Privacy-Utility Tradeoff: Studying the Boundaries of Anonymization in Health Data Visualization Design

Jialin Sun

Shandong University of Science and Technology, Shandong, China

Received: 21 Jul 2025
Revised: 25 Jul 2025
Accepted: 30 Jul 2025
Published: 08 Aug 2025

Abstract*:* With the increasing prevalence of health data, how to maximize the utility of data while ensuring privacy protection has become a core issue in health data visualization design. This paper explores the issue of anonymization boundaries in health data visualization design, focusing on the trade-off between privacy and utility. Through theoretical analysis and experimental verification, the application effects of different anonymization techniques in visualization are studied, and the optimal balance between privacy protection and visualization utility is proposed. Experimental results show that under moderate anonymization boundary settings, privacy protection and data utility can be effectively balanced, preventing privacy leaks while retaining sufficient analytical value. This study provides theoretical support for the selection of privacy protection technologies and the setting of anonymization boundaries in health data visualization and offers new perspectives and directions for future research.

Keywords: Health data; Privacy protection; Utility trade-off; Anonymization techniques; Visualization design; Differential privacy

1 INTRODUCTION

With the rapid development of digital health technology, the collection, storage and analysis of health data have gradually become an important means to improve medical quality and optimize health management. From electronic medical records to the large amount of data generated by health monitoring equipment, this information has provided great help to doctors, researchers and public health decision makers. However, the privacy and sensitivity of these health data make it an urgent problem to balance privacy protection and data utility [1]. Health data not only involves personal medical information, but also includes biometrics, lifestyle and other content. Once leaked or abused, it will cause immeasurable risks to individuals. Therefore, when using this data for scientific research and decision support, privacy protection principles must be strictly followed.

However, with the widespread application of health data, the contradiction between privacy protection and data utility has become increasingly prominent. In many cases, to achieve efficient data analysis and visualization, researchers or decision makers often need to process and de-identify the data, but this process may damage the accuracy and effectiveness of the data [2]. Such as in the process of anonymizing health data, excessive concealment of information may cause data distortion and reduce the analytical value of the data. Therefore, how to maximize the utility of data while protecting privacy has become a core issue in health data visualization design [3].

This study aims to explore privacy protection in health data visualization, focusing specifically on the impact of data anonymization boundaries on visualization utility. Specifically, this study analyzes how to protect individual privacy through appropriate anonymization in health data visualization while ensuring that the visualization quality and analytical utility of the data are not excessively compromised. The significance of this study lies in providing a theoretical basis for data visualization design and proposing solutions to the trade-off between privacy protection techniques and data utility.

This paper is organized as follows: Section 2 reviews relevant literature, discussing health data privacy protection, data visualization utility analysis, and existing anonymization technology research; Section 3 provides a detailed introduction to health data anonymization technologies and their current applications; Section 4 further explores the privacy-utility trade-off and proposes relevant theoretical models; Section 5 focuses on analyzing anonymization boundaries in health data visualization design and explores the impact of different anonymization strategies on data utility; Section 6 validates this theoretical framework through experiments and data analysis; and finally, Section 7 summarizes the research findings and outlines future research directions.

2 RELATED WORKS

With the widespread application of health data, privacy protection has gradually become a focus of attention. In this context, researchers have proposed a variety of solutions to protect data privacy while minimizing the impact of data utility. In health data visualization, how to reasonably anonymize data has become a key research topic. This Section will focus on reviewing the relevant technologies and regulations for health data privacy protection, the utility analysis methods for data visualization, and the exploration of existing research in the field of health data visualization anonymization [4].

First, the privacy protection of health data has always been a difficult problem in the field of data processing. As an important means of privacy protection, data anonymization removes or obscures personal identity information so that data can no longer be directly associated with a specific individual. Common data anonymization technologies include k-anonymity, l-diversity, t-precision, etc. They process data through different algorithms to ensure that when the data is made public or shared, the identity of the individual cannot be traced back to the data itself [5]. Such as k-anonymity requires that each piece of data be consistent with at least k other data entries in terms of sensitive attributes, thereby reducing the possibility of identifying an individual through a single data point. In addition, differential privacy, as a privacy protection technology that has emerged in recent years, has been widely used in the process of data publishing and analysis. It introduces noise to ensure that the data analysis results cannot statistically leak any information about individuals. Although these technologies provide a theoretical basis for privacy protection, how to achieve anonymization without compromising the utility of the data remains a difficult problem [6].

Second, data visualization, as an effective way to display data, plays an important role in health data analysis. The main goal of health data visualization is to transform complex medical data into easy-to-understand and easy-to-interpret graphical forms, helping decision makers, researchers and doctors to extract valuable information from them and then make decisions or scientific research. Visualization methods include but are not limited to bar charts, line charts, heat maps, geographic information system (GIS) maps, etc., which can show the distribution, trends, changes and geographic spatial characteristics of data [7]. However, with the growth and complexity of health data, how to strike a balance between privacy protection and data utility in the visualization process has become a challenge. Excessive anonymization may lead to distortion of visualized information, thereby affecting the effect of data analysis. Therefore, in data visualization design, how to find the best balance between privacy protection and utility has become a key issue in the design process.

In existing research, there has been some exploration of anonymization technology in health data visualization. Many studies focus on how to optimize data anonymization processing through algorithms to maximize the analytical utility of data. Such as some scholars have proposed a health data visualization framework based on differential privacy, which ensures that the statistical characteristics and visualization effects of the data are as unaffected as possible while protecting privacy by adding noise [8]. At the same time, there are also studies dedicated to evaluating the impact of different anonymization methods on the utility of data visualization and exploring how to select appropriate anonymization technology to improve visualization effects while ensuring data security. The trade-off analysis between privacy protection and utility has become a research hotspot in this field. In the existing literature, some studies have proposed a trade-off framework between privacy protection and data utility by constructing mathematical models or optimization algorithms, and conducted experimental verification, which provides theoretical guidance for data visualization design [9].

In summary, the issue of health data privacy protection and data visualization utility has become an important topic in academia and practice. Although many studies have provided theoretical basis and technical support for solving this problem, how to effectively balance privacy protection and data utility in the anonymization process of health data is still an issue that needs to be explored in depth.

3 HEALTH DATA ANONYMIZATION TECHNOLOGY

In the field of health data processing, anonymization is a core privacy protection method. Its purpose is to strip or weaken the identifiable association between data and specific individuals, thereby reducing the risk of privacy leakage during data sharing and analysis. According to different technical implementation methods, anonymization methods can be divided into various types, among which the representative ones include k-anonymity, differential privacy and pseudo-anonymization [10]. K-anonymity reduces the probability of a single individual being uniquely identified by grouping records so that each record is identical to at least k-1 other records in terms of key identifier attributes; differential privacy introduces random noise into the data or query results to ensure that even if an attacker has external background knowledge, he cannot significantly infer the specific information of an individual; pseudo-anonymization focuses on removing direct identification marks, such as name, ID number, etc., but retains indirect identifiers that may be useful in statistical analysis. Different technologies have significant differences in protection strength, applicable scenarios and impact on data structure [11].

In practical applications, anonymization technology has been verified in multiple health data processing scenarios. Such as when some hospitals publish disease epidemic trend reports, they use k-anonymity to group patient information to prevent individuals from being re-identified through feature combinations; when research institutions share genomic databases, they use differential privacy to add controlled noise to statistical results, thereby ensuring the feasibility of research while protecting the privacy of data subjects [12]. However, anonymization technology also faces many challenges in the application of health data. Health data is often highly granular and highly correlated, especially in the fields of genomic data, medical imaging, and continuous monitoring data. Simple anonymization operations can easily be cracked by high-dimensional data feature associations. In addition, the implementation of anonymization may also be restricted by factors such as the high frequency of data updates and the strong availability of external data sources, making the effect of privacy protection easily weakened over time.

In the process of data anonymization, the conflict between privacy protection and data utility is inevitable. Excessive anonymization may cause data distortion, thereby reducing its value in analysis and decision-making. Such as in geographic epidemiological studies, if the patient's location is blurred to an overly large geographic unit for privacy protection, the disease transmission path may be misjudged, affecting the accuracy of public health interventions. This data distortion is particularly evident at the visualization level. The noise or grouping introduced by anonymization can weaken the spatial resolution, temporal precision, and numerical authenticity of the data, causing the trends, clusters, or outliers presented in the charts to deviate from the actual situation. This loss of utility not only affects the accuracy of research results but can also mislead decision-making. Therefore, how to reduce data identifiability while preserving its structural characteristics and statistical significance as much as possible is a core issue that needs to be addressed in visualization design using health data anonymization technology.

4 THE PRIVACY-UTILITY TRADE-OFF

The tension between privacy and utility is particularly prominent in the use and sharing of health data. Privacy needs are diverse, subject to legal and ethical constraints and reflect individual sensitivity and expectations regarding privacy protection. In health data applications, users' privacy needs are not singular but vary depending on the nature of the data, the usage scenario, and the participants involved. Such as individuals may prefer that sensitive information such as their medical records and medical history remain private. Especially when sharing medical data or participating in public health research, they often want to provide effective data support while ensuring privacy. Different user groups, diverse cultural backgrounds, and privacy regulations in the legal environment all contribute to the diversity of privacy needs. User privacy needs are not solely dependent on technical measures such as anonymization or encryption; they are also influenced by factors such as the degree of information exposure and the purpose of the data.

The inherent sensitivity of health data determines the importance of its privacy protection. Health data not only includes personal disease information and treatment records, but also includes biometrics, genetic information, lifestyle, and other aspects. With technological advances, an increasing amount of high-dimensional health data is being collected and analyzed, and this data is extremely sensitive. Once leaked or used maliciously, data can have long-term negative consequences for individuals, such as identity theft and insurance discrimination. Therefore, ensuring that user privacy is protected to the greatest extent possible during data sharing and use has become a core issue in data protection. Privacy protection goes beyond simply preventing the leakage of personal identities and addresses the risk of inferring sensitive personal information through data.

On the other hand, utility requirements are equally diverse. In health data visualization design, utility requirements encompass multiple aspects, including data analysis, research, and decision support. The utility of health data is reflected not only in its direct analytical value but also in its support for medical research, disease prediction, and public health policymaking. For researchers, the utility of health data lies in its ability to reveal scientific questions such as disease patterns and treatment efficacy assessment. For policymakers, the utility of data lies in its ability to provide a basis for decision-making in public health management and resource allocation. However, excessive privacy protection can lead to data distortion, affect data analysis results, and even negate the original meaning of the visualization design.

In data visualization design, utility requirements manifest themselves as multi-dimensional analytical requirements. On the one hand, data accuracy is a key criterion for measuring utility, and data visualizations must truly reflect the underlying information. On the other hand, ease of use and intuitiveness are equally important. Researchers and decision-makers need to quickly and accurately extract meaningful insights from data, requiring data visualizations to be both highly accurate and understandable. Therefore, when designing health data visualizations, a key challenge is how to balance the multidimensional utility of data while avoiding utility loss due to excessive privacy protection measures.

To strike a balance between privacy protection and data utility, scholars have proposed privacy-utility trade-off models. These models attempt to quantitatively analyze the impact of privacy protection on data utility and establish an optimal path between the two. A common model framework is an optimization model based on differential privacy. This model protects privacy by introducing noise into the data and then balances privacy protection with utility by controlling the noise level. Model validation typically relies on experimental data and feedback from real-world application scenarios. Such as in public health data analysis, models adjust noise parameters to measure the impact of privacy protection on visualization and data analysis accuracy. Furthermore, the model's application examples demonstrate how data visualization effects vary under different privacy protection strategies. This research demonstrates that appropriate privacy protection measures can effectively reduce the risk of privacy breaches while maintaining manageable impacts on data utility. When utilizing differential privacy methods, a good balance between privacy and utility can be achieved by properly setting the noise level.

In summary, the trade-off between privacy and utility is not only a technical challenge in health data visualization design but also a topic of ongoing exploration in theory and practice. By establishing a reasonable trade-off model and integrating it with specific application scenarios, it can effectively guide how to strike a balance between privacy protection and utility optimization in practical operations.

5 THE BOUNDARIES OF ANONYMIZATION IN HEALTH DATA VISUALIZATION DESIGN

In the context of health data visualization, the anonymization boundary refers to the maximum or minimum tolerance for de-identifying and obfuscating data to protect personal privacy during data processing. This boundary represents both a watershed in technical implementation and a critical point where designers must strike a balance between privacy protection and information presentation. Too low anonymization can result in data still being identifiable, making it difficult to meet privacy regulations, while too high anonymization can cause visualizations to lose detail and analytical value. Therefore, the anonymization boundary acts like a "valve" in health data visualization—by adjusting its position, the strength of privacy protection and data usability can be controlled within a certain range.

The relationship between the anonymization boundary and privacy protection is dynamic and closely linked. With the continuous advancement of data analysis technology and external data sources, anonymization methods once considered secure may become re-identifiable or compromised in new contexts. This means that the anonymization boundary is not static and needs to be redefined as the technological environment and legal standards evolve. In health data visualization design, adjusting this boundary is not only a technical issue but also involves ethical judgment and legal compliance. Such as in the release of public health epidemic data, city-level geographic aggregation may preserve sufficient trend information while protecting privacy. However, in rural areas with small sample sizes, the same aggregation scale may still pose identification risks and require further blurring.

Different anonymization boundaries often significantly impact the utility of visualizations. When the degree of anonymization is low, visualizations can retain more details of the original data, such as precise geographic location, time series variations, and fine-grained grouping characteristics. This is very beneficial for tasks such as pattern recognition and anomaly detection. However, this design may also allow sensitive individual characteristics to be indirectly identified in the visualization results. When the degree of anonymization is increased, visualizations may lose detail. Such as when the time granularity is expanded to months or quarters, or when geographic locations are blurred to provinces or states, the clustering characteristics of the data become blurred, potentially obscuring trends and differences. This reduction in utility can lead to delayed responses or misjudgments in real-world decision-making. Therefore, the anonymization boundary must be carefully selected in the design to ensure a balance between privacy protection and utility preservation.

In real-world cases, the effectiveness of setting anonymization boundaries varies significantly. Such as when publishing hospitalization rate heat maps, some hospitals blurred patient addresses to the first two digits of their postal codes and set the time window to one week. This approach protected privacy while maintaining sufficient spatial and temporal resolution to facilitate analysis of localized disease outbreaks. Such cases are generally considered successful because they effectively protect personal information without compromising core analytical tasks. However, there are also failures. Such as a local public health department, responding to an infectious disease outbreak, over-aggregated data to the state level and set the time granularity to quarterly. As a result, the visualization failed to reflect the localized outbreak trends, delaying the deployment of prevention and control measures. Such cases demonstrate that if the setting of anonymization boundaries deviates from actual needs, it can undermine the actual value of data visualization and even lead to adverse consequences at critical moments.

Therefore, the design of anonymization boundaries should be based on a comprehensive assessment of data sensitivity, analytical task objectives, and audience needs. By dynamically adjusting the degree of anonymization, combining with legal regulations, technical capabilities, and practical application scenarios, the dual goals of privacy protection and utility maximization can be achieved in health data visualization. This is not only a matter of technical optimization but also the result of a long-term negotiation between data ethics and information dissemination.

6 EXPERIMENT AND DATA ANALYSIS

This study's experimental design aims to systematically evaluate the relationship between the utility of health data visualization and the strength of privacy protection under different anonymization boundaries. The dataset consists of a mixture of real-world public health databases and simulated data, encompassing a wide range of health information types, including disease incidence, patient characteristics, and geographic and time series information. The public data is partially derived from a de-identified infectious disease surveillance platform, while the simulated data is constructed based on real-world distributional characteristics, allowing for flexible experimental control without compromising privacy. The experimental design sets multiple anonymization boundaries, ranging from low-strength (E.g., street-level location and daily time) to high-strength (E.g., province-level location and seasonal time), and generates corresponding visualizations for each setting, including heat maps, time trend plots, and cluster analysis plots. Privacy protection is quantified using a re-identification risk assessment, while utility is measured using accurate metrics, pattern fidelity, and task completion rate.

The experimental results show that the impact of different anonymization boundaries on visualization performance exhibits nonlinear characteristics. When the anonymization strength is low, visualization can fully present the temporal and spatial distribution characteristics of the data, but the risk of privacy leakage increases significantly. When the anonymization strength is too high, the privacy risk is significantly reduced, but the visualization effect is significantly distorted. In the tasks of identifying disease clusters and detecting abnormal fluctuations, information loss reduces the credibility of analytical conclusions. Notably, at moderate anonymization boundaries (Such as aggregating geographic information to the district and county level and adjusting the temporal granularity to weekly), the risk of re-identification is effectively reduced while preserving most of the trends and spatial distribution characteristics required for analysis. This result suggests that there is a relatively optimal range for selecting anonymization boundaries that achieve acceptable levels of privacy and utility. Furthermore, different anonymization techniques have different effects on visualization utility. Such as when the noise amplitude is appropriately controlled, the visualization trends of differential privacy-based methods are more consistent with the original data, while k-anonymity methods based on strong aggregation are more likely to lose details.

In-depth analysis of the experimental results shows that the relationship between anonymization boundaries and visualization utility depends not only on data characteristics but also on the analytical task objectives. For public health monitoring tasks that require accurate tracking of short-term fluctuations, medium or low levels of anonymization are more appropriate; for long-term trend analysis or cross-regional comparisons, high-intensity anonymization can also meet the needs with lower privacy risks. This finding suggests that the setting of anonymization boundaries should be task-driven rather than uniformly fixed. In addition, the privacy-utility balance curve proposed in the theoretical model was partially verified in the experiment, indicating that quantitative risk assessment and utility indicators can provide an operational reference framework for actual visualization design. At the practical level, this means that public health agencies should establish a dynamic anonymization boundary setting mechanism when publishing data, and flexibly adjust strategies based on the purpose of data use, audience type, and risk assessment results, to achieve effective protection of sensitive health data without sacrificing key analytical value.

7 CONCLUSION AND OUTLOOK

This study systematically explores the trade-off between privacy protection and data utility in health data visualization design, focusing on the anonymization boundary. Experiments validate the dynamics of visualization effects and privacy risks at different levels of anonymization. The results demonstrate that privacy and utility are not simply contradictory but rather can be optimally balanced by properly setting anonymization boundaries. At a moderate level of anonymization, the risk of data re-identification is significantly reduced while sufficient fine-grained information is retained to support analysis and decision-making. This "optimal balance" provides a feasible design basis for the secure sharing and effective utilization of health data. Furthermore, experiments revealed that different anonymization techniques perform differently in visualization. Differential privacy, under moderate noise control, effectively balances privacy and trend fidelity, while strong aggregation-based methods offer more robust protection but are susceptible to loss of detail. This suggests that designers need to flexibly select technical strategies based on task objectives.

Although this study has achieved preliminary results in both theoretical and experimental aspects, it still has certain limitations. First, the dataset used is limited in type and size. While it covers some real public health data and simulated data, it does not fully encompass all types of health information, particularly high-dimensional genetic data and multimodal medical imaging. Second, the evaluation of anonymization effectiveness and utility primarily relies on existing quantitative metrics, which may not fully reflect the perceived experience and decision-making impact of actual use in different application scenarios. Furthermore, the controlled experimental environment of this study still differs somewhat from real-world data release and usage scenarios, which may affect the applicability of the research conclusions in complex real-world settings.

Future research can be further expanded in several directions. First, health data of more types and sources should be included, especially multimodal and cross-platform data, to verify the applicability and stability of anonymization boundaries in more complex data structures. Second, methods combining artificial intelligence with privacy-preserving computing technologies (Such as federated learning and secure multi-party computation) can be explored to achieve efficient visualization and analysis without directly exposing the original data, thereby achieving a higher level of synergistic optimization of privacy and utility. Furthermore, research on user privacy perceptions and trust mechanisms should be strengthened, integrating technical evaluation with user psychological models to ensure that the setting of anonymization boundaries not only meets technical and legal standards but also meets public acceptance and expectations. Ultimately, through interdisciplinary collaboration, building a compliant and efficient health data visualization ecosystem will contribute to the sustainable and secure development of data-driven healthcare and public health.

REFERENCES

[1] Alkaabi, S. H. (2024). VISUALIZING PRIVATELY PROTECTED DATA: EXPLORING THE PRIVACY-UTILITY TRADE-OFFS.

[2] Bhattacharjee, K., Chen, M., & Dasgupta, A. (2020, June). Privacy-preserving data visualization: reflections on the state of the art and research opportunities. In *Computer Graphics Forum* (Vol. 39, No. 3, pp. 675-692).

[3] Valdez, A. C., & Ziefle, M. (2019). The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies*, *121*, 108-121.

[4] Im, E., Kim, H., Lee, H., Jiang, X., & Kim, J. H. (2024). Exploring the tradeoff between data privacy and utility with a clinical data analysis use case. *BMC Medical Informatics and Decision Making*, *24*(1), 147.

[5] Nanayakkara, P., Bater, J., He, X., Hullman, J., & Rogers, J. (2022). Visualizing privacy-utility trade-offs in differentially private data releases. *arXiv preprint arXiv:2201.05964*.

[6] Sarmin, F. J., Sarkar, A. R., Wang, Y., & Mohammed, N. (2024). Synthetic data: Revisiting the privacy-utility trade-off. *arXiv preprint arXiv:2407.07926*.

[7] Kamal, O. S., Sohail, S. A., & Bukhsh, F. A. (2024). Optimizing Privacy-Utility Trade-Off in Healthcare Processes: Simulation, Anonymization, and Evaluation (Using Process Mining) of Event Logs. In *14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications SIMULTECH 2024* (pp. 289-296). SCITEPRESS.

[8] Franzen, D., Müller-Birn, C., & Wegwarth, O. (2024). Communicating the privacy-utility trade-off: Supporting informed data donation with privacy decision interfaces for differential privacy. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1-56.

[9] Dewri, R., Ray, I., Ray, I., & Whitley, D. (2011). Exploring privacy versus data quality trade-offs in anonymization techniques using multi-objective optimization. *Journal of Computer Security*, *19*(5), 935-974.

[10] Pilgram, L., Meurers, T., Malin, B., Schaeffner, E., Eckardt, K. U., Prasser, F., & GCKD Investigators. (2024). The costs of anonymization: case study using clinical data. *Journal of medical Internet research*, *26*, e49445.

[11] Appenzeller, A., Leitner, M., Philipp, P., Krempel, E., & Beyerer, J. (2022). Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, *12*(23), 12320.

[12] Sarmin, F. J., Sarkar, A. R., Wang, Y., & Mohammed, N. (2025). Synthetic data: revisiting the privacy-utility trade-off: F. Jahan Sarmin et al. *International Journal of Information Security*, *24*(4), 156.