Opportunities and Challenges for Privacy-Preserving Visualization of Electronic Health Record Data

Aritra Dasgupta, Eamonn Maguire, Alfie Abdul-Rahman and Min Chen

Abstract—In this paper, we reflect on the use of visualization techniques for analyzing electronic health record data with privacy concerns. Privacy-preserving data visualization is a relatively new area of research compared to the more established research areas of privacy-preserving data publishing and data mining. We describe the opportunities and challenges for privacy-preserving visualization of electronic health record data by analyzing the different disclosure risk types, and vulnerabilities associated with commonly used visualization techniques.

1 Introduction

Privacy-preserving data analysis techniques aim to prevent disclosure of sensitive personal information, while making the anonymized data usable for analysis. With electronic health record data, minimization of such disclosure risks is of high priority. In recent years, there has been a lot of work focusing on visualization of health record data, both at the individual record level [19, 12], and also at an aggregate level to look at cohort analysis and identifying temporal trends of treatments and patient plans [18, 3]. In both these cases, data privacy is at risk because of the complex ecosystem of the health-care industry, involving both trusted and untrusted users [9].

Hospitals and care-providers are typically the data owners in this ecosystem. Other stakeholders include analysts, insurance companies, clinicians, pharmaceutical companies, etc. They can become data custodians and use the data for their own analysis. In many cases some of the data are also publicly available, leading to availability of the data to users who are external to the ecosystem. There are threats to the protection of sensitive information at each of these levels (Figure 1).

So far, the issue of privacy-preservation of health care data has mostly been researched from a data publishing or data mining point of view. These scenarios typically produce an anonymized static data table or a modified mining algorithm. They aim to prevent leakage of sensitive information through a joining of public and private databases through quasi-identifiers such as age, gender, race, address, etc., which co-occur in such databases [2]. However, privacy-preserving data publishing and mining mostly constitute the non-interactive setting of privacy-preservation, where, once released, the data owner does not have any control over the data or the mining results. The advantages of an interactive query interface for privacy-preservation has been mentioned in the data mining literature [10]. Similarly, an interactive visualization alternative to the privacy-preserving paradigm may work better because: i) potential higher utility because of interaction and ii) more flexibility to the data owners and custodians in tuning visual parameters for protecting against an attacker's background knowledge.

However, visual representations and the associated interaction mechanisms also have their own challenges in protecting unintended disclosure. Contrary to the conventional visualization goal of maximizing user insight, a privacy-preserving visualization needs to constrain insight that can divulge sensitive information. In this paper, we reflect on the opportunities and challenges for visualization as a means to preserve sensitive insights from the data and only divulge non-sensitive patterns. To this end, we analyze the risk types and comment on the potential vulnerabilities in traditional visualization techniques in terms of privacy.

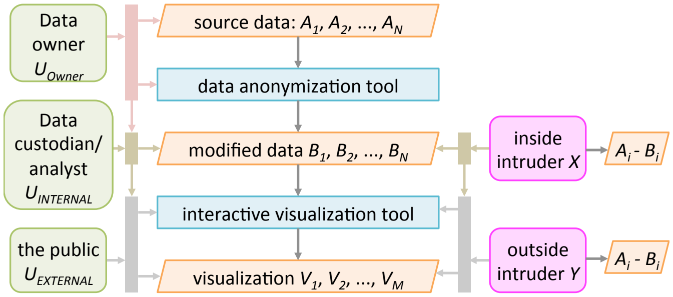

Data owner U_{Owner}
source data: $A_1, A_2, ..., A_N$
data anonymization tool
Data custodian/ analyst $U_{INTERNAL}$
modified data $B_1, B_2, ..., B_N$
inside intruder X
$A_i - B_i$
interactive visualization tool
the public $U_{EXTERNAL}$
visualization $V_1, V_2, ..., V_M$
outside intruder Y
$A_i - B_i$

Fig. 1. A schematic illustration of a privacy-preservation environment.

2 Data-based Privacy Risks and Preservation Approaches

In this section we give a brief overview of the research in the field of privacy-preserving data publishing and mining, and highlight their relevance to visualization.

Risk Types: The privacy risks mainly stem from the availability of external information about the persons in the database, and also the attacker's background knowledge. The two main risk types are due to *record linkage* and *attribute linkage*. Record linkage happens when a private database like a health record database can be joined with a publicly available data base like the census database. This happens through attributes like age, gender, zip code, etc., commonly known as quasi-identifiers. In the case of attribute linkage, through exploiting relationships between sensitive attributes and quasi-identifiers, attackers can know the value of an attribute, like the name of a disease of an individual.

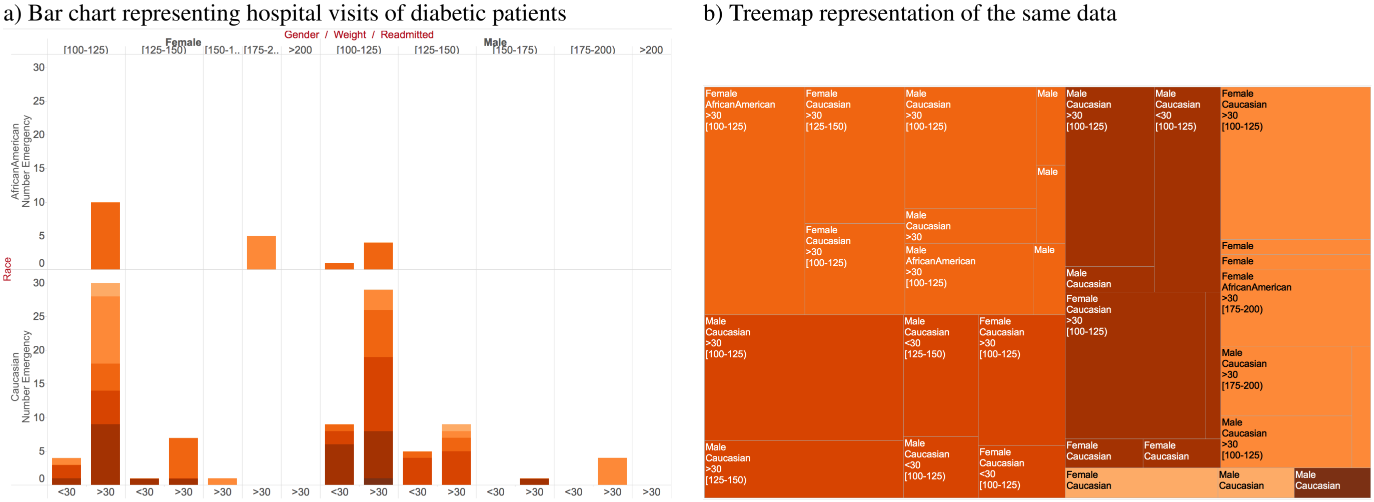
Anonymization Techniques for Input Privacy: Protection against attribute linkage and record linkage is achieved by broadly two privacy preservation approaches: *methods that modify data attribute values* and *methods that do not modify any attribute values.*

For the first case, there are two sub-categories: a) Generalization and suppression, where data values are either hidden or generalized based on some hierarchy or aggregation, and b) Perturbation, where the goal is to add noise to the data, so that the aggregate information is recoverable but the individual information cannot be recovered.

For the second case, different anatomization approaches have been proposed, where quasi-identifier data and sensitive attribute data are released in separate tables. A visualization counter-part of this approach would be coordinated multiple views for the different classes of attributes. In this paper, however, we focus on single views of the data.

- *A. Dasgupta is with New York University.
E-mail: adasgupt@nyu.edu*
- *Eamonn Maguire is with Oxford University.
E-mail: eamonn.maguire@st-annes.ox.ac.uk*
- *Alfie Abdul-Rahman is with Oxford University.
E-mail: alfie.abdulrahman@oerc.ox.ac.uk*
- *Min Chen is with Oxford University.
E-mail: min.chen@oerc.ox.ac.uk*

Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records. Copyright retained by the authors.



a) Bar chart representing hospital visits of diabetic patients

b) Treemap representation of the same data

Fig. 2. Aggregate data visualizations such as bar charts (color mapped to age groups of patients) and tree maps (color mapped to age groups of patients and area mapped to their number of emergency visits) might be necessary but not sufficient to fulfill the conditions for privacy-preservation. In a) the bar chart reveals patterns like: for African American women having a high frequency of *readmitted* (> 30) and *highest number of emergency visits*, the only age group is 50 to 60 represented by dark orange. In b) the same information is revealed, but it does not pop out as much as in the bar chart. With the presence of externally available information, involving quasi-identifiers age and race, such patterns can be construed as both sensitive, and having a high potential to reveal individual identity.

Issue of Output Privacy : Both attribute linkage and record linkage are related to the risk involving the input privacy, that is the privacy of the raw data that is an input to a mining algorithm. A relatively less researched area is how attackers can use the output produced by the mining algorithm to further mine sensitive patterns. Approaches like differential privacy [8] account for this scenario by protecting against malicious queries. However, as pointed out by Aris et al. ([10]), there needs to be more research on how sensitive knowledge can be protected.

These two scenarios of input privacy and output privacy, point to two factors that need to be accounted for while evaluating privacy risks in visualization: i) the cost of the operations that lead to record or attribute linkage and ii) the different tasks that can be done by interacting with a visualization, that can lead to unforeseen disclosure, similar to the mining scenario. These are discussed in the next section.

3 A Framework for Evaluating Privacy Risks

The research on privacy-preservation of electronic health data has been largely focused on data anonymization. Privacy risks are evaluated based on the level of anonymity and the computational performance of anonymization algorithms and tools. While such efforts are no doubt highly valuable, they are not exactly the same as what Shannon defined as "valuations of secrete systems". In particular, Shannon defined the *amount of secrecy* as follows.

"*There are some systems that are perfect – the enemy is no better off after intercepting any amount of material than before. Other systems, although giving him some information, do not yield a unique 'solution' to intercepted cryptograms. Among the uniquely solvable systems, there are wide variations in the amount of labor required to effect this solution and in the amount of material that must be intercepted to make the solution unique.*" [20]

Shannon's definition for secrecy can be extended to privacy. If we assume that an intruder may have some additional information beyond what can be extracted from the data available to him/her, it will be difficult to confirm any system is perfect for preserving privacy. Hence the evaluation of privacy risks should focus on "the amount of labor required to" break privacy. In a way, this bears some resemblance to evaluating encryption methods based on the computation resources required to break a piece of ciphertext.

However, when interactive visualization is law-abiding users and intruders, human factors will play a more significant role in a privacy-preservation environment than in a typical encryption environment. As Dasgupta *et al.* pointed out [5], visualization may introduce additional uncertainty in the screen-space, which may increase the privacy of the data, while reducing its utility. Meanwhile, interactive visualization may increase the efficiency and effectiveness of law-abiding users as well as intruders in exploring and integrating the data. Therefore, the evaluation of privacy risks is not only about computation resources but also human factors.

Figure 1 shows a schematic illustration of a privacy-preservation environment. For the simplicity and generality of our discussion, we consider the source data as a collection of N *data records*, $\mathbf{A} = \{A_1, A_2, \ldots, A_N\}$, each of which can be as simple as a univariate value or as complex as an arbitrary dataset. In order to preserve the privacy of each record, $\mathbf{A}$ is then anonymized, yielding the modified data as $\mathbf{B} = \{B_1, B_2, \ldots, B_N\}$.

Intuitively, a privacy-preserving visualization should be built on $\mathbf{B}$ rather than $\mathbf{A}$. However, it has been shown that a privacy-preserving visualization built on $\mathbf{A}$, proves to be more useful in some cases, than when built on the anonymized data $\mathbf{B}$. [7]. This is assuming that the visualization process is supported by an interactive visualization tool, which can produce many different visual representations of $\mathbf{B}$. Collectively these visual representations, $\mathbf{V} = \{V_1, V_2, \ldots, V_M\}$ enable users to perform their tasks, but may also be abused by an intruder to break the privacy of the data. Since theoretically, anyone, except superusers, should only have information about $\mathbf{B}$, gaining the information about any aspect of $A_i - B_i$ suggests an invasion of privacy.

Let us consider two abstract levels of privacy, namely accessing to (i) $\mathbf{B}$ as well as $\mathbf{V}$, and (ii) $\mathbf{V}$ only. Note that in practice, there can be many different levels of privacy. The two-level abstraction allows us to formulate a general framework for evaluating privacy risks. At Level (i), internal users, e.g., data analysts, can access the data records in $\mathbf{B}$ through textual interface (e.g., database queries, text files, spreadsheets) as well as visualization. An intruder who operates at this level is referred to as an *inside intruder*, and denoted as X. At Level (ii), visualizations are used for wider dissemination to external users, e.g., the public. An intruder who operates at this level is referred to as an *outside intruder*, and denoted as Y.

At each level, we can utilize Shannon's notion of "the amount of labor required" to define the following:

S_{int} — the amount of time required on average for an inside intruder X to obtain a piece of private data (i.e., a part of $A_i - B_i$) correctly among N data records.

T_{int} — the amount of time required on average for an internal user

$U_{INTERNAL}$ to perform a specific task or a set of specific tasks satisfactorily.

S_{ext} — the amount of time required on average for an outside intruder X to obtain a piece of private data (i.e., a part of $A_i - B_i$) correctly among N data records.

T_{ext} — the amount of time required on average for an external user $U_{EXTERNAL}$ to perform a specific task or a set of specific tasks satisfactorily.

Similar to the evaluation of any data visualization techniques, T_{int} and T_{ext} are task-dependent. However, this does not in any way make such evaluation unfeasible or unusable. One can always define one or more typical tasks as the reference tasks. One can reasonably conclude that it is possible to obtain S_{int}, T_{int}, S_{ext} and T_{ext} quantitatively through controlled user studies, or to estimate these qualitatively based on survey questions in Likert-type scales.

Hence S_{int} and S_{ext} are two absolute measures of "the amount of labor required" for breaking the privacy, while S_{int}/T_{int} and S_{ext}/T_{ext} are two relative measures that taking into account of the performance of the expected tasks. For example, given two different types of visualization techniques, one can measure or estimate these four quantities for each of the two techniques, conducting a comparison based the two sets of measures. In the context of electronic health record data, two general questions are particularly interesting. (1) How much can an inside intruder benefit from interactive visualization in comparison with dealing solely with textual data? (2) How much can an outside intruder benefit from interaction facilities (e.g., brushing) in comparison with viewing pre-rendered visualization images only?

Motivating example: For example, consider the visualization of hospital records of patients as shown in Figure 2. The different attributes that are represented are age, gender, race, weight, number of emergency visits, and number of times re-admitted. Since age is one of the quasi-identifier attributes (attackers can have knowledge about individual's age from external data bases), for protecting the individual age information, the data is binned into different age categories, which is represented by a transition of light orange to dark orange.

Higher number of emergency visits and readmitted greater than 30 are pointers to people being diagnosed with diabetes. Non-zero frequency for both these categories almost certainly has correlations with a diabetic condition, and points to an attribute linkage scenario. Therefore, if an attacker is able to narrow down to an age group that only exhibits these conditions, then the probability of record linkage becomes high.

Intuitively, we can assume that aggregate data visualizations provide an inherent protection against revealing sensitive attributes. However, as shown in Figure 2, that might not always be true. With respect to S_{int} or T_{int}, inside intruders with intentions to breach privacy can use the bar chart or the tree map to visualize patterns, which are otherwise difficult using textual data. Using the co-occurrence of certain age groups with only high readmitted and emergency visit numbers, they can try narrow down to the identity of an individual. With respect to S_{ext} or T_{ext}, one might already know the race and age of a person using external information. In that case, if only certain combinations of race and age reveal patterns like high frequency of readmitted and emergency cases, the attackers can guess the identity of the individual with high probability.

4 Vulnerabilities in Visualization Techniques

Before one can design control user studies or conducting surveys, it is necessary to take the first step to examine the vulnerabilities of some common visualization techniques. In this section, we consider five classes of visualization techniques using Keim's categorization [14] and one additional category of text visualization. We comment on the possible linkage attacks with such visualizations using example scenarios.

4.1 Standard 2D and 3D Visualizations

In this section, we examine the standard 2D and 3D visualizations in the context of electronic health records. Some examples of 2D visualizations are line charts and scatter plots where there is a one-to-one mapping between the data point and the screen coordinate. Advantages in using line graphs in visualizing health records are two-fold: i) we would be able to see trends and patterns easily, and ii) we would also be able to make predictions about possible future events through the examination of past data, such as the possible detection of hypertension by consistently having high blood pressure of 140/90 and above during check-ups. However, as patterns can be easily inferred from a line graph, intruders may be able to identify a patient through record linkage by combining the trend and data in a line graph with an external system. For example, a patient can be easily identified by examining a combination of their check-up appointments for high blood pressure, medical prescriptions as well as their health insurance claim costs. An example of electronic health record visualization system is *LifeLines* [19] that incorporates line graphs to represent a patients past, current, and future medical records.

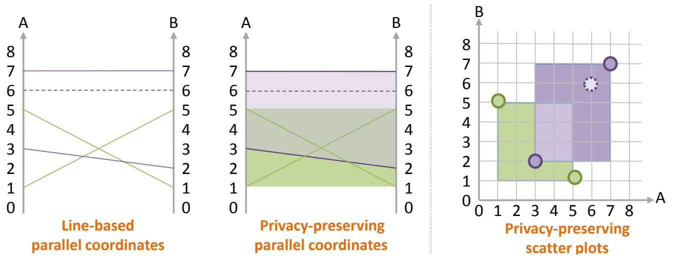


Fig. 3. Illustrating the k-anonymity privacy model based on pixel binning [7] that ensures at least k records belong to a group, where $k = 2$. The edges of clusters represent real data points and are more vulnerable to disclosure than the non-edge points, shown by dotted lines.

Most health record visualization systems now incorporates a combination of 2D and 3D visualizations to provide further details about the patients records, such as ultrasound images [19] and MRI scans. However, by generating a full 3D rendering of a patients head an intruder might be able to identify the specific patient.

Instead of showing individual items, 2D visualizations like bar charts can show aggregate information. While intuitively aggregate visualizations can be considered harmless from a privacy attack point-of-view, such an assumption might not always be true. In the case of the bar chart in Figure 2 a), some patterns stand out, like the correlation between high re-admission rate and number of emergency visits for male and female African Americans aged 50 to 60. Also, with weight over 175, there is only one category with non-zero frequency in re-admission greater than 30, and these are Caucasian males, aged 40 to 50. This implies that with knowledge of quasi-identifiers such as race and age, deducing the diabetic condition would not be hard. Also, finding outliers and correlations among diabetes indicators such as emergency visits and readmission rate is not hard, given that only certain age groups have non-zero frequency for some of the bins.

4.2 Geometrically transformed visualizations

Geometrically transformed visualization techniques like parallel coordinates and scatter plot matrices aim at finding interesting transformations of multidimensional data sets. By looking at the image, or by interaction when there is too much clutter, one can find multidimensional relationships in electronic health records. Since each record is encoded in the visualization, there are both attribute and record linkage risks. Privacy-preserving parallel coordinates and scatter plots have been proposed by generalization through k-anonymity [7], where records are visualized as clusters. When the position visual variable provides the primary encoding, then we can exploit the difference in resolution between the screen space and data space to inherently lose information through binning, etc. This when used as a parameter for controlling a privacy-preserving algorithm, can produce visualizations with both high privacy and utility.

However, it has been shown that cluster-based k-anonymous parallel coordinates and scatter plots have certain vulnerabilities from record linkage and attribute linkage [6]. An example of such vulnerability is shown in Figure 3. In this case, the cluster edges can represent real data points. If an attacker is aware about, say the age of a person, and the pixel coordinate of that data point coincides with a cluster border, then the location of the record is revealed. On the other hand, if the pixel coordinate is a non-edge point within a cluster, that provides higher privacy. With respect to attribute linkage, one can geometrically derive the number of possible cluster configurations given different values of k and use that for guessing the linkage between adjacent attributes. Reordering and brushing can enable an attacker to choose a different adjacency configuration of quasi-identifiers and browse through subset of records. Dasgupta et al.[6] have proposed different screen-space metrics that aim to constrain such interactions based on the privacy risks.

4.3 Pixel-based visualizations

Pixel-based visualizations offer a spatially cost-effective means of depicting a large number of data values, while facilitating the observation of global visual patterns (e.g., clusters, outliers and trends). A classic pixel-based visualization is a map of 3-attribute dataset. The values of one attribute (e.g., hospital bed usage) are mapped to colors according to a color map [13]. The values of the other two attributes (e.g., year and month) are mapped to two spatial dimensions. Each pixel can be recursively replaced by another pixel map (e.g., replacing a year-month pixel with a week-day pixel map) [15]. In addition to the classical layout, there are other pixel-based designs [16]. When interaction with each pixel is enabled, the typical 3-attribute pixels become the entrance points for individual data records. With some simple database queries, data from most types of electronic health records can easily be converted to a tabular form suitable for pixel-based visualization. This category of visualization techniques is thus highly usable in healthcare and medical research for visualizing data that are very close to the raw data records with little data processing or data transformation.

Because of its support of observation of global visual patterns, an intruder might utilize such a visualization for identifying outliers and small clusters, which are the weak points for privacy preservation. It is particularly risky when it is used by an inside intruder as a means of overview first and details on demand. For example, an opportunistic inside intruder could use interaction to explore outlier pixels one-by-one and bring up the corresponding anonymous data record to examine quasi-identifiers. In addition, when two relatively-vulnerable quasi-identifiers are used as the spatial dimensions, a pixel based visualization could become an effective tool for outside intruder who does not have the capability to drill into each data record. Perhaps pixel-based visualization becomes most vulnerable is when it is used a part of the coordinated multi-view visualization.

4.4 Hierarchical and Network Visualizations

Hierarchical visualizations such as the tree map [21] or network visualizations can be used to visualize relationships among different patient attributes. As shown in Figure 2b, a tree map can be used to represent the relationship among quasi-identifiers like race, age, gender; and hospital record specific information like number of emergency visits or number of times readmitted. As opposed to a bar chart, when tree map visualization is used, some of the trends are less obvious, and this is good from a privacy perspective. There is no immediate indication of the categories with zero frequency, and so the ones with non-zero frequency pop out less. They have to be found out through a sequential search. Also, because of the resolution limitation, some categories are not labeled, especially those with low areas. In a way, this is implicit privacy-preservation, as some of the metadata is hidden without applying any transformation on the data. In other words, compared to a bar chart, one has to potentially spend more time to gain information that can reveal sensitive trends in this case.

On the other hand, multivariate graphs have a higher potential for divulging information, especially for the attribute linkage scenario. This is because, even without interaction, one can analyze the connections among the different attributes to infer about the relationships. Multivariate relationships can be controlled by adding noise to the connections or aggregating the connection information for protecting sensitive trends. Also the mapping of other visual variables like color, thickness of lines can be intentionally manipulated to avoid encoding association with sensitive attributes.

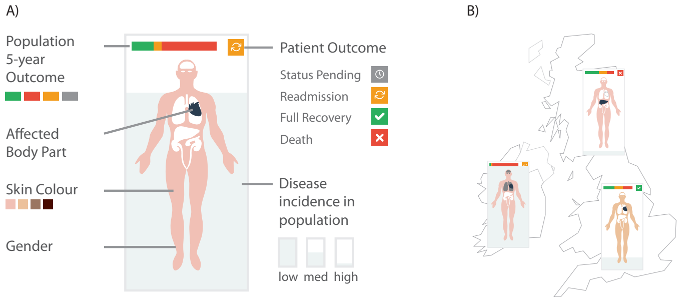


Fig. 4. A) An example glyph design showing some patient survival parameters for an unspecified disease. B) The glyphs are placed on to a map of the UK and Republic of Ireland to show three patients with differing diseases.

4.5 Glyph-based Visualizations

Glyphs are generally defined as *small visual* objects that have a *meaning*, involve *learning* and are often *metaphoric*[4]. In visualization, their *meaning* relates to two or more data attributes mapped to some visual code such as color, shape, or texture. Glyphs have the potential to represent many attributes of a data record in a single image. A further advantage to glyphs is their ability to preserve the spatial information of the data record through placement of the glyph in 2D/3D displays. For example, a glyph could be designed to show disease incidence including demographic information, patient outcomes, and spatially located to show geographical information. Figure 4 shows such an example.

There is a significant risk with this particular visualization technique due to the number of variables that can be encoded. In our example in Figure 4, more information can be potentially determined about the patients than what is explicitly encoded in the glyph. In the glyph shown over the Republic of Ireland, there is a patient with: neural, respiratory, cardiovascular, and ophthalmological problems; the incidence for the disease is very low among the population; and the person is white. There are few diseases that cause all of these problems in the Republic of Ireland. A quick search will reveal that the disease is most likely to be Hurlers Syndrome, a disease that predominantly affects the Irish traveler community (one in ten Irish travelers are carriers compared to one in nearly two hundred in the overall population) [1]. Further analysis will reveal that this condition is largely caused by inbreeding within the traveler community.

In this example, from a few variables, much more information could be inferred from the organs affected, the location, and disease incidence rates. If there were few Irish travelers in that particular area, there is a strong likelihood that the patient could be identified. One can imagine a “Guess Who” scenario where a few data attributes can very quickly narrow down the patient health record search space to a much smaller pool of individuals. Case in point, within genomics projects, from small DNA sequences (approximately 78 base pairs out of approximately three billion base pairs in the human genome) from the Y chromosome, researchers at MIT were able to extract the genealogical information (surname, relatives) and religious background of fifty people from the 1000 Genomes Project [11]. In summary, care must be taken to ensure that the data attributes with high information content are preserved.

4.6 Text Visualization

This category of visualization techniques includes word cloud [22], inkblots [1], word net [17], and so on. It was not mentioned explicitly in [14], and one may include this class of visualization in Standard

[1] https://recombine.com/diseases/hurler-syndrome

2D and 3D Visualizations. Because these techniques allow direct depiction of textual components of electronic health records, we treat them separately. Unlike other categories, text visualization can bring a substantial amount of raw text onto a display. In some cases, text is accompanied by statistical significance; in other cases, the connections between discretely-placed words can be easily established. These techniques can provide data custodians and analysts with an efficient tool for exploring a huge amount of data, such as doctors notes and patients feedback written in free-text.

Although the records of these texts may have been undergone a process of anonymization, it is necessary to recognize the fact that text analysis is a semantically-rich process, and automated techniques have not yet reached a satisfactory level in practice. Hence raw free-text can be highly vulnerable, and one natural defence is offered by the fact that reading text is time-consuming. Thus any visualization techniques that were designed to improve analysts speed of reading and exploring texts could be abused by intruders to attack the privacy the data. In particular, text visualization should be used with an extreme care in public dissemination. For example, a word cloud may be designed to encourage the public to pay attention to those big-letter words, but can easily expose special cases through small-letter words, which are often not carefully checked by the creators of such visualizations. Visualization techniques such as InkBlots and word net almost offer direct exposure of the raw text. In addition to the risk of exposing specific content, the writing styles and errors can also become the attacking points of intruders.

5 Conclusion

In this paper, we have have presented an overview of the opportunities and challenges involving privacy-preserving visualization of electronic health record data. Privacy-preserving visualization is a relatively nascent research area compared to its data mining counterpart. Therefore, it is important to reflect on how visualizations can be utilized for analyzing health-care data but not at the cost of privacy. To address the challenges described in this paper, future research needs to investigate how traditional anonymization techniques like generalization, perturbation, etc., need to be applied in conjunction with tuning visualization-specific parameter for controlling privacy of visual representations. In addition, there needs to be a thorough analysis of how interaction techniques can be leveraged in case of different attack scenarios. We believe by reflecting upon the challenges described here, we can take more concrete steps towards achieving a synergy between optimal privacy and high utility for privacy-preserving, interactive visualizations.

References

[1] A. Abbasi and H. Chen. Categorization and analysis of text in computer mediated communication archives using visualization. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 11–18. ACM, 2007.

[2] D. Agrawal and C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the ACM Symposium on Principles of database systems*, pages 247–255. ACM, 2001.

[3] W. Aigner and S. Miksch. Carevis: integrated visualization of computerized protocols and temporal patient data. *Artificial intelligence in medicine*, 37(3):203–218, 2006.

[4] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Eurographics State of the Art Reports*, pages 39–63, 2013.

[5] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum*, 31(3pt2):1015–1024, 2012.

[6] A. Dasgupta, M. Chen, and R. Kosara. Measuring privacy and utility in privacy-preserving visualization. In *Computer Graphics Forum*, volume 32, pages 35–47. Wiley Online Library, 2013.

[7] A. Dasgupta and R. Kosara. Adaptive privacy-preservation using parallel coordinates. *Transactions on Visualization and Computer Graphics*, 17(12):2241–2248, 2011.

[8] C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[9] B. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):14, 2010.

[10] A. Gkoulalas-Divanis, G. Loukides, and J. Sun. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, 50:4–19, 2014.

[11] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[12] W. Horn, C. Popow, and L. Unterasinger. Support for fast comprehension of icu data: Visualization using metaphor graphics. *Methods of information in medicine*, 40(5):421–424, 2001.

[13] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1):59–78, 2000.

[14] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[15] D. A. Keim, M. Ankerst, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *Proceedings of the 6th conference on Visualization'95*, page 279. IEEE Computer Society, 1995.

[16] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu. Pixel bar charts: a visualization technique for very large multi-attribute data sets. *Information Visualization*, 1(1):20–34, 2002.

[17] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[18] A. Perer and J. Sun. Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, volume 2012, page 716. American Medical Informatics Association, 2012.

[19] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: using visualization to enhance navigation and analysis of patient records. In *Proceedings of the AMIA Symposium*, page 76. American Medical Informatics Association, 1998.

[20] C. E. Shannon. Communication theory of secrecy systems*. *Bell system technical journal*, 28(4):656–715, 1949.

[21] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99, 1992.

[22] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near martha? exploring name entities in literary text collections. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 107–114. IEEE, 2009.