Balancing Data Transparency and Privacy Protection in Epidemiological Visualizations: A Targeted Narrative Review of Privacy-Preserving Data Visualizations

Julian Haessner
University of North Carolina State

Philipp Haessner
University of Florida

Joseph Thomas
University of Central Arkansas

Epidemiology relies on data visualization to identify disease patterns, causes, and effects. These visualizations often contain sensitive information that risks re-identification if not properly anonymized, while excessive anonymization may lead to information loss, distort findings, and reduce usability. This review summarizes strategies for balancing transparency with privacy preservation in epidemiological visualizations. Through targeted database searches, we identified common anonymization strategies, adaptable for various visualizations. Each visualization type requires a different approach, and often a combination of methods to yield the strongest privacy protection. Privacy-preserving visualization remains underdeveloped, and further empirical validation and user studies are needed.

Keywords: data visualization, sensitive data, patient privacy, privacy protection, data anonymization, information loss, epidemiology, public health

INTRODUCTION

The rapid advancement of digital technologies has led to significant increases in both the volume and diversity of data across various sectors, including retail, banking, e-commerce, and healthcare (Ram Mohan Rao, Murali Krishna, and Siva Kumar). Healthcare data is particularly sensitive, as it contains patient records, treatment plans, medical test results, and prescriptions (Avraam et al.). In 2020, global healthcare data surpassed 25,000 petabytes and is projected to grow by 20-40% each year. This significant growth raises concerns and challenges regarding data security and privacy (Jin et al. 2019).

Privacy is a fundamental human right encompassing individual control over personal information (Ram Mohan Rao et al. 2018). Regulations, such as the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States, require medical data to be stored and shared securely (Jin et al.). Yet, even under these frameworks, breaches and misuse remain common.

Unlike established areas of privacy-preserving data publishing and mining, protecting sensitive information within visualizations is an emerging challenge. Visualizations help identify trends and outliers, but may also expose individuals if safeguards are absent. Excessive anonymization, on the other hand, risks distorting results and reducing usability. The core challenge is balancing interpretability with privacy (Avraam et al. 2021).

OBJECTIVES & METHODS

This paper reviews privacy-preserving visualization methods relevant to epidemiology and public health, focusing on standard visualization and anonymization techniques. We conducted a selective search across multiple databases (e.g., PubMed, IEEE Xplore, Google Scholar) using search terms related to epidemiology, data visualization, and privacy. Results were narratively synthesized to present a high-level overview of widely used privacy-preserving data visualization methods. The aim of this targeted overview is not to deliver a systematic or comprehensive review, but rather to serve as a practical resource for researchers and practitioners seeking to design visualizations that protect individual privacy while preserving interpretability for public health decision-making.

RESULTS

Privacy Concerns in Healthcare Data Collection and Sharing

Healthcare data is a frequent target for attackers who seek to exploit sensitive information. Breaches occur daily due to the wide availability of disclosed information and the attackers' ability to combine health records with external knowledge. Additionally, individuals may inadvertently share information through mobile apps or due to incomplete consent review (Jin et al. 2019; Ram Mohan Rao et al. 2018).

Risks fall into two broad categories:

1. Re-identification risk (How attackers gain access):
    - Record linkage: Linking health records to public datasets using quasi-identifiers (e.g., age, race, gender) can re-identify individuals (Bhattacharjee, Chen, and Dasgupta 2020).
    - Attribute linkage: Correlations between quasi-identifiers and attributes (e.g., demographic details with diagnoses) allow inference of sensitive information (Bhattacharjee et al. 2020).
2. Consequence risks (What happens after disclosure):
    - Discrimination: Disclosed data may be used in statistical analyses that reveal hidden patterns, leading to biases such as denial of insurance or unequal treatment.
    - Surveillance: Continuous monitoring, common in retail analytics, becomes problematic in healthcare, where patients may be monitored to study health status or behaviors without their knowledge or consent (Ram Mohan Rao et al. 2018).

Epidemiology, which depends heavily on sensitive health data, is especially vulnerable to these risks.

The Role of Data Visualization in Epidemiology

Epidemiology studies patterns, causes, and effects of health conditions in populations. Over the past century, the field has faced increasing challenges in effectively analyzing and disseminating large volumes of complex health data (Lau et al.).

Data visualization plays a central role in this effort by simplifying complex, often spatial and temporal datasets, helping researchers detect patterns, policymakers act on findings, and allowing the public to better understand health risks (Carroll et al. 2014; Unwin 2020).

Effective visualizations can enhance trustworthiness, improve information delivery, and reduce cognitive burden (Park et al. 2022). Their importance has grown further with the rise of digital distribution platforms and accessible visualization software (Midway 2020). Interactive methods add value by enabling the exploration of multiple facets of disease and social determinants of health (Chishtie et al.).

At the same time, new challenges have emerged. The rising volume and increasing specialization of epidemiology (e.g., genetic, molecular, psychiatric) necessitate careful consideration of how data are collected, analyzed, and represented (Frérot et al.). Visualizations must also account for the diverse needs of users, varying levels of health literacy, workflow integration, and the importance of building trust (Carroll et al.). Furthermore, there is limited guidance on privacy-preserving practices in epidemiology, and a need for training and resources to help healthcare professionals use visualizations effectively and understand the challenges related to data protection (Chishtie et al., 2022; Kim et al.).

Researchers and practitioners rely on anonymization techniques that protect sensitive information while preserving data utility to address these challenges. The following section reviews key methods for safeguarding data visualizations in epidemiology.

Anonymization Techniques for Protecting Sensitive Data

Anonymization ensures that an individual cannot be uniquely identified within a larger group, often referred to as the "anonymity set". Because individuals usually show distinct behaviors or attributes, achieving perfect anonymity is challenging (Pfitzmann and Hansen 2008). Nevertheless, various techniques are available to minimize identifiability and protect sensitive information. Key techniques include:

- Generalization: Replace specific data values with broader ones to reduce the granularity of the information. Examples include grouping ages into decades or fixed ranges, showing only the first three digits of a ZIP code, or merging detailed regions into larger areas such as counties or states (Samarati and Sweeney 1998; Yaseen et al. 2018).
- Suppression: Remove or mask sensitive values (e.g., with asterisks, number signs, or nulls) within cells, or drop entire rows when counts fall below a threshold, to reduce links between quasi-identifiers and sensitive attributes (Cox 1980; Samarati and Sweeney 1998; Xu et al. 2014). Newer methods (e.g., ℓρ-suppression, suppression slicing) target only high-risk values and scale better for larger datasets (Elanshekhar and Shedge 2017; Gunawan et al. 2022).
- Aggregation: Summarizes data across groups (e.g., averages, counts, sums) to reduce identifiability. Novel aggregation methods strengthen this anonymization technique by using cryptography or delinking and dispersing data before aggregation (Almalki and Soufiene 2021; He et al. 2007; Zhang, Sarvghad, and Miklau 2020).
- Geographic masking: Perturbs or displaces geocoordinates (e.g., random noise, donut masking) to protect location privacy, with displacement typically scaled to population density to balance confidentiality and spatial accuracy (Armstrong, Rushton, and Zimmerman 1999; Kwan, Casas, and Schmitz 2004; Zandbergen 2014).
- Perturbation: Adds controlled noise or randomness to sensitive data values to prevent re-identification while preserving aggregate statistics (Wilson and Rosen 2005). Advanced perturbation methods (e.g., scaling, shearing, rotation, M6, NOS2R, NOS2R2) further improve the balance between privacy protection and data utility (Rahman, Paul, and Sattar 2023; Roman 2023; Turgay and İlter 2023).
- Synthetic data: Use model-generated data that mimics real datasets to protect individual identities while enabling research and AI development (Chen et al. 2021; Jordon et al. 2022). Synthetic data can closely approximate real EMRs but may face challenges in dataset linkage, extension, and compliance with clinical standards (Benaim et al. 2020; Chen et al. 2021; Jordon et al. 2022).

Other methods, such as visual blurring, pseudonymization, data swapping, or cryptography, exist but are less commonly applied in visualization. For example, visual blurring distorts identifiable information (e.g., faces, names, medical conditions) in medical images and documents, while balancing privacy with diagnostic quality (Chen et al., 2006; Vishwamitra et al.). When applied carefully, blurring has been shown to have minimal impact on the accuracy of deep learning models trained on visual data (Jiang et al. 2022). Similarly, pseudonymization replaces identifiers with artificial labels or codes, which can be personal (e.g., artificial ID numbers, nicknames) or transaction-based (randomly generated), with each pseudonym linked to exactly one holder (Pfitzmann and Hansen 2008). It reduces identifiability while maintaining data

usability for analysis, supports patient-controlled e-health systems (Riedl et al., 2007, 2008), and is also applied in genomic research to prevent the re-identification of biospecimens (Aamot et al.). Another approach is swapping, which exchanges values between records to protect confidentiality while preserving overall statistics (Dalenius and Reiss 1982; Reiss 1984). It can be extended to summary statistics or census blocks, though it involves a trade-off between privacy and data accuracy (Fienberg and McIntyre 2004). Finally, cryptography utilizes encryption and watermarking to secure medical texts and images by converting data into unreadable formats and embedding hidden identifiers, ensuring confidentiality. Given the sensitivity of medical data, these methods must be fully reversible (Boussif, Aloui, and Cherif 2018; Kester et al. 2015).

Privacy-Preserving Anonymization Approaches

Frameworks such as *k*-anonymity, *l*-diversity, *t*-closeness, probabilistic and deterministic anonymization, and differential privacy build on the core techniques introduced in Section 5. These structured approaches systematically combine and extend those methods to make sensitive records indistinguishable within a dataset while preserving analytical value. Anonymization approaches typically modify quasi-identifiers in patient records (e.g., age, race, and gender) through various anonymization techniques (Bhattacharjee et al. 2020):

- *K*-anonymity: Ensures that each record is indistinguishable from at least *k–1* others (Sweeney 2002). While widely used, it is vulnerable to background knowledge attacks, and if sensitive attributes lack diversity, values can still be inferred (Bhattacharjee et al. 2020).
- *L*-diversity: Extends *k*-anonymity by requiring that each equivalence class contains at least *l* diverse values for sensitive attributes, thereby reducing the risk of attribute disclosure further (Jin et al. 2019; Li, Li, and Venkatasubramanian 2006; Machanavajjhala et al. 2007).
- *T*-closeness: Further refines *l*-diversity by requiring that the distribution of sensitive attributes within each equivalence class remains close to the overall dataset distribution, limiting inference attacks. A smaller *t* value indicates more similarity between the equivalence class and the original distribution and greater privacy protection (Li et al. 2006).
- Probabilistic anonymization: Perturbs individual values with stochastic noise to reduce re-identification risk while preserving aggregate statistics. Its effectiveness depends on carefully controlling noise levels and protecting parameters such as the random seed because excessive or repeated perturbation can distort the data (Avraam et al. 2021; Avraam, Jones, and Burton 2022).
- Deterministic anonymization: Uses methods such as k-nearest neighbors (KNN) to replace individual-level observations with centroids of nearby values. This approach redistributes continuous variables while preserving overall patterns. Deterministic anonymization protects sensitive data in diverse domains (Avraam et al. 2022).
- Differential privacy: Adds noise at a predetermined rate to all observations so that attackers cannot determine whether an individual's data is included (Bhattacharjee et al. 2020; Jin et al. 2019). This guarantees that the analysis results remain nearly the same whether or not any person's data is present, while protecting against reconstruction and linkage attacks (Bhattacharjee et al. 2020).

The following section explores how these approaches and techniques can be explicitly applied to data visualizations in epidemiology.

Selected Visualizations in Epidemiology

Epidemiology relies on a wide range of visuals to explore, analyze, and communicate patterns, causes, and effects of health conditions in populations. Commonly used visuals include bar charts, histograms, scatter plots, heat maps, geographic maps, box plots, time series, network visualizations, and hierarchical visualizations (Archana, Hegadi, and Manjunath 2018; Carroll et al. 2014; Kim et al. 2024). The following subsections outline the role of each visualization type, associated privacy risks, and common anonymization approaches. Table 1 provides a comparative summary.

TABLE 1
PRIVACY RISKS AND ANONYMIZATION TECHNIQUES ACROSS COMMON EPIDEMIOLOGICAL VISUALIZATION

Visualization	Privacy Risks	Common Anonymization Techniques / Approaches
Bar Charts	● Small bar counts ● Rare categories ● Few visible groups	● Aggregation (merge categories) ● Generalization (broader ranges, e.g., 10-year age groups) ● Suppression (hide low-count bars) ● Perturbation (noise in bar heights)
Histograms	● Low-frequency bins ● Inappropriate binning (too narrow/many)	● Suppression (drop low-count bins) ● Generalization (broader bins) ● Probabilistic anonymization (calibrated noise) ● Deterministic anonymization (centroid replacement) ● Differential privacy (PRIVHIST, NoiseFirst, etc.)
Scatterplots	● Exact coordinates ● Sensitive cluster edges	● Grid-based generalization ● Suppression of low-count cells ● Deterministic anonymization ● Probabilistic anonymization (e.g., DAWA and Geometric Truncated)
Heatmaps	● Low-count cells ● Density hotspots	● Suppression (thresholding, *k*-anonymity) ● Generalization (larger grids) ● Probabilistic anonymization (noise in cell counts) ● Deterministic anonymization (KNN)
Geographic Maps	● Inference from precise geolocation ● Re-identification via fine-grained coordinates	● Generalization/Suppression (spatial *k*-anonymity) ● Geographic masking (aggregation, perturbation, circular/variable-radius/weighted/donut masks) ● Blurring (pixel averaging) ● Pseudonymization (address coding) ● Synthetic data
Boxplots	● Outliers ● Extreme whisker values	● Suppression (low-frequency values) ● Generalization (coarser grids) ● Deterministic anonymization ● Probabilistic anonymization (variance-based noise) ● Differential privacy (Laplace mechanism, ϵ-control)

Time Series	• Unique temporal patterns • Rare events • Correlations across intervals • Re-identification via N-grams	• Aggregation (broader time intervals) • Microaggregation (grouping similar time series) • Cryptographic methods (secret sharing, garbled circuits) • Perturbation (compression, geometric, noise) • *k*-anonymity for N-grams • Synthetic data
Network Visualizations	• Sensitive nodes or relationships • Unique structural positions	• Aggregation (node merging) • Generalization (edge bundling) • Suppression (node/edge deletion)
Hierarchical Visualizations (Treemaps)	• Multivariate relationships • Rare/small categories	• Aggregation (merge small categories) • Generalization (broader ranges, coarsened colors) • Suppression (hide small nodes) • Perturbation (noise in multivariate connections/encoding)

Bar Charts

A bar chart is commonly used to display disease incidence or counts; however, privacy risks can arise if certain patterns stand out, such as bars with small counts or when only a few categories are present (Bhattacharjee et al.). Data aggregation summarizes individual data points in bar charts into broader groups or categories to reduce the visibility of rare cases (Dasgupta et al. 2014). Generalization widens category ranges, such as reporting ages in 10-year intervals instead of 5-year steps. This helps eliminate or reduce categories, making them less inferable in a breach. If low-frequency categories remain after aggregation and generalization, they can be concealed through suppression, which hides categories or bins below a minimum threshold. Perturbation offers another option, which is adding controlled noise to bar heights. However, excessive noise can reduce the perceptual accuracy of bar charts, especially for tasks that involve retrieving specific values or comparing ranges (Zhang et al. 2020). Structured approaches such as *k*-anonymity, *l*-diversity, and *t*-closeness combine aggregation, generalization, suppression, and sometimes perturbation to limit re-identification.

Histograms

Histograms group observations into ranges through binning, reducing granularity, but can also introduce privacy risks if bins contain low counts or bin width and number are poorly chosen. Low-frequency bins should be suppressed to prevent disclosure, while generalization (e.g., increasing bin ranges) reduces the number of bins and strengthens privacy (Avraam et al. 2021). Deterministic and probabilistic anonymization can further enhance privacy in histograms. Probabilistic methods add calibrated noise, obscuring actual values while preserving distributions. Deterministic methods replace individual values with centroids of nearby neighbors before binning (Avraam et al. 2021). More advanced approaches, such as PRIVHIST and related algorithms, apply the principle of differential privacy by calibrating noise to each bin based on a privacy parameter (ϵ) (Ghazi et al. 2022; Suresh 2019; Wang et al. 2024). On the other hand, adding excessive noise can significantly distort distributions, especially for small bins, underscoring the need to combine suppression, generalization, and controlled perturbation for adequate privacy protection in histograms (Chen et al. 2024).

Scatterplots

Scatterplots visualize relationships between two continuous variables but pose privacy risks because they reveal exact coordinate pairs. Suppression and grid-based generalization can mitigate this risk by grouping data into coarser grids and suppressing low-count cells (Avraam et al. 2022; Panavas et al. 2023). However, excessive generalization or suppression may distort dispersion, and cluster edges may still expose sensitive points if attackers have background knowledge (Dasgupta et al. 2014). Deterministic anonymization preserves statistical properties while masking exact coordinates, whereas probabilistic anonymization introduces controlled noise to obscure boundaries such as convex hulls in bounded distributions (Avraam et al. 2021). Algorithms such as DAWA and Geometric Truncated calibrate the appropriate noise level for scatterplot privacy (Panavas et al.,).

Heatmaps

Heatmaps display the density of values using colored grids, where darker colors represent higher counts. Heatmaps can compromise privacy if low-density cells or unique patterns make it possible to single out individuals. Suppression is commonly applied by filtering out grid cells with fewer than a minimum threshold (e.g., k = 5 users per cell) to enforce k-anonymity (Oksanen et al., 2015; Sainio, Westerholm, and Oksanen. Generalization can complement suppression by enlarging grid sizes (e.g., merging cells into broader 30×30 grids) to reduce granularity before removing low-count cells (Avraam et al. 2021). Additionally, probabilistic anonymization introduces controlled random noise (e.g., 6.25% variance) to further obscure cell values. Deterministic approaches apply KNN (e.g., k = 3) to preserve overall density patterns while reducing disclosure risk (Avraam et al. 2021).

Geographic Maps

Geographic maps represent spatial information, ranging from static to dynamic 2D/3D displays, as well as choropleth maps. Because geolocation data reveal sensitive details, such as home, school, and work addresses, as well as movement histories, they pose significant privacy risks. During the COVID-19 pandemic, location data were used for contact tracing, highlighting risks associated with inferring sensitive health information (Iyer et al.). Several privacy-preserving techniques can mitigate these risks.

Spatial *k*-anonymity generalizes or suppresses precise geolocation data so that each individual is indistinguishable from at least *k–1* others (Iyer et al. 2021). For instance, a larger area containing multiple individuals is shown instead of exact coordinates. Geographic masking obscures precise locations through either aggregation (replacing exact coordinates with broader areas, such as streets or districts) or perturbation (shifting coordinates within a controlled radius) (Armstrong et al.). Variants include circular, variable-radius, weighted, or donut masks, which balance privacy and spatial accuracy depending on the perturbation radius (Jeffery, Ozonoff, and Pagano 2014; Kwan et al. 2004). Blurring provides another option by averaging pixel values with those of neighboring pixels, allowing specific locations to blend into their surrounding areas. Pseudonymization replaces identifiers, such as addresses, with unique codes, allowing clinical records to be linked without revealing exact locations (Hampton et al.). Finally, when detailed demographic or location data are unavailable, synthetic datasets can be used to mimic real-world spatial patterns while preserving clustering and heterogeneity for epidemic modeling (Tildesley and Ryan).

Boxplots

Boxplots summarize data distributions through five-number summaries (minimum, first quartile, median, third quartile, and maximum) and highlight skewness and outliers. Privacy risks primarily arise from outliers and extreme values at the whiskers, which can facilitate re-identification (Avraam et al.). Suppression that hides low-frequency values (e.g., counts <3) mitigates outlier risks but may distort ranges. Generalization broadens value ranges (e.g., using coarser bins) to make extreme points less identifiable. Deterministic anonymization (e.g., KNN) and probabilistic anonymization (adding slight variance-based noise) preserve the overall distribution without excluding outliers, as new outliers differ from the originals (Avraam et al.). Lastly, differential privacy applies noise to summary statistics (e.g., quartiles, median), often via the Laplace mechanism, with privacy parameter ϵ controlling the trade-off between privacy and

accuracy (Ramsay and Diaz-Rodriguez 2024). This parameter is a crucial consideration, as excessive noise can stretch whiskers and distort the ranges of boxplots (Avraam et al., 2021; Ramsay and Diaz-Rodriguez).

Time Series

Time series data consist of sequences of information collected at consistent intervals to reveal trends and patterns. Privacy concerns in time series data arise from unique temporal patterns, rare events, or correlations across intervals that can be linked back to individuals. Aggregation summarizes data points over time into broader intervals to reduce identifiability. Microaggregation extends this by grouping similar time series and replacing individual values with group statistics such as means or medians (Thouvenot, Nogues, and Gouttas 2017). Cryptographic techniques, such as additive secret sharing and garbled circuits, enable privacy-preserving computation of aggregated statistics by ensuring that individual time series data remain concealed throughout the analysis. In additive secret sharing, data are split into random shares that only reconstruct meaningful values when combined.

In contrast, garbled circuits allow multiple parties to jointly compute functions (e.g., averages) without disclosing their inputs (Liu et al. 2020). Additionally, data perturbation is commonly employed in time series data to modify data, obscuring sensitive values while preserving trends. Techniques include compression-based perturbation (reducing dimensionality), geometric transformations (such as rotation, translation, and scaling), and noise addition. Gaussian noise can be added to raw data before aggregation, or Laplacian noise can be applied to the aggregated values (Hong et al., 2013; Imtiaz et al.). Lastly, *k*-anonymity can be adapted to time series using N-grams, which are repeating symbol sequences representing the data. The sequences are adjusted to have at least *k* occurrences, preventing unique traces from being re-identified (Zare-Mirakabad, Kaveh-Yazdy, and Tahmasebi 2013). However, overgeneralization risks distorting temporal patterns. Synthetic data offers an alternative, where time series records (e.g., electrocardiograms) are generated based on KNN in the frequency domain, preserving statistical properties while protecting privacy (Bennis and Gourraud 2021).

Network Visualizations

Networks represent entities (nodes) and their relationships (edges), making them powerful for understanding movement, contact patterns, and correlations in disease spread. Epidemiologists can utilize networks to identify highly connected individuals or groups (key nodes) that are central to transmission and target interventions accordingly (Christakis and Fowler). However, displaying sensitive attributes within nodes or edges introduces privacy risks, as individual identities or relationships can be inferred.

Several network anonymization techniques have been proposed, often combining generalization, aggregation, suppression, and data masking (Hay et al. 2007). Node merging combines multiple nodes into a "super-node," aggregating their connections while obscuring intra-node relationships. Edge bundling reduces clutter and hides specific endpoints by grouping edges between node clusters, limiting disclosure of sensitive links. Node and edge deletion can address isolated privacy leaks by removing nodes or edges that uniquely expose sensitive information, though at the cost of data loss (Chou, Bryan, and Ma 2017; Zheleva and Getoor 2007).

Hierarchical Visualizations

Hierarchical visualizations, such as treemaps, represent nested relationships between attributes (e.g., age, gender, and medical conditions) (Hugine, Guerlain, and Turrentine 2014; Scheibel et al. 2020). Compared to simpler charts, such as bar charts or histograms, immediate disclosure risks are lower because treemaps inherently obscure some details, as categories with minimal areas may not be labeled or clearly visible (Archana et al., 2018; Dasgupta et al.). Although treemaps obscure some details, their multivariate structure can still reveal subtle patterns that attackers might exploit. Aggregation can summarize connecting information or merge smaller categories into broader groups to mitigate these risks. Additionally, generalization can potentially adjust category ranges or coarsen color gradients to obscure fine-grained distinctions.

In contrast, suppression can hide or omit nodes/rectangles that fall below a threshold area to prevent exposure of rare categories (Avraam et al. 2021). Adjusting visual encodings (e.g., line thickness, color intensity) can also reduce the risk of sensitive attribute disclosure (Hugine et al. 2014). Additional techniques, such as adding controlled noise to multivariate connections, can further reduce the risk of sensitive attribute disclosure (Avraam et al. 2021; Hugine et al. 2014).

CONCLUSION & FUTURE RESEARCH

Epidemiology increasingly relies on data visualizations to interpret and communicate complex health information, but these tools pose significant privacy risks if proper anonymization techniques and strategies are not employed. This review presents one of the first structured syntheses of how privacy-preserving techniques are applied to specific visualization types in epidemiology, highlighting both their risks and the most effective anonymization strategies. Each visualization type serves a unique purpose and requires careful consideration, often involving a combination of privacy-preserving techniques. Our findings provide a practical resource for researchers and practitioners seeking to design visualizations that strike a balance between confidentiality and interpretability, with clear implications for organizational risk management and policy compliance.

Building on this foundation, future research should test these techniques in applied contexts, evaluate compliance with regulatory frameworks such as HIPAA and GDPR, and explore the integration of multiple anonymization techniques within a single visualization. Developing standardized evaluation metrics and automated toolkits could also help operationalize privacy-preserving visualization in epidemiological practice. Ultimately, engaging healthcare professionals and end-users through surveys and experiments will ensure that privacy-preserving visualizations are both protective and practically useful in guiding clinical and public health decisions.

REFERENCES

Aamot, H., Kohl, C.D., Richter, D., & Knaup-Gregori, P. (2013). Pseudonymization of patient identifiers for translational research. *BMC Medical Informatics and Decision Making*, *13*(1), 75.

Almalki, F.A., & Soufiene, B.O. (2021). EPPDA: An efficient and privacy-preserving data aggregation scheme with authentication and authorization for IoT-based healthcare applications. *Wireless Communications and Mobile Computing*, (1), 5594159.

Archana, R.A., Hegadi, R.S., & Manjunath, T.N. (2018). A study on big data privacy protection models using data masking methods. *International Journal of Electrical and Computer Engineering*, *8*(5), 3976.

Armstrong, M.P., Rushton, G., & Zimmerman, D.L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525.

Avraam, D., Jones, E., & Burton, P. (2022). A deterministic approach for protecting privacy in sensitive personal data. *BMC Medical Informatics and Decision Making*, *22*(1), 24.

Avraam, D., Wilson, R., Butters, O., Burton, T., Nicolaides, C., Jones, E., Boyd, A., & Burton, P. (2021). Privacy preserving data visualizations. *EPJ Data Science*, *10*(1), 2.

Benaim, A.R., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., … Khoury, J. (2020). Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. *JMIR Medical Informatics*, *8*(2), e16492.

Bennis, Z., & Gourraud, P.-A. (2021). Application of a novel anonymization method for electrocardiogram data. *[Conference paper]*, 1–5.

Bhattacharjee, K., Chen, M., & Dasgupta, A. (2020). Privacy-preserving data visualization. *Reflections on the state of the art and research opportunities*, *39*, 675–692. Wiley Online Library.

Boussif, M., Aloui, N., & Cherif, A. (2018). Secured cloud computing for medical data based on watermarking and encryption. *IET Networks*, *7*(5), 294–298.

Carroll, L.N., Au, A.P., Detwiler, L.T., Fu, T.-C., Painter, I.S., & Abernethy, N.F. (2014). Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics*, *51*, 287–298.

Chen, J., Zhou, S., Qiu, J., Xu, Y., Zeng, B., Fang, W., … Chen, Y. (2024). A histogram publishing method under differential privacy that involves balancing small-bin availability first. *Algorithms*, *17*(7), 293.

Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497.

Chen, T.-J., Chuang, K.-S., Chang, J.-H., Shiao, Y.-H., & Chuang, C.-C. (2006). A blurring index for medical images. *Journal of Digital Imaging*, *19*(2), 118–125.

Chishtie, J., Bielska, I.A., Barrera, A., Marchand, J.-S., Imran, M., Tirmizi, S. F.A., . . . Senthinathan, A. (2022). Interactive visualization applications in population health and health services research: Systematic scoping review. *Journal of Medical Internet Research*, *24*(2), e27534.

Chou, J.-K., Bryan, C., & Ma, K.-L. (2017). Privacy preserving visualization for social network data with ontology information. *[Conference paper]*, 11–20. IEEE.

Christakis, N.A., & Fowler, J.H. (2009). Social network visualization in epidemiology. *Norsk Epidemiologi (Norwegian Journal of Epidemiology)*, *19*(1), 5.

Cox, L.H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, *75*(370), 377–385.

Dalenius, T., & Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, *6*(1), 73–85.

Dasgupta, A., Abdul-Rahman, A., Chen, M., & Maguire, E. (2014). Opportunities and challenges for privacy-preserving visualization of electronic health record data. *[Conference paper]*.

Elanshekhar, N., & Shedge, R. (2017). An effective anonymization technique of big data using suppression slicing method. *[Conference paper]*, 2500–2504. IEEE.

Fienberg, S.E., & McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In *Privacy in Statistical Databases* (pp. 14–29). Springer.

Frérot, M., Lefebvre, A., Aho, S., Callier, P., Astruc, K., & Glélé, L.S.A. (2018). What is epidemiology? Changing definitions of epidemiology 1978–2017. *PLOS ONE*, *13*(12), e0208442.

Ghazi, B., Kamath, P., Kumar, R., & Manurangsi, P. (2022). Anonymized histograms in intermediate privacy models. *Advances in Neural Information Processing Systems*, *35*, 8456–8468.

Gunawan, D., Nugroho, Y.S., Al Irsyadi, F.Y., Utomo, I.C., Andreansyah, I., & Islam, S. (2022). ℓρ-suppression: A privacy preserving data anonymization method for customer transaction data publishing. *[Conference paper]*, 171–176. IEEE.

Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A., … Miller, W.C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069.

Hay, M., Miklau, G., Jensen, D., Weis, P., & Srivastava, S. (2007). Anonymizing social networks. *[Conference paper]*.

He, W., Liu, X., Nguyen, H., Nahrstedt, K., & Abdelzaher, T. (2007). PDA: Privacy-preserving data aggregation in wireless sensor networks. *[Conference paper]*, 2045–2053. IEEE.

Hong, S.-K., Gurjar, K., Kim, H.-S., & Moon, Y.-S. (2013). A survey on privacy preserving time-series data mining. *[Conference paper]*, 44–48.

Hugine, A.L., Guerlain, S.A., & Turrentine, F.E. (2014). Visualizing surgical quality data with treemaps. *Journal of Surgical Research*, *191*(1), 74–83.

Imtiaz, S., Horchidan, S.-F., Abbas, Z., Arsalan, M., Chaudhry, H.N., & Vlassov, V. (2020). Privacy preserving time-series forecasting of user health data streams. *[Conference Paper]*, 3428–3437. IEEE.

Iyer, R., Rex, R., McPherson, K.P., Gandhi, D., Mahindra, A., Singh, A., & Raskar, R. (2021). Spatial k-anonymity: A privacy-preserving method for COVID-19 related geospatial technologies. *arXiv Preprint arXiv:2101.02556*.

Jeffery, C., Ozonoff, A., & Pagano, M. (2014). The effect of spatial aggregation on performance when mapping a risk of disease. *International Journal of Health Geographics*, *13*(1), 9.
Jiang, J., Skalli, W., Siadat, A., & Gajny, L. (2022). Effect of face blurring on human pose estimation: Ensuring subject privacy for medical and occupational health applications. *Sensors*, *22*(23), 9376.
Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. *IEEE Access*, *7*, 61656–61669.
Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., … Weller, A. (2022). Synthetic data—What, why and how? *arXiv Preprint arXiv:2205.03257*.
Kester, Q.-A., Nana, L., Pascu, A.C., Gire, S., Eghan, J.M., & Quaynor, N.N. (2015). A cryptographic technique for security of medical images in health information systems. *Procedia Computer Science*, *58*, 538–543.
Kim, D., Cánovas-Segura, B., Campos, M., & Juarez, J.M. (2024). Visualization of spatial–temporal epidemiological data: A scoping review. *Technologies*, *12*(3), 31.
Kwan, M.-P., Casas, I., & Schmitz, B. (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*(2), 15–28.
Lau, B., Duggal, P., Ehrhardt, S., Armenian, H., Branas, C.C., Colditz, G.A., … Hofman, A. (2020). Perspectives on the future of epidemiology: A framework for training. *American Journal of Epidemiology*, *189*(7), 634–639.
Li, N., Li, T., & Venkatasubramanian, S. (2006). T-closeness: Privacy beyond k-anonymity and l-diversity. *[Conference paper]*, 106–115. IEEE.
Liu, X., Zheng, Y., Yi, X., & Nepal, S. (2020). Privacy-preserving collaborative analytics on medical time series data. *IEEE Transactions on Dependable and Secure Computing*, *19*(3), 1687–1702.
Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3–es.
Midway, S.R. (2020). Principles of effective data visualization. *Patterns, 1*(9).
Oksanen, J., Bergman, C., Sainio, J., & Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography*, *48*, 135–144.
Panavas, L., Crnovrsanin, T., Adams, J.L., Ullman, J., Sargavad, A., Tory, M., & Dunne, C. (2023). Investigating the visual utility of differentially private scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, *30*(8), 5370–5385.
Park, S., Bekemeier, B., Flaxman, A., & Schultz, M. (2022). Impact of data visualization on decision-making and its implications for public health practice: A systematic literature review. *Informatics for Health and Social Care*, *47*(2), 175–193.
Pfitzmann, A., & Hansen, M. (2008). *Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management—a consolidated proposal for terminology* (Version V0.31, p. 15).
Rahman, M., Paul, M.K., & Sattar, A.H.M.S. (2023). Efficient perturbation techniques for preserving privacy of multivariate sensitive data. *Array*, *20*, 100324.
Rao, R.M., Krishna, S.M., & Kumar, A.P.S. (2018). Privacy preservation techniques in big data analytics: A survey. *Journal of Big Data*, *5*(1), 33.
Ramsay, K., & Diaz-Rodriguez, J. (2024). Differentially private boxplots. *arXiv Preprint arXiv:2405.20415*.
Reiss, S.P. (1984). Practical data-swapping: The first steps. *ACM Transactions on Database Systems (TODS)*, *9*(1), 20–37.
Riedl, B., Grascher, V., Fenz, S., & Neubauer, T. (2008). Pseudonymization for improving the privacy in e-health applications. In *IEEE Conference Proceedings* (pp. 255–255). IEEE.
Riedl, B., Neubauer, T., Goluch, G., Boehm, O., Reinauer, G., & Krumboeck, A. (2007). A secure architecture for the pseudonymization of medical data. In *IEEE Conference Proceedings* (pp. 318–324). IEEE.

Roman, A.-S. (2023). Evaluating the privacy and utility of time-series data perturbation algorithms. *Mathematics*, *11*(5), 1260.
Sainio, J., Westerholm, J., & Oksanen, J. (2015). Generating heat maps of popular routes online from massive mobile sports tracking application data in milliseconds while respecting privacy. *ISPRS International Journal of Geo-Information*, *4*(4), 1813–1826.
Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information. In *Proceedings of the IEEE Symposium*, *98*, 10–1145.
Scheibel, W., Trapp, M., Limberger, D., & Döllner, J. (2020). A taxonomy of treemap visualization techniques. In *Conference Proceedings* (pp. 273–280).
Suresh, A.T. (2019). Differentially private anonymized histograms. *Advances in Neural Information Processing Systems*, *32*.
Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570.
Thouvenot, V., Nogues, D., & Gouttas, C. (2017). Data-driven anonymization process applied to time series. In *Conference Proceedings* (pp. 80–90).
Tildesley, M.J., & Ryan, S.J. (2012). Disease prevention versus data privacy: Using landcover maps to inform spatial epidemic models. *PLoS Computational Biology*, *8*(11), e1002723.
Turgay, S., & İlter, İ. (2023). Perturbation methods for protecting data privacy: A review of techniques and applications. *Automation and Machine Learning*, *4*(2), 31–41.
Unwin, A. (2020). Why is data visualization important? What is important in data visualization. *Harvard Data Science Review*, *2*(1), 1.
Vishwamitra, N., Knijnenburg, B., Hu, H., & Caine, Y.P.K. (2017). Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Conference Proceedings* (pp. 39–47).
Wang, X., Mo, L., Zheng, X., & Dang, Z. (2024). Streaming histogram publication over weighted sliding windows under differential privacy. *Tsinghua Science and Technology*, *29*(6), 1674–1693.
Wilson, R.L., & Rosen, P.A. (2005). Does protecting databases using perturbation techniques impact knowledge discovery? In *Advanced Topics in Database Research*, *4*, 96–107. IGI Global.
Xu, Y., Ma, T., Tang, M., & Tian, W. (2014). A survey of privacy preserving data publishing using generalization and suppression. *Applied Mathematics & Information Sciences*, *8*(3), 1103.
Yaseen, S., Abbas, S.M.A., Anjum, A., Saba, T., Khan, A., Malik, S.U.R., … Bashir, A.K. (2018). Improved generalization for secure data publishing. *IEEE Access*, *6*, 27156–27165.
Zandbergen, P.A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*, *2014*(1), 567049.
Zare-Mirakabad, M.-R., Kaveh-Yazdy, F., & Tahmasebi, M. (2013). Privacy preservation by k-anonymizing n-grams of time series. In *IEEE Conference Proceedings* (pp. 1–6). IEEE.
Zhang, D., Sarvghad, A., & Miklau, G. (2020). Investigating visual analysis of differentially private data. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1786–1796.
Zheleva, E., & Getoor, L. (2007). Preserving the privacy of sensitive relationships in graph data. In *Springer Conference Proceedings* (pp. 153–171). Springer.