in the records. In the biomedical domain, many text de-identification tools are focused on extracting identifiers from different types of medical documents and use simple identifier removal or replacements according to the HIPAA safe harbor method for de-identification.[7–10] Several studies and reviews have evaluated the re-identification risks of linking de-identified data by the HIPAA safe harbor method with external data such as voter registration lists.[11–14] Many studies have proposed or applied formal anonymization methods on medical data.[15–22] While still the dominant approach in practice, the main limitation of microdata release with de-identification is that it often relies on assumptions of certain background or external knowledge (eg, availability of voter registration lists) and only protects against specific attacks (eg, linking or re-identification attacks).

A complementary research problem to microdata (ie, original data) release is to release only privacy-preserving statistical macrodata (ie, derived statistics), which could also be used to construct synthetic data. Differential privacy[23–25] has emerged as one of the strongest unconditional privacy guarantees for statistical data release. It makes few assumptions on the background or external knowledge of an attacker, and thus provides a strong provable privacy guarantee. A statistical aggregation or computation satisfies ε-differential privacy, ie, is ε-differentially private, if the outcomes are formally 'indistinguishable' ('indistinguishable' is formally and quantitatively defined in Dwork)[23] (outcome probability differs by no more than a multiplicative factor e^{ε}) when run with and without any particular record in the dataset, where ε is a privacy parameter that limits the maximum amount of influence a record can have on the outcome. A common mechanism to achieve ε-differential privacy is the Laplace mechanism, which adds calibrated noise to a statistical measure, as determined by a given privacy parameter ε and the sensitivity of the statistical measure to the inclusion and exclusion of any record in the dataset. A more stringent privacy parameter requires more

Copyright 2012 by American Medical Informatics Association.

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

noise to be added and thus provides a higher level of privacy. A data custodian can specify an overall privacy parameter (ie, privacy budget) (the 'privacy budget' intuitively refers to an expendable resource that can be utilized to get statistical information from a dataset given a privacy requirement. A lower budget typically requires more noise to be added to each statistical measure or allows fewer measures to be computed) that can be used for a sequence of statistical measures, ie, each computation utilizes a portion of the budget, and the overall result guarantees differential privacy according to the composition properties of differential privacy.[26 27] While interactive mechanisms for specialized studies exist,[28] it remains a hard problem to find efficient and effective algorithms for non-interactive data release (ie, to find an optimal set of statistical measures) that ensures differential privacy given a privacy budget while guaranteeing and maximizing data utility[29] for targeted applications of the data. Applying differential privacy to health data presents practical challenges in addition to technical challenges,[30] but the quantification of privacy risk that it entails more than justifies research in this area.

SHARE is a prototype we have developed for releasing statistical health information with differential privacy guarantees. The released data allow researchers to deduce important medical findings without compromising the privacy of individuals. The usage of formal privacy techniques gives formal guarantees of privacy, which are typically lacking in honest brokers and data releasers' data toolboxes. We present SHARE's design and report on its application in guaranteeing privacy of shared, real-world clinical data.

MATERIALS AND METHODS

Overview

SHARE takes as input structured biomedical data (eg, coded diagnoses, demographic data), a privacy budget, and outputs aggregated statistics (eg, means, histograms) with differential privacy guarantees. It implements several state-of-the-art algorithms that are designed for different types of data. The basic component, DPCube, releases aggregated count statistics in the form of multidimensional histograms (data cubes). For longitudinal data, it contains a specialized component, DPTrie, for accurately releasing count statistics of longitudinal patterns in the form of a prefix tree (trie). The released statistics can serve as a sanitized synopsis of the original database. It can also be used to generate a synthetic dataset that mimics the original data. Together, they support a variety of online analytical processing queries and other learning tasks. We present two use case studies evaluating DPCube and DPTrie on two real-world biomedical datasets.

SHARE is also integrated with the health information de-identification (HIDE)[31–34] system we have developed previously for releasing both differentially private statistical data and de-identified records for unstructured (eg, narrative text) and structured data. Figure 1 shows an overall conceptual view of the SHARE system integrated with the HIDE system. The integrated system offers an end-to-end solution. A data custodian for a medical institution typically has access to structured and unstructured components of electronic health records (EHR). For unstructured records, our previous system HIDE uses a statistical learning approach, the state-of-the-art conditional random field framework,[35 36] as the basis for tagging protected health information (PHI) and other useful elements. A patient-centric view of the data is created by linking all of the relevant variables for the same patient from the structured and unstructured data. The patient-centric view may be used to release: the original text with anonymized substitutions in place of the original PHI and anonymized data tables containing individual records generated by the de-identification component using HIPAA safe harbor methods or more advanced statistical anonymization methods, and differentially private aggregated statistics through the SHARE system. Please refer to Gardner and Xiong,[33] Gardner *et al*[34] and Jurczyk *et al*[37] for details of the PHI extraction and linking components of HIDE. In this article, we focus on SHARE functionalities using structured data inputs. Below we describe the DPCube and DPTrie components in detail.

Differentially private histogram release

The DPCube component builds a differentially private multidimensional histogram (data cube) for an input dataset given a privacy budget. It implements the multidimensional partitioning algorithm we have previously designed.[38 39] The DPCube algorithm consists of three steps. In step 1, the algorithm generates a differentially private equiwidth cell (unit) histogram. In step 2, it partitions the data space using the cell histogram from the first step and generates partitions (optimized histogram buckets). In step 3, the partitions are used to generate a differentially private subcube histogram. The privacy budget is allocated to steps 1 and 3 for noise perturbation to ensure differential privacy of the resulting histograms. The main goal of the algorithm is to generate a v-optimal histogram,[40] which minimizes the cumulative weighted variance of the histogram buckets and thus improves query precision or utility of the disclosed histogram. A v-optimal histogram can be approximated by a variety of heuristics. The initial implementations of DPCube used greedy partitioning of either the median value of attribute or information gain-based split points. Given a user-issued query, an estimation component can answer the query using the subcube histogram or apply inference or estimation techniques to boost the accuracy further using both histograms.[39 41] The histograms can serve as a sanitized synopsis of the original database and, together with an optional synthesized dataset based on the histograms, are useful to support count queries and other types of online analytical processing queries and learning tasks. Figure 2A shows an example illustrating the process of DPCube for releasing two-dimensional histograms built on age and income attributes from the original data.

For data that contain set-valued attributes (eg, a list of symptoms extracted from a pathology report), SHARE includes an implementation of the top-down partitioning algorithm[42] that is designed to cope with the high-dimensional set-valued data.

Differentially private longitudinal pattern release

Many biomedical datasets are collected for longitudinal studies that involve repeated observations of the same variables over periods of time. Due to high dimensionality and self-correlation of the data, DPCube and other existing histogram methods are not well suited for this type of data. SHARE includes an adaptation of the prefix tree-based algorithm[43] to maintain the longitudinal patterns of the data accurately. We named this component DPTrie and briefly describe it with an example below.

Figure 2B shows an example longitudinal dataset of blood pressure history for several patients and the prefix tree generated from their original records. A prefix tree groups temporal patterns with the same prefix into the same branch in the tree. Each level of the tree corresponds to a time point in the longitudinal data. The key value pair in each tree node represents a prefix pattern, and the number of patients in the dataset

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

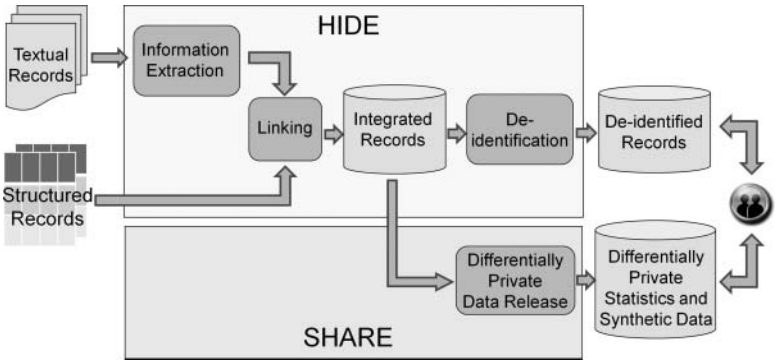


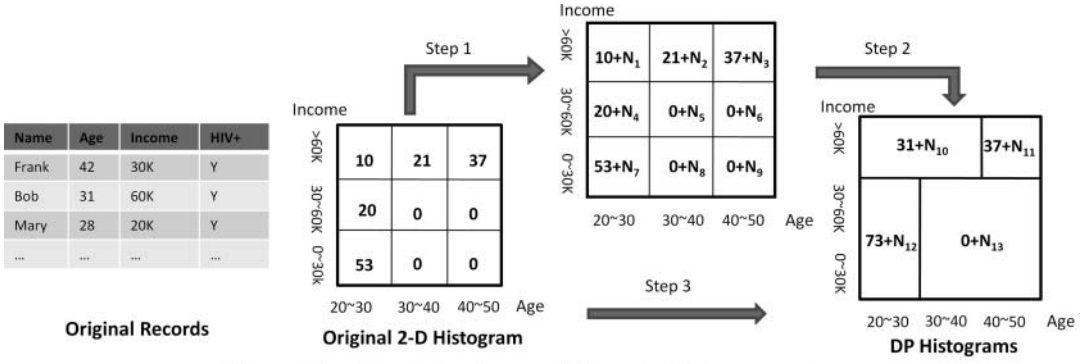
Figure 1 System overview of statistical health information release (SHARE): it is integrated with health information de-identification (HIDE) to provide both de-identification and differentially private statistical data release for unstructured and structured records. This figure is only reproduced in colour in the online version.

corresponding to that pattern. For example, at the first level of the tree, 'L:0' means that there are 0 patients with low blood pressure (90 or lower, denoted by L) at time t_1. Similarly, there are 70 normal (above 90 and below 130, denoted by N) and 30 with high blood pressure (130 or higher, denoted by H). Among the 70 patients with N at t_1, 40 of them had N at t_2 (corresponding to pattern NN) and 30 of them had H at t_2 (corresponding to pattern NH). At each level, if a pattern is associated with no or a low number of patients, it does not needed to be expanded. For example, if no patient has L pattern at t_1, the node L is not expanded. The counts at each node are perturbed using the standard Laplace mechanism[16] before release to guarantee differential privacy.

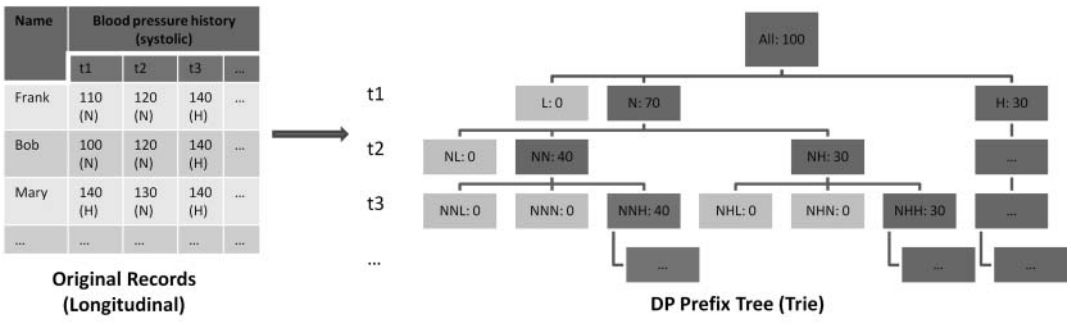
Case studies

We evaluated the feasibility and utility of the SHARE components using two real world datasets. The surveillance, epidemiology and end results (SEER) dataset[44] contains cancer statistics representing approximately 28% of the US population. The SEER research data include incidence, demographics (age, sex, race), year of diagnosis, and geographical area. It is common to use aggregated population statistics to determine mortality rates over ranges of time or to analyze demographic patterns. We demonstrate the feasibility and utility of differentially private data cubes generated by the DPCube component for such studies.

The Emory electronic medical record (EeMR) prescription dataset contains all e-prescription (eRx) information written by physicians at Emory University and affiliated hospitals over a period of 2 years. It also contains demographic information about each physician including age, sex, and locations of residence over the period in which they were in residency in the hospital system. We evaluated both the DPCube component and the DPTrie component on this longitudinal dataset to assess whether the latter better preserved the longitudinal patterns.



(a) **DPCube: Differentially Private Multidimensional Histogram Release**



(b) **DPTrie: Differentially Private Longitudinal Pattern Release (L: low, N: Normal, H: High)**

Figure 2 Overview and examples of differentially private histogram release (DPCube) and longitudinal pattern release (DPTrie) in statistical health information release (SHARE). This figure is only reproduced in colour in the online version.

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

RESULTS

We performed a variety of experiments addressing heterogeneous queries on the two datasets. We present empirical results and illustrate them with figures followed by a discussion.

SEER statistics

For the SEER[44] breast cancer dataset, the goal was to release data cubes that are close to the original data distribution while also giving a guaranteed level of privacy. The following seven dimensions (and cardinality) were selected to generate full data cubes representing high dimensions: sex (two), age (130), diagnosis year (36), behavior code (two), laboratory confirmation (nine), death code (two), other death code (two). The following three dimensions were selected to generate reduced data cubes representing low dimensions: age, diagnosis year, and behavior code. After filtering out patients with unknown data, the dataset contained 22 174 breast cancer patient records between 1973 and 2008. In all of our DPCube experiments, we allocated the privacy budget equally between the released cell histogram and subcube histogram and the overall released data cubes are differentially private. The year of diagnosis, age at diagnosis, and the death status were sliced from the full data cubes as bases for analyses (all figures in this section use blue to indicate other cause of death, and green to indicate death as a result of cancer).

We also compared the DPCube algorithm with a baseline algorithm that simply generates a noisy cell (unit) histogram by adding noise to the count of every unit according to the total privacy budget. We note that there are several state-of-the-art algorithms[45 46] for generating multidimensional histograms. They have been compared in our earlier work[38 47] and others,[46] which showed that DPCube results in comparable accuracy relative to the alternative methods given a privacy requirement and can even show an advantage for certain data distributions and queries. Our goal in this article is to demonstrate the feasibility and utility of releasing differentially private data cubes for real-world medical data, rather than to repeat detailed comparisons with other approaches.

Figure 3A, B shows the original histograms and differentially private histograms from the full data cubes generated by the baseline algorithm and the DPCube algorithm for a privacy budget of 0.5 with respect to the variables (year of diagnosis, cause of death) and (age of diagnosis, cause of death), respectively. The histograms show both individuals who died as a result of cancer as well as those who are either still living or died of other causes. DPCube produces distributions that are closer to the original distributions than those by the baseline algorithm. In general, we observed that for high-dimensional data cubes such as the full data cubes, a smaller privacy budget (ie, higher noise) causes the distributions to become closer to uniform and thus provides lower utility. When $\varepsilon=0.1$, the absolute values of differentially private counts become less meaningful due to the high noise; however, we observed that they still preserve the original distribution to some extent. It is a task for regulators and honest brokers to determine the acceptable level of utility (or error) for a certain guaranteed level of privacy.

We performed further error analysis for both the full data cubes and the reduced data cubes. Figure 4 shows the counts and absolute error of the number of deaths from breast cancer with respect to the year of diagnosis. For the full data cubes, the DPCube algorithm results in errors between 2227 and 3056, while the baseline gives errors between 6787 and 7447 for the yearly death counts of individuals diagnosed between years 1973 and 2008. These results confirm that the DPCube algorithm produces distributions closer to the original with the same level of privacy as the baseline approach. For the reduced data cubes, the DPCube algorithm provides a slight improvement over the baseline approach. We also calculated the Kullback–Leibler (KL) divergence between the released noisy data cubes and the original data cubes. The KL divergence is used as a standard non-symmetric measure of the difference between two probability distributions. Figure 4 shows that DPCube achieves lower (better) KL divergence for varying differential privacy budgets.

Longitudinal study using EeMR dataset

The second case study evaluated the longitudinal data support of SHARE using the EeMR prescription dataset. The results show that differentially private statistics can be released that support complex queries involving aggregations of demographic and longitudinal information from the data. The dataset contains the national provider ID and the average number of eRx per patient for each month for each physician. Many physicians started at different times, therefore the data were normalized so that each physician started at month 1. This preprocessing allows for the detection of trends for the counts of eRx writing for physicians in residence. Physicians with fewer than 9 months of residency were removed from the dataset. After filtering, the dataset consisted of 517 physician temporal sequences. The data were smoothed into 'quarters', for which we took the average over 3-month spans for each physician. We randomly augmented the data by sampling with replacement 10 000 entries in order to get a large enough dataset to apply the differential privacy principles described in this article. (These data were selected to represent as real-world data as possible, but the augmentation was necessary to evaluate the differential privacy on histograms, which requires large datasets. Assuming our dataset is representative of the population, similar datasets taken from a larger pool of clinicians should show similar results.) The data were normalized to indicate physicians who averaged zero, low (0–3), medium (3–6), or high ($\geq$6) medication counts per patient visit in each month.

Figure 5 shows the trends of a random selection of four physicians in the dataset. Most physicians tend to write more eRx per patient visit on average over time, but there are some trends downward exhibiting 'zig-zag' patterns. We performed experiments to check if demographic information could be used to classify physicians by trend on the original dataset, without success. This led us to believe that demographic information alone is not a good indicator of trend. Even though we were unable to classify or cluster trends with physician demographics, we were able to see clear trends in the longitudinal data. The goal was to provide differentially private release of the data that still preserve the ability to perform trend queries or aggregate analysis. The utility of these trends can be evaluated by measuring the error for temporal queries of varying length. An example temporal query of length 3 would be 'How many physicians averaged 2 eRx per visit the first month, 4 the second, and 6 the third?' One measure for determining the accuracy of a differentially private longitudinal data release is to measure the average error for temporal queries involving different lengths of time.

We first applied DPCube to the aggregations over these data. The error was significantly worse than that of the standard cell-based approach (baseline). By examining the data, the results confirmed our analysis that the DPCube approach is ill

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

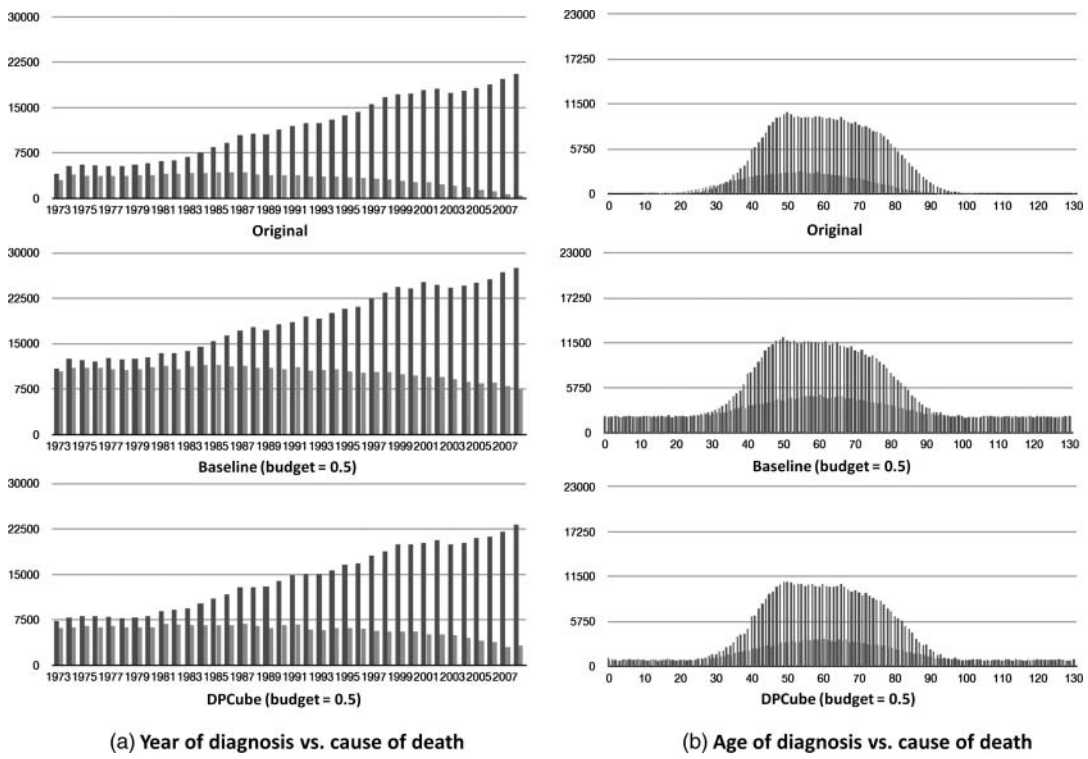


Figure 3 Histograms of death cause after cancer diagnosis relative to the year of diagnosis and age of diagnosis generated from full data cubes for the surveillance, epidemiology and end results (SEER) dataset. All figures use green to indicate death as a result of cancer and blue to indicate other causes of death. This figure is only reproduced in colour in the online version.

suited for datasets with extremely skewed local distributions. In addition, treating each time point as a single attribute creates a highly dimensional dataset with an exponential number of cells with respect to the number of time points, which DPCube is not designed for. We then applied the DPTrie component and present the results below.

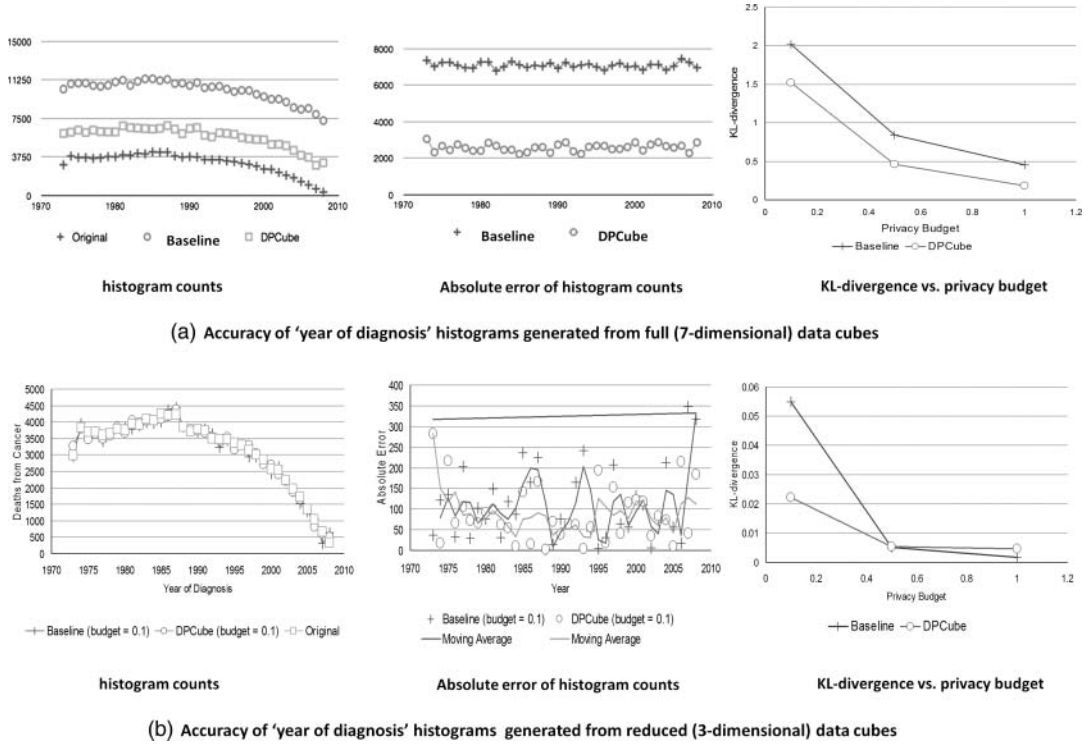


Figure 4 Comparison of DPCube and baseline for number of cancer deaths relative to the year of diagnosis generated from the full data cubes (seven-dimensional) and reduced data cubes (three-dimensional). KL, Kullback–Leibler. This figure is only reproduced in colour in the online version.

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

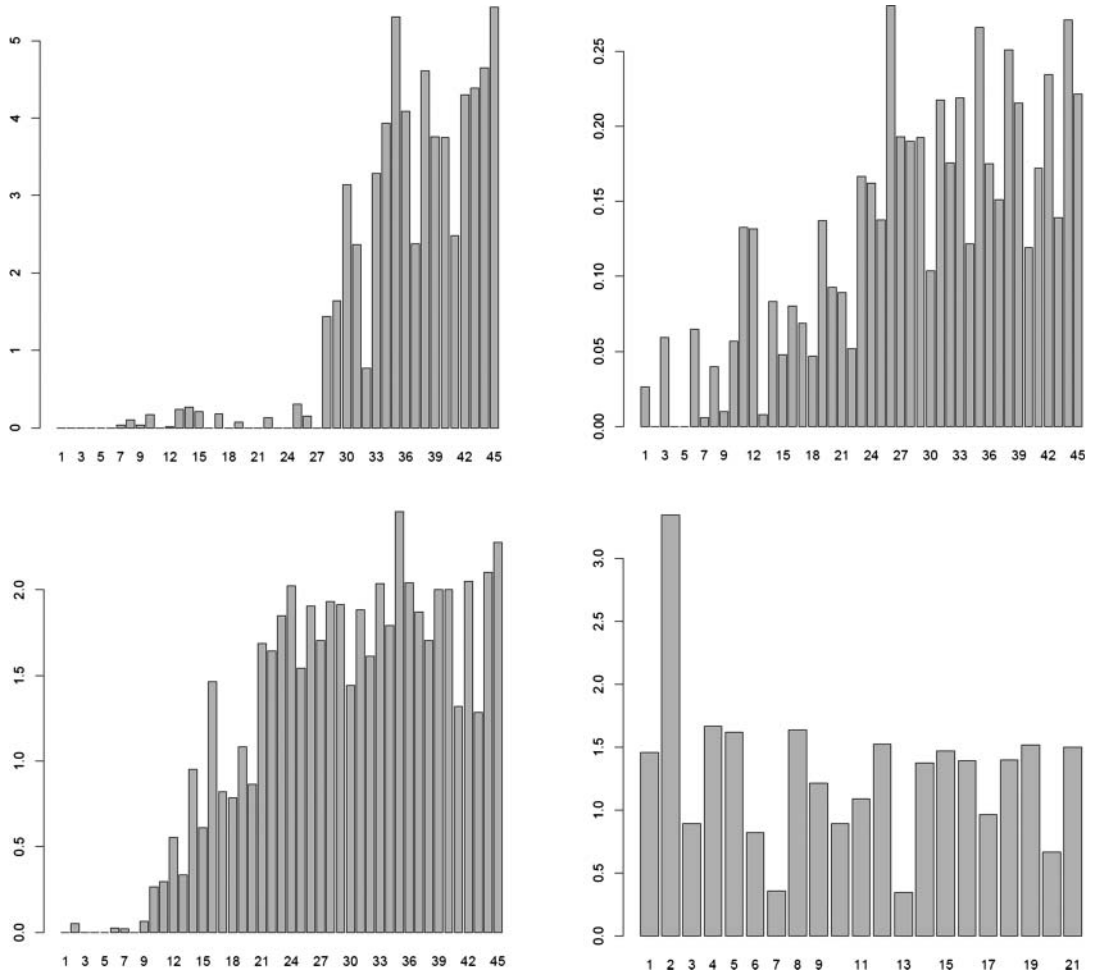


Figure 5 Random selection of physicians where X value is month of residence and Y value is the average number of e-prescriptions (eRx) per visit in each month.

Figure 6A shows the average perturbed counts returned for temporal queries of random patterns of a given length are comparable to the original counts. As expected, the average count decreases over query length as the query patterns become more distinguishing. To examine the error more closely, we generated temporal queries according to different slopes, corresponding to physicians who tend to write more or fewer eRx over a period of time (query length) during their residence. The slopes are defined to be negative (−1.5, −0.5), approximately zero or flat (−0.5, 0.5), or positive (0.5, 1.5). For example, we could find out the number of physicians who have particular patterns with an increasing eRx rate over their entire residence. Figure 6B, C shows the average absolute error and average relative error with respect to different slopes and the query length. We observed that the error decreases as query length increases because the number of physicians returned will be smaller. The error is quite small over long patterns. On the other hand, queries of positive slope have a higher error because of its higher count as the majority of the physicians exhibit a positive eRx rate over time. In general, the relative error varies between 0 and 30%. We believe that the slope statistics on this dataset are useful and can be released in a differentially private manner.

DISCUSSION

We have presented the system design of the SHARE system and various experiments on real-world and augmented datasets. We envision the released data statistics could be freely used for queries preparatory to research studies and prospective clinical trials to gauge whether or not there are enough data satisfying the needs of the researcher in the original dataset that could warrant seeking institutional review board approval and access rights to the original data. Population-level or larger-scale observational research studies could potentially be done on the privacy-preserving data release to determine trends or possible predictors for disease outcomes. Before publication or dissemination it would probably be necessary to perform the study on the original data, but a mechanism that is guaranteed to preserve privacy above a certain predetermined level would allow for potentially more studies without the need for formal approval or large pools of patients that give consent for some studies. Comparative effectiveness studies could also be possible following a similar workflow. Typically, comparative effectiveness studies require the use of a variety of data sources. These studies could have increased power if a number of institutions release differentially private data statistics. If analyses on these cubes suggest promising trends across diverse data sources, then it may be possible to reach out to other researchers who could verify the findings on their own data, for which they may have higher access rights. Comparing analyses on disclosed data cubes versus distributed analyses[48] is also an important area for further research.

There are also some questions or challenges that need to be addressed before differentially private data release can be

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

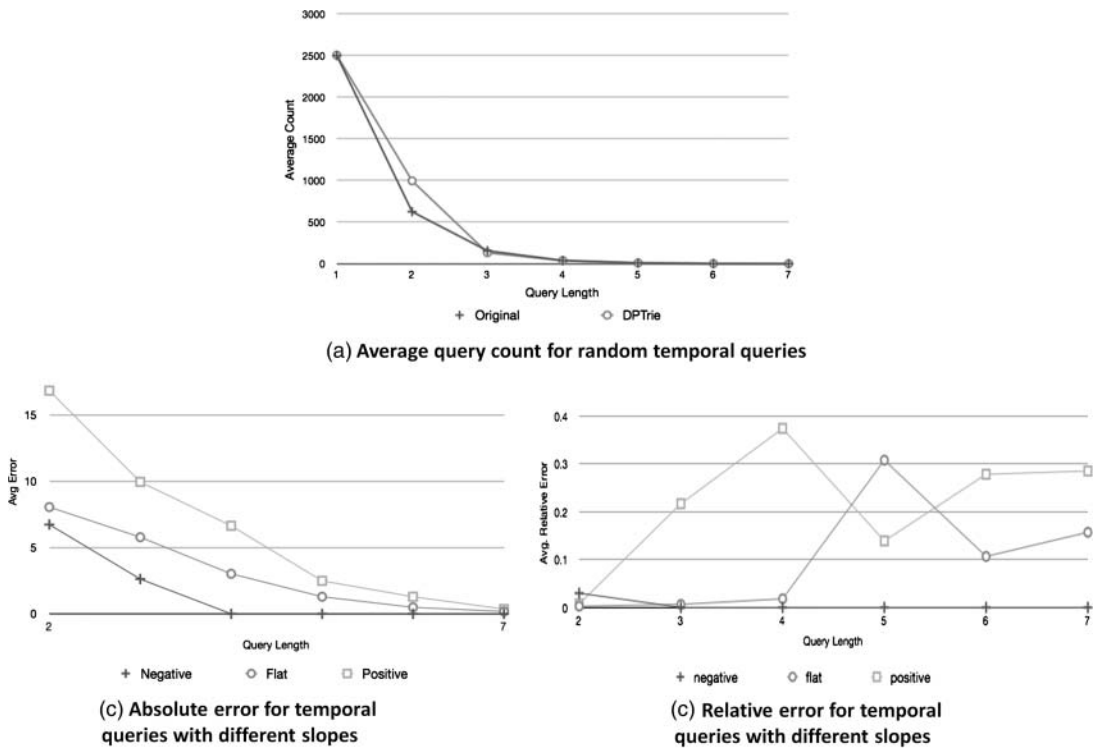


Figure 6 Average counts and query errors of longitudinal queries with respect to time length for the Emory electronic medical record (EeMR) dataset. This figure is only reproduced in colour in the online version.

applied on a large scale. In practice, a data custodian would specify an initial privacy budget based on how sensitive the information is and the desired level of privacy. It remains an open question how to configure and even explain the level of differential privacy in intuitive ways to the data custodians and patients. In addition, each of the statistical data releases would utilize a portion of this privacy budget due to the composition properties of the differential privacy. The budget bound inevitably still places limitations on the practical implementation of the system. Finally, as we have observed, many of the heuristic algorithms are data dependent. Even a subtle difference in the algorithmic parameters can have a significant effect on the resulting data. Domain knowledge about the data to be released (eg, whether they are high dimensional, whether they are longitudinal) and the targeted applications are important and should be used, when possible, to guide the selection and design of proper data release algorithms.

While we have shown that differential privacy can be successfully applied in two medical datasets, future research is warranted. The case studies presented here and most existing work have shown that differentially private histogram release methods work well for single or low-dimensional data. It remains a challenge to build high-dimensional histograms efficiently that are accurate due to the high dimensionality and data sparseness. For the prefix tree-based release of longitudinal data, one key question is how to allocate the privacy budget among all the different levels of the tree. We plan to investigate various allocation schemes to maintain more accurate longitudinal patterns. We are also interested in devising methods to incorporate domain knowledge in the release process.

CONCLUSION

We have presented the SHARE system for releasing statistical health information with differential privacy guarantees. Integrated with HIDE, it enables custodians to implement a variety of privacy-preserving medical data publishing options. It gives honest brokers the ability to share privacy-preserving aggregated statistics and longitudinal data. However, there are still several aspects that need to be explored further. Our ultimate goal is to create a framework and system that can be used in practice on a large scale.

Funding This material is based on work supported by the National Science Foundation under grant no. 1117763 and an Emory URC grant. XJ and LO-M were funded in part by NIH grants U54 HL108460, 1k99LM 011392–01, R01HS019913, and UL1RR031980.

Contributors The authors are ranked according to their contributions. JG implemented the prototype, extended the DPCube algorithm design, conducted the experiments, and drafted parts of the manuscript. LX led the prototype and algorithm design, and was responsible for the overall manuscript drafting and revision. YX designed the original DPCube algorithm. JG and ARP contributed to the Emory EeMR dataset construction and experiment design and reviewed the manuscript. XJ and LO-M contributed to the literature review, the SEER experiment design, and reviewed and revised the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The SEER data are publicly available. We plan to share the EeMR dataset in an aggregated form with differential privacy.

REFERENCES

1. **Advisory C for USPIT, PITAC, (PITAC)**. President's Information Technology Advisory Committee. Revolutionizing health care through information technology. National Coordination Office for Information Technology Research and Development, 2004.
2. **Stead WW,** Lin HS, eds. *Computational technology for effective health care: immediate steps and strategic directions*. Committee on Engaging the Computer Science Research Community in Health Care Informatics; National Research Council. Washington DC: The National Academies Press, 2009.
3. **Nass SJ,** Levit LA, Gostin LO. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. Washington DC: National Academy Press, 2009.
4. **Fung BCM,** Wang K, Chen R, *et al*. Privacy-preserving data publishing: a survey of recent developments. *ACM Computing Surveys* 2010;**42**:1–534.

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

5. **Malin B.** An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J Am Med Inform Assoc* 2005;**12**:28–34.
6. **Ohno-Machado L,** Bafna C, Boxwala A, *et al*. iDASH. Integrating data for analysis, anonymization, and sharing. *J Am Med Inf Assoc* 2011;**19**:196–201. PMID: 22081224 PMCID: PMC3277627.
7. **Meystre SM,** Friedlin FJ, South BR, *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med Res Methodol* 2010;**10**:70.
8. **Sweeney L.** Replacing personally-identifying information in medical records, the scrub system. *Proc AMIA Annu Fall Symp* 1996;333–7.
9. **Szarvas G,** Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc* 2007;**14**:574–80.
10. **Uzuner OO,** Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;**14**:550–63.
11. **El Emam K,** Jabbouri S, Sams S, *et al*. Evaluating common de-identification heuristics for personal health information. *J Med Int Res* 2006;**8**:e28.
12. **El Emam K,** Brown A, AbdelMalik P. Evaluating predictors of geographic area population size cut-offs to manage re-identification risk. *J Am Med Inform Assoc* 2009;**16**:256–66.
13. **Benitez K,** Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;**17**:169–77.
14. **El Emam K,** Jonker E, Arbuckle L, *et al*. A systematic review of re-identification attacks on health data. *PLoS One* 2011;**6**:e28071.
15. **Ohrn A,** Ohno-Machado L. Using Boolean reasoning to anonymize databases. *Artif Intell Med* 1999;**15**:235–54. PMID: 10206109.
16. **Sweeney L.** K-anonymity: a model for protecting privacy. *Int J Uncertainty, Fuzziness and Knowledge-based Syst* 2002;**10**:557–70.
17. **Ohno-Machado L,** Silveira PS, Vinterbo S. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. *Int J Med Inf* 2004;**73**:599–606. PMID: 15246040.
18. **El Emam K,** Dankar FK. Protecting privacy using k-anonymity. *J Am Med Inform Assoc* 2008;**15**:627–37.
19. **Mohammed N,** Fung BCMM, Hung PCKK, *et al*. Anonymizing healthcare data: a case study on the blood transfusion service. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '09*. New York, USA: ACM Press, 2009:1285.
20. **Malin B,** Benitez K, Masys D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA privacy rule. *J Am Med Inform Assoc* 2011;**18**:3–10.
21. **Emam KE,** Paton D, Dankar F, *et al*. De-identifying a public use microdata file from the Canadian national discharge abstract database. *BMC Med Inform Decis Making* 2011;**11**:53.
22. **El Emam K,** Arbuckle L, Koru G, *et al*. De-identification methods for open health data: the case of the heritage health prize claims dataset. *J Med Int Res* 2012;**14**:e33.
23. **Dwork C.** Differential privacy. *Int Colloquium on Automata, Languages and Programming* 2006;**4052**:1–12.
24. **Dwork C.** Differential privacy: a survey of results. In: Agrawal M, Du DZ, Duan Z, *et al*., eds. *TAMC, volume 4978 of Lecture Notes in Computer Science*. Xi'an, China: Springer, 2008:1–19.
25. **Dwork C.** A firm foundation for private data analysis. *Commun ACM* 2011;**54**.
26. **McSherry FD.** Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: *Proceedings of the 35th SIGMOD international conference on Management of data (SIGMOD '09)*; 19–30. Binnige C, Dageville B, eds. New York, NY, USA: ACM, 2009. doi:10.1145/1559845.1559850 http://doi.acm.org/10.1145/1559845.1559850
27. **McSherry F.** Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun ACM* 2010;53:89–97. http://doi.acm.org/10.1145/1810891.1810916
28. **Vinterbo S,** Sarwate A, Boxwala A. Protecting count queries in study design. *J Am Med Inform Assoc* Published Online First: 17 Apr 2012. doi:10.1136/amiajnl-2011–000459.
29. **Dwork C,** Naor M, Reingold O, *et al*. On the complexity of differentially private data release: efficient algorithms and hardness results. In: *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing—STOC '09*. New York, NY, USA: ACM Press, 2009:381.
30. **Dankar F,** Emam KE. The application of differential privacy to health data. In: EDBT/ICDT 2012 Joint Conference. Berlin, Germany, 2012:1–9.
31. **Gardner J,** Xiong L. HIDE: an integrated system for health information DE-identification. *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2008:254–9.
32. **Gardner J,** Xiong L, Li K, *et al*. HIDE: heterogeneous information DE-identification. In: *Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology—EDBT '09*. New York, NY, USA: ACM Press, 2009:1116–19.
33. **Gardner J,** Xiong L. An integrated framework for de-identifying unstructured medical data. *Data Knowledge Eng* 2009;**68**:1441–51.
34. **Gardner JJ,** Xiong L, Wang F, *et al*. An evaluation of feature sets and sampling techniques for de-identification of medical records. *The 1st ACM International Health Informatics Symposium (IHI)*. 2010:1–8.
35. **Lafferty JD,** McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 2001:282–9.
36. **Okazaki N.** CRFsuite: a fast implementation of Conditional Random Fields (CRFs). 2007. http://www.chokkan.org/software/crfsuite/ (accessed June 2008).
37. **Jurczyk P,** Lu JJ, Xiong L, *et al*. FRIL: a tool for comparative record linkage. *AMIA Annu Symp Proc*. 2008:440–4.
38. **Xiao Y,** Xiong L, Yuan C. Differentially private data release through multidimensional partitioning. *Secure Data Manag* 2011;**6358**:150–68.
39. **Xiao Y,** Gardner J, Xiong L. DPCube: Releasing Differentially Private Data Cubes for Health Information. In *28th IEEE International Conference on Data Engineering (ICDE)*, 2012.
40. **Poosala V,** Haas PJ, Ioannidis , *et al*. Improved histograms for selectivity estimation of range predicates. In: *Proceedings of the 1996 ACM SIGMOD international conference on Management of data (SIGMOD '96)*, ACM, 1996:294–305.
41. **Hay M,** Rastogi V, Miklau G, *et al*. Boosting the Accuracy of Differentially-Private Histograms Through Consistency. In: *Proceedings of the International Conference on Very Large Data Bases*; 2009:3:15.
42. **Chen R,** Mohammed N, Fung BCM, *et al*. Publishing set-valued data via differential privacy. *Proceedings of the VLDB Endowment* 2011;**4**:1087–98.
43. **Chen R,** Fung BCM, Desai BC, *et al*. Differentially private transit data publication: A case study on the montreal transportation system. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*; August 2012, ACM Press: Beijing, China.
44. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2009), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2011.
45. **Hayy M,** Rastogiz V, Miklauy G, *et al*. Boosting the accuracy of differentially-private histograms through consistency. The 36th International Conference on Very Large Data Bases, 13–17 September 2010, Singapore. In: Proceedings of the VLDB Endowment, Vol 3, No 1.
46. **Cormode G,** Procopiuc M, Shen E, *et al*. Differentially private spatial decompositions. IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April 2012, IEEE Computer Society 2012.
47. **Xiao Y,** Xiong Fan L, *et al*. DPCube: differentially private histogram release through multidimensional partitioning. Eprint arXiv: 1202.5358.
48. **Wu Y,** Jiang X, Kim J, *et al*. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012;**19**:758–64.

Downloaded from jamia.bmj.com on January 17, 2013 - Published by group.bmj.com

SHARE: system design and case studies for statistical health information release

James Gardner, Li Xiong, Yonghui Xiao, et al.

J Am Med Inform Assoc published online October 11, 2012
doi: 10.1136/amiajnl-2012-001032

Updated information and services can be found at:
http://jamia.bmj.com/content/early/2012/10/10/amiajnl-2012-001032.full.html

These include:

References This article cites 23 articles, 8 of which can be accessed free at:
http://jamia.bmj.com/content/early/2012/10/10/amiajnl-2012-001032.full.html#ref-list-1

P<P Published online October 11, 2012 in advance of the print journal.

Email alerting service Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

Advance online articles have been peer reviewed, accepted for publication, edited and typeset, but have not not yet appeared in the paper journal. Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/