Health Information Data Breaches

Angeline Teo & Ritu Shah

Northeastern University

Abstract

This paper describes the use of a visualization tool to enhance healthcare data security and protect patient data. This paper highlights two domain tasks for a healthcare organization, wherein their end goals consist of conducting a risk assessment and developing a data security strategy. The tool enables healthcare organizations to identify connections with organizations that have experienced data breaches, which can help them assess potential risks and develop targeted data security strategies. By using the tool to conduct risk assessments and develop data security strategies, healthcare organizations can comply with privacy regulations, reduce reputational damage, and increase trust with their patients. In addition, the tool can also be useful for healthcare users who are choosing a healthcare entity, as it allows them to view entities that have been affected by data breaches or information exposure, and make an informed decision about where to entrust their data.

1 Introduction

Healthcare organizations face significant challenges in safeguarding patient data due to the constantly evolving threat of cyberattacks. A data breach could not only result in reputational damage and legal consequences, but also cause harm to patients. Healthcare organizations need to conduct regular risk assessments to identify potential vulnerabilities and threats to patient data security, and develop data security strategies to mitigate those risks. The use of a visualization tool may be effective for healthcare organizations to enhance their security systems and safeguard patient data.This visualization tool will showcase health data breaches reported by the U.S. Department of Health and Human Services that are currently being investigated by the Office for Civil Rights. Attributes that will be visible are the name of the entity, the state the breach occurred, the type of covered entity, the number of affected individuals, type of breach, location of breach information, and the date of breach. This can be used to identify connections with organizations that have experienced data breaches, allowing healthcare organizations to conduct more targeted risk assessments and develop more effective data security strategies. The visualization tool can also assist healthcare users in selecting a healthcare entity, enabling them to make informed decisions about whether or not to entrust their data to an organization.

This visualization will support two domain tasks. The first is conducting a risk assessment. Healthcare organizations need to conduct regular risk assessments to identify potential vulnerabilities and threats to patient data security. By using the visualization tool to identify connections with organizations that have experienced data breaches, the healthcare organization can conduct a more targeted risk assessment that focuses on specific areas of concern. This task requires specialized knowledge of healthcare data security and risk management principles, as well as the ability to analyze data and identify potential risks. Another domain task is developing a data security strategy. Once potential risks have been identified through the risk assessment, the healthcare organization needs to develop a data security strategy to mitigate those risks. This task requires a deep understanding of healthcare data security regulations and best practices, as well as the ability to develop and implement effective security measures. By using the visualization tool to identify connections with organizations that have experienced data breaches, the healthcare organization can develop a more targeted data security strategy that addresses specific areas of concern.

2 Related Work

A group of researchers wanted to examine the transition from paper health records to electric health records and the increase in data breaches [1]. They used public data to examine the nature and extent of data breaches from 2010 to 2017. The visualizations presented in this paper will be beneficial as a resource since similar data is used within our chosen dataset and cover various aspects of the data. In Figure 1, a line plot is created that shows the year vs. the number of breaches. There are three different lines that represent a HIPAA entity type which is shown in their own distinct colors. For our visualization, we could create a line plot that shows the covered entity type with the associated number of breaches and distinguish them with different colors. In Figure 2, a line plot is created that shows the year vs. the number of breaches. There are six different lines to represent the media location and breach data. This could be helpful for our dataset since we also work with similar data.

Researchers discusses how although digital healthcare services have made it easier and more accessible for patients to view their records and receive treatment, the present-day healthcare industry has been a prominent victim of internal and external breach attacks [2]. They performed an in-depth analysis of healthcare data breaches and drew inferences to improve healthcare data confidentiality. A visualization in this paper could be advantageous as a resource because of the way they incorporated channels to better envision their data. In Figure 5, the visualization compares the year vs. the number of associated data breaches using a grouped bar chart. Each bar for a certain year represents the type of breach location but is distinguished by its own distinct color. This concept allows readers to clearly see which breach was more popular to take place during a certain year or time period. For example, in 2019, a data breach via email was the most common breach while in 2010, it was one of the least common breaches. This shows the impact of how the increase in technology use has made companies and digital health platforms more at risk of data breaches.

3 Use Case

A healthcare organization is taking proactive steps to enhance their security system and safeguard patient data. They recognize that a data breach could not only result in reputational damage and legal consequences, but also cause harm to their patients. They plan to use a visualization tool to identify any connections they have with organizations that have experienced data breaches. By doing so, they can assess the potential impact on their own systems and procedures, and develop strategies to mitigate any risks. This would not only help them to comply with privacy regulations but also increase trust with their patients.

In addition to being a useful tool for healthcare organizations, the visualization tool could also be relevant to healthcare users who are choosing a healthcare entity. They could use the tool to view the entities that have been affected by data breaches or information exposure. By seeing this information, users could make an informed decision about whether they feel comfortable entrusting their data to that organization or whether they should consider an alternative option.

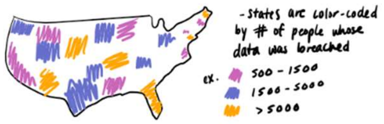
4 Data

The data was collected by the U.S. Department of Health and Human Services that are currently being investigated by the Office for Civil Rights. The original dataset can be found in this link: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. The U.S. Department of Health and Human Services is a sector of the United States government that promotes the well-being and health of all Americans by supporting services that emphasize the science behind medicine, public health, and social services. The data demonstrates a list of data breaches reported within the last 24 months that have affected 500 or more individuals with unsecured health information. It shows the name of the covered entity, also known as the organization or company that was affected by a data breach, the type of covered entity, and the type of breach. Additionally, the state, the date of the submitted breach, and the number of individuals affected are shown.

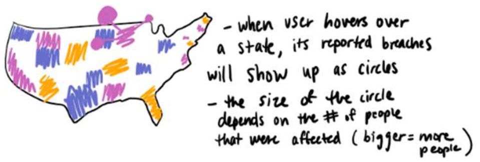
Since this data is directly submitted to the U.S. government, there should be no bias as it is information that should be neutralized without any preference. The data is covered from all places around the United States and does not only show one part of the country. The only aspect that is disadvantageous is the option of "Other" which is included in the "Type of Breach" column. This option does not fully specify what type of breach was committed. This information should be fully open about every type of data breach that exists to create more awareness among users of healthcare and allow readers of the data to see if a specific type was more common compared to other breaches or if it would have an impact on the statistics of the data.

The data doesn't have any missing or unexpected values. There are no outliers, and the data contains all the attributes we want to visualize. However, the values under "Name of Covered Entity" in the data are messy, containing characters that do not make logical sense, extraneous information, and inconsistent capitalization and formatting (ex. "doing business as" is written as "d/b/a", "dba", "doing business as"). In terms of data cleaning, there was an additional column included called "Web Description" that had no data, so this column was removed since it was not essential to our visualization. There were also white spaces towards the end of certain words which were eventually removed. All the data had consistent wording and spelling and did not require any change. There were no new, derived attributes added to the data.

5 Design Process

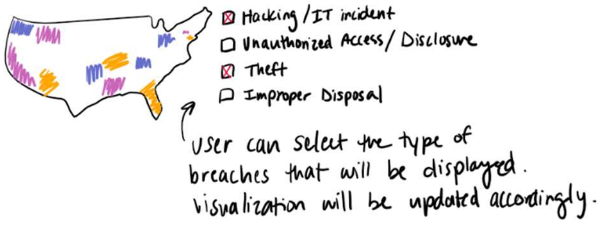


- states are color-coded by # of people whose data was breached



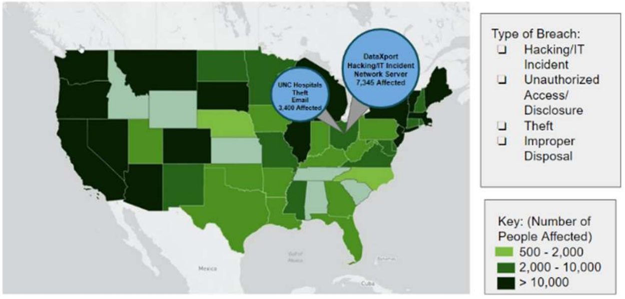
- when user hovers over a state, its reported breaches will show up as circles

- the size of the circle depends on the # of people that were affected (bigger = more people)



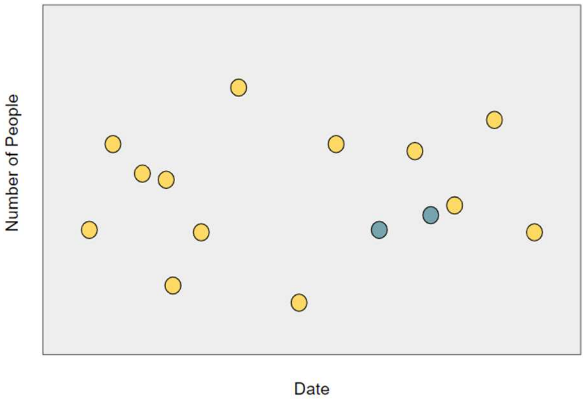
☒ Hacking/IT incident
□ Unauthorized Access/Disclosure
☒ Theft
□ Improper Disposal

user can select the type of breaches that will be displayed. visualization will be updated accordingly.



Type of Breach:
□ Hacking/IT Incident
□ Unauthorized Access/ Disclosure
□ Theft
□ Improper Disposal

Key: (Number of People Affected)
🟩 500 - 2,000
🟩 2,000 - 10,000
■ > 10,000



Breaches will be displayed geographically on a U.S. map. The states will be color-coded using a gradient to distinguish how many people were affected by breaches (the darker the hue is, the more people were affected, and vice versa). When a user hovers over a state, the state will highlight and its individual breaches will pop-up and enlarge as circles with pointers to the state. Each circle will contain more details about the breach, such as the organization, the type of breach, what was breached, and the number of people affected. The size of these circles will reflect how many people were affected (the bigger the circle is, the more people were affected, and vice versa). The user can also select the type of breach they want to see, and the visualization will update accordingly.

A separate view will display a scatter plot plotting the date of breach and the number of people affected. When a user hovers over a state, the points on the scatter plot that coordinate with the breaches present in that state will change color.

6 Final Design

7 Discussion

8 Conclusion

References

[1] McCoy, T. H., Jr, & Perlis, R. H. (2018). Temporal trends and characteristics of reportable health data breaches, 2010-2017. *JAMA*,*320*(12),1282–1284. https://doi.org/10.1001/jama.2018.9222

[2] Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., & Ahmad Khan, R. (2020). Healthcare data breaches: Insights and implications. *Healthcare*, *8*(2), 133. https://doi.org/10.3390/healthcare8020133

Appendix: Data Abstraction

Each row in the dataset represents an item. The column “Name of Covered Entity” is a categorical attribute. The column “State” is a categorical attribute. The column “Covered Entity Type” is a categorical attribute. The column “Individuals Affected” is a quantitative attribute. The column “Breach Submission” is an ordered attribute. The column “Type of Breach” is a categorical attribute. The column “Location of Breach” is a categorical attribute. The column “Business Associate Present” is a categorical attribute.

Appendix: Task Abstraction

- Conducting a risk assessment:
    - High-level: Discover; the visualization will need to be explored by the user to glean any valuable insights.
    - Medium-level: Browse; the location is already known (where the end user is located), but the target is unknown (are there any data breaches in the specified location?)
    - Low-level: Identify; to help the end user determine whether or not to conduct a risk assessment, the visualization will provide information about data breaches in the target location. If there are data breaches connected to the organization, steps to develop a data security strategy may be pursued.
- Developing a data security strategy:
    - High-level: Discover; the visualization will need to be explored by the user to glean any valuable insights.
    - Medium-level: Browse; the location is already known (where the end user is located), but the target is unknown (are there any data breaches in the specified location?)
    - Low-level: Summarize; the visualization will provide information about data breaches by state and its associated information (organization that was breached, what kind of information was breached, when the data was breached) to help the end user develop a data security regulations and practices that are equipped to handle information that has been shown to be at higher risk.