CSCE 5360

Project Group #44

Rimma Shilkina

Final Project

Visualization of Healthcare Data Breaches in the United States and European Union to find it correlation with Legal Regulations

*Introduction*

Cybersecurity is one of the priorities for healthcare industry. Since 2009 more than 2100 healthcare data breaches were reported, and 2021 became the year with the most healthcare data breaches in the United States. Here is some related statistics:

* By the end of 2020, security breaches cost $6 trillion dollars for healthcare companies.
* Nearly 80 million people were affected by the Anthem Breach (January 2015).
* In 2021, data breaches in healthcare cost businesses an average of $9.3 million per incident. That’s a 29.5% rise compared to 2020. (Source: CompliancyGroup)
* 95% of identity theft comes from stolen healthcare records. (Source: Globe NewsWire)
* Breaches due to negligence happen twice as often as targeted ones. (Source: JOCS Vol. 2 Iss. 1)
* The healthcare industry invests less than 6% of its budget on cybersecurity. The US spends 16% of its federal budget on cybersecurity. (Source: Healthcare IT News)1

These numbers show that from one side, there is a significant increase in healthcare data breaches, but, from the other side, it is obvious that companies are not very keen to increase their cybersecurity expenses and take more actions to protect private information of their clients. One of possible reasons is absence of the “strong”, unified, and effective legal regulations protecting personal data in the USA. There is no comprehensive federal law in the United States. The most important legislation is HIPAA - “The Health Insurance Portability and Accountability Act” of 1996. It is a federal law that requires creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge.” 2 It includes 3 Rules: The Privacy Rule, The Security Rule, and The Breach Notification Rule. Many companies struggle to interpret HIPAA and comply with it.

As opposite, GDPR - The General Data Protection Regulation - is the personal data protection and privacy legislation for the European Union and the European Economic Area. It is recognized as the most strict and protective law related to data privacy. The United States does not have such broad, detailed, effective regulation on the federal level. In 2018 the State of California passed the California Consumer Privacy Act (CCPA) which is an analogue to GDPR.

*Goal of the Project*

The goal of this project is to show trends in healthcare data breaches by using data visualization tool(s) not only in the United States but also in European Union to highlight the correlation between number of healthcare data breaches, their severity, and legal regulations. In other words, does the existence of GDPR affect the situation with data breaches in healthcare system, does the EU have less data breaches incidents/breached records?

*Datasets description.*

Two separated datasets (i.e., there is no any types of joins) were used for the Project.

For the first part of the project, related to the United States, the dataset of the U.S. Department of Health and Human Services Office for Civil Rights was used. It contains 909 records of known data breaches for the last two years which are currently under investigation. Also, it has archive data, that could be used to see trends/data in progress. Dataset includes seven attributes like state, numbers of individuals affected, when data breach was reported, type of breach, etc. Below is a screenshot of few rows of that dataset:

Table

Description automatically generated

For the second part, the dataset created by some researchers from MIT and C6 Bank was used. It contains world data for 2018 and 2019, but it is possible to extract European Union countries data. Information for that dataset was gathered from press reports, industry studies, and regulatory agencies’ reports.3 It contains 430 records with 8 attributes. Below is a screenshot of several rows from the dataset:

Table

Description automatically generated

*Implementation and Results*

For the first part of the Project – visualizing US data – it was interesting to see a US map with a number of individuals affected by state. Since some records had NULL for the state, a filter was applied to exclude NULL values. For better understanding the situation, top 10 states were shown. It was expected that Texas, California, New York, and Florida will be in the list (states with the biggest population), but Wisconsin is a second in the list. That was an interesting finding. Before searching on the Internet what has happened in Wisconsin, it is possible to see when it is happened. The graph shows that in July 2022 there was an event or events with more than 4 million records breached. Some results from Google search confirmed that Wisconsin mailing vendor, OneTouchPoint, was a subject of a ransomware attack and reported more than 1 million affected individuals, but some of the state’s healthcare provider clients self-reported data breach, including Aetna ACE Health Plan. Goodman Campbell Brain and Spine also suffered a major ransomware attack.

As Texas is number one in the list, it is interesting to find some numbers. Assuming that affected records/individuals were not duplicated through several data breaches, then every fourth person in Texas was affected by healthcare data breach (29,945,493 (Texas population in 2022) / 8,003,201 = 3.7).

It is interesting to see number of individuals affected by month not only for the state of Wisconsin, but also for the United States in general. From both graphs, July 2022 was the worst month with total more than 8 million records breached.

Graphical user interface, application

Description automatically generated

Another important characteristic of data breach is where breached records were located. Network servers and e-mails are location with the most affected records. This could indicate following reasons of breaches. For emails the most probable reason of data leakage is sending messages to wrong recipients (i.e., by mistake or misconfiguration). Network servers’ issues could be with improperly configured access controls for servers designed for data sharing between different organizations.

For the second part of Project, it is important to understand that GDPR is designed to protect data privacy for persons of European Union and European Economic Area, but it has global effect (i.e., a company in the US must comply with it if it is processing personal data of EU and EEC persons). It was a challenge to find a good dataset regarding GDPR data breaches. Probably, one of the reasons is, again, the GDPR itself, i.e., strict rules regarding data privacy, including statistics. But it is possible to gain some information from the MIT and C6 Bank dataset. First histogram shows that Europe has the biggest number of records breached, but Europe is not the region with biggest population. North America is the second by breaches but 4th in terms of population. Both regions have more developed legislation for data security and privacy protections.

Table

Description automatically generatedA picture containing chart

Description automatically generated

The histogram above shows all sectors of economy. After filter were applied, the Europe is disappeared from the histogram. Does not Europe have healthcare data breaches? The next graph shows sectors of Europe economy where data breaches occurred.

Graphical user interface, chart, bubble chart

Description automatically generated

Business and Technology are the biggest sectors with breached records. Most likely, when company reports data breach event in healthcare it reports it as in “Business” sector. Unfortunately, this dataset is not informative enough for the goal of the Project, i.e., it is impossible to compare healthcare data breaches. But it is possible to compare in general (across all sectors of economy). Also, GDPR regulates not only healthcare sector but protects data privacy in general. So, it is possible to use the second dataset to compare data breaches in the key EU countries, UK, and the USA for all sectors of economy. Important note, population of the USA significantly exceeds population of any country of EU and UK.

Chart, bubble chart

Description automatically generated

As example, population of the US in 2022 approximately 332.4 million while France population is nearly 65 million. An interesting fact from the last diagram is that France among other EU countries in this diagram has the biggest number of data breaches. Although Germany has the largest population in the EU, it has 169 times less breached records than France and 180 times less than the US (according to the dataset). The reason of such differences between France and Germany could be that although GDPR is legislation for all countries in EU and EU Economic Area, but enforcement and reporting is per country basis. France might be more open to provide statistical data than Germany.

*Conclusion*

It is a big challenge to find cybersecurity data (or data related to privacy) in general and for the European Union specifically. But even from the datasets used for this project it is possible to obtain some information, to see healthcare data breaches situation in the United States, to look at global data breaches trends in 2018-2019 for all sectors of economy, and to figure out how correlated data breaches (more precisely, number of breached records) with strictness of legal regulations of data privacy. It is obvious, that for the better understanding of the situation more datasets should be explored. But still, resulted data visualizations shows that despite the GDPR accepted at the EU, countries of the Union have significant data breaches. It can be assumed that GDPR realizes differently in each EU country, and, probably, some countries have additional rules and regulations that allow to decrease number of data breaches incidents. Although, the United States does not have strict federal law which protects data privacy, the situation with data breaches does not look worse than in the EU.

References

1. <https://techjury.net/blog/healthcare-data-breaches-statistics/#gref>
2. <https://www.cdc.gov/phlp/publications/topic/hipaa.html#:~:text=The%20Health%20Insurance%20Portability%20and,the%20patient's%20consent%20or%20knowledge>
3. <https://dl.acm.org/doi/pdf/10.1145/3439873>
4. <https://worldpopulationreview.com/continents>