© University Press 2023
ISSN 2719-6550
ISSN 2719-7417 online

"Journal of Education, Technology and Computer Science"
No. 4(34)/2023
www.eti.ur.edu.pl

Received: 15.03.2023
Accepted for printing: 15.12.23
Published: 29.12.2023

DOI: 10.15584/jetacomps.2023.4.22

PAWEŁ DYMORA[1], **MIROSŁAW MAZUREK**[2], **MARIUSZ NYCZ**[3]

Modeling and Statistical Analysis of Data Breach Problems in Python

[1] ORCID: 0000-0002-4473-823X, Ph.D. Eng., Rzeszów University of Technology, Faculty of Electrical, and Computer Engineering, Department of Complex Systems, Poland
[2] ORCID: 0000-0002-4366-1701, Ph.D. Eng., Rzeszów University of Technology, Faculty of Electrical, and Computer Engineering, Department of Complex Systems, Poland
[3] ORCID: 0000-0002-6297-5730, Ph.D. Eng., Rzeszów University of Technology, Faculty of Electrical, and Computer Engineering, Department of Complex Systems, Poland

Abstract

The subject of the work is electronic medical record linkage threat analysis and modeling with the use of the submitted data breaches list published by the U.S. Department of Health and Human Services. Multipronged data analysis with the use of statistics utilities and data visualization has been conducted. The model forecasting the number of data breaches based on a time series mathematical model has also been built. The article reviews the tools and techniques used in data security analysis and presents practical examples of modeling and analysis that can be used in practice to improve data protection. It was shown how important it is to protect personal data, especially medical data, and what tools can be used in the educational process of data analytics for students to effect data analysis, trend assessment, and data prediction.

Keywords: Data Breach, Data Analysis, Cybersecurity, Electronic Health Record, Time Series, Statistical analysis, Forecasts

Introduction

One of the duties of healthcare providers is to keep medical records. Such records contain information on the examination and treatment of the patient, such as symptoms, staff observations, diagnoses, treatment plans, and course of treatment. It may contain information about the patient's private life, such as lifestyle, habits, or recreational activities (Beltran-Aroca, Girela-Lopez, Collazo--Chao, Montero-Pérez-Barquero, Muñoz-Villanueva, 1992). Medical services

have a duty of confidentiality regarding the information provided by their clients – this is the foundation of trust in the doctor-patient relationship (Boyd, 1992; Kleinman, Baylis, Singer, 1997).

The progressive development of information technology has also affected the medical industry. It is becoming increasingly common to replace traditional paper medical records with systems such as Electronic Health Record – HER (Benefits of EHRs, 2022; Seh et al., 2020). Data security consists of confidentiality, integrity, and availability. According to the European Union, a data breach occurs when any of these three components is compromised as a result of an incident (What is a data breach and what do we have to do in case of a data breach? 2022). According to the US Department of Health and Human Services, a data breach is an unauthorized access or disclosure of data, as defined by the Privacy Rule set of standards, that compromises the security and privacy of protected (medical) information (Breach Notification Rule, 2022; Schlackl, Link, Hoehle, 2022).

Between 2005 and 2019, 249.09 million people were affected by health data leaks (Seh et al., 2020). For example, in 2018, there were 2216 data leaks in 65 countries, of which health services were affected by 536 leaks; this means that, of all the industries included in the study, the medical industry was the most affected (2018 Data Breach Investigations Report, 2022). As the study shows, the cost per such spill is calculated in millions of dollars. The subject of the study is the analysis and modeling of threats to medical records, i.e. data (often sensitive) stored by entities in the medical industry, such as hospitals, outpatient clinics, and medical networks, but also entities related to the business medical industry, such as health insurances, companies offering medical care. The analysis was based on a U.S. data source - lists of reported leaks published by the U.S. Department of Health and Human Services (Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information, 2022).

The security of electronic data in healthcare and beyond has been the subject of many scientific studies in recent years. A study (Ayyagari, 2014) analyzed 2,633 leaks in the medical and other industries, which resulted in data breaches of more than 500 million people. The results show the significant contribution of hacking attacks to these breaches but at the same time the increasing importance of leaks based on the "hu-man element". In contrast, the paper (Angst, Block, D'Arcy, Kelley, 2017) analyzed data leaks in healthcare, focusing on the evaluation of the implementation of IT security measures.

The authors of the paper (Schlackl, Link, Hoehle, 2022) highlighted the multitude of different studies and approaches developed in different environments and the need to integrate the knowledge gained on this topic. They performed a systematic search of the scientific literature on the causes and consequences of breaches of information confidentiality and integrity. The paper (Seh et al., 2020)

analyzed data on data leaks in healthcare. The data for the analysis came from multiple sources, such as the PRC database, HIPAA, and OCR reports.

In the paper, the authors (McLeod, Dolezel, 2018) aimed to build a model that is a profile of factors related to data leakage in healthcare. The theoretical basis was the Swiss Cheese Model. Backward stepwise logistic regression models were built – as output, the models estimated the probability that data leakage would occur in a given healthcare unit. The paper, therefore, seeks to attempt to draw constructive conclusions for healthcare entities on how to improve the level of security of their data.

In the educational process in the faculty of data analysis, a particularly important issue is the selection and use of appropriate IT tools based on the theory of mathematical statistics, and multidimensional data analysis as well as areas of data mining. The article reviews the tools and techniques that can be used in the training process of future data analysts, and analytical system architects that can be used in data security analysis and presents practical examples of modeling and analysis that can be used in practice to improve data protection. It is shown how important it is to protect personal data in particular medical data. The main threats, selected methods, and measures to counter such threats and errors are indicated, pointing out the low level of user education and awareness as the main problem. This study presents compelling social indicators of such magnitude that they cannot be ignored. The following paper attempts to fit a model to a time series of the number of data leaks and a time series of the number of leaks of each type. The modeling aimed to obtain a model capable of producing reliable predictions of the number of data leaks in the future.

Time series modeling

The model fits the time series of the number of data leaks and the time series of the number of leaks of each type were carried out based on the number of leaks, which were counted monthly. The period considered was the time segment from the beginning of 2010 to the end of 2022. A tool was prepared for the automatic retrieval and aggregation of data. A modeling scheme based on autocorrelation and development trend extraction was developed. The model was implemented. A simulation was run, the results were presented in a graph and the model was validated.

Mathematical model

The model for the total number of leaks was built from smaller models – sub-series, i.e. series denoting the number of leaks of one particular type (e.g. monthly numbers of leaks due to an IT/hacking incident). The sub-series were divided into two groups: group I and group II. Group I included those series showing signs of non-random trends (e.g. a trend). On the other hand, group II included those series that did not show any non-random features (so-called noise).

The partial series in the two groups were modeled in a slightly different way. It was assumed that X(t) is the actual number of leaks of a given type per month t, and x(t) is the model–predicted number of leaks per month t. Group I series was modeled using the sum of linear regression and the moving average error of this regression last in the month (1), (2), (3).

$$\varepsilon_j(t) = (at + b) - X_j(t) \quad (1)$$

$$\underline{\varepsilon}_j(t) = \frac{\varepsilon_j(t-w)+\varepsilon_j(t-w+1)+\cdots+\varepsilon_j(t-1)}{w} \quad (2)$$

$$x_j(t) = max\,(at + b + \underline{\varepsilon_j}(t), 0) \quad (3)$$

For moments of time t, where we are unable to calculate the ε(t) due to the fact that we do not know X(t-1) (future values), values ε(t) we calculate from the equation (1), by replacing unknown values X values calculated from the model x.

We model the Group II sub-series using only a moving average. The model's prediction of total leakage for month t is obtained by summing the predicted values for each subseries.

Modeling framework

An ACF autocorrelation function plot was used to investigate non--coincidental patterns in the series. Modeling involves the following steps:

- Determination of model parameters and learning and forecasting intervals.
- Aggregation of data into monthly time series of counts.
- Dividing the sub-series into groups based on autocorrelation plots.
- Simulation of sub-series models.
- Aggregation of the sub-series forecasts, obtaining a summary forecast of the number of leakages.
- Visualize the results, and compare the modeling results with real values.
- Model validation.

A computational environment with the necessary tools was created for computer data analysis and numerical model building. The Python programming language was used (with additional libraries). Analysis and modeling were carried out in the JupyterLab environment.

Data analysis

The data used for the analysis cover several years. They show the changes that have taken place during this time, including the laws governing the phenomena under study. A set divided into three parts according to the date the leakages were reported was used for the analyses:

- concerning leakages reported from the beginning of 2010 until the end of 2014;

– on leakages reported from the beginning of 2015 until the end of 2019;
– on leakages reported in 2020 and beyond.

For these periods, the abundance of individual values for the features stored in the data was examined. A series of leakage counts of each type and for each data storage location in consecutive months was created. Correlations between them and between each of them and changes in selected macroeconomic values related to health care in the United States were examined.

Analysis of the number of data breaches victims

In order to examine the distribution of the number of affected people, basic statistical indicators were calculated for the sample described; these are presented in Table 1.

Table 1. Statistical measure values for the number of leakage victims

Examined period	2010–2014	2015–2019	2020–present	Summary
Number of leaks	1207	1838	2072	5135
Average	39 802,28	105 369,8	65 489,32	73 522,81
Standard deviation	311 450,7	1 915 294	251 748,1	1 166 876
Minimum	500	500	500	500
1st quartile	1 000	1 016,5	1 450,25	1 125,5
Median	2 365	2 675	5 000	3 116
3rd quartile	7 461	10 865	28 768,5	13 734
Maximum	6 121 158	78 800 000	4 142 440	78 800 000

Since the beginning of 2020 (i.e. in three years), approximately 72% more spills have been reported than from the beginning of 2010 to the end of 2014 (four years) and approximately 13% more spills than from the beginning of 2015 to the end of 2019 (also four years). The analysis shows how the mean and median have changed. It can be seen that, in both cases, the smallest values are for the years 2010–2014. However, the mean reached its highest value for the years 2015–2019. The median, on the other hand, reaches its highest value for the most recent period – from 2020 onwards. This discrepancy may be due to the fact that in the period 2015–2019, one of the spills involved a number of over 78 million victims – the maximum value number of victims for this period. Anomalies such as this may overstate the average. This conjecture is confirmed by the very high standard deviation for the period 2015–2019.

Categorical features analysis

In the next step, the values taken by the categorical characteristics available in the data were examined, i.e.: type of entity (institution) affected, type of leakage, and storage location of the data affected by the incident.

The study was conducted by visualizing the distribution of values in a bar chart for each characteristic and for each of the three periods into which the data

were divided (Figure 1, 2, 3). Healthcare clearinghouses account for a negligible proportion of incidents in each period. It is noteworthy that there is an increasing number of spills where the healthcare provider is affected. For the characteristic "type of spill", significant changes were observed over the three periods studied. In the first period – 2010 to 2014, the leading type of data leakage was theft (theft), and a significant proportion was loss (loss) – these are categories whose representatives often have little to do with IT. An IT or hacking incident in this period was a phenomenon comparable in scale to leaks due to loss. The following two periods show a large increase in the number of leaks due to an IT or hacking incident. Interestingly, the number of leaks due to theft decreases. In the first of the periods studied, the counts for each location of information were most evenly distributed, the slight leader being data stored in paper form. In the following two periods, a large increase in the counts of two categories was observed: web server and email, i.e. digital form. It is worth noting the decreasing number of data leaks stored on a desktop computer (PC).

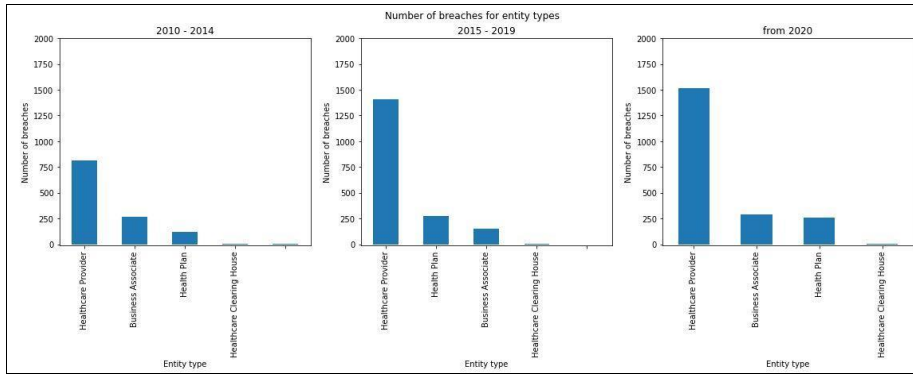

Figure 1. Number of breaches for entity types

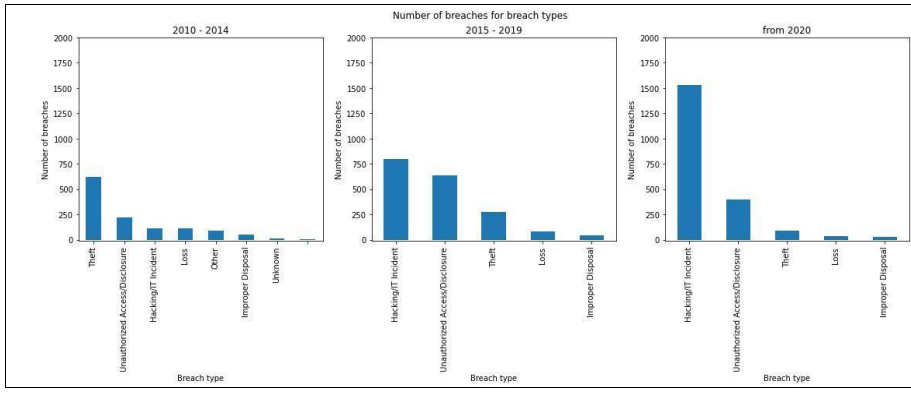

Figure 2. Number of breaches for breach types

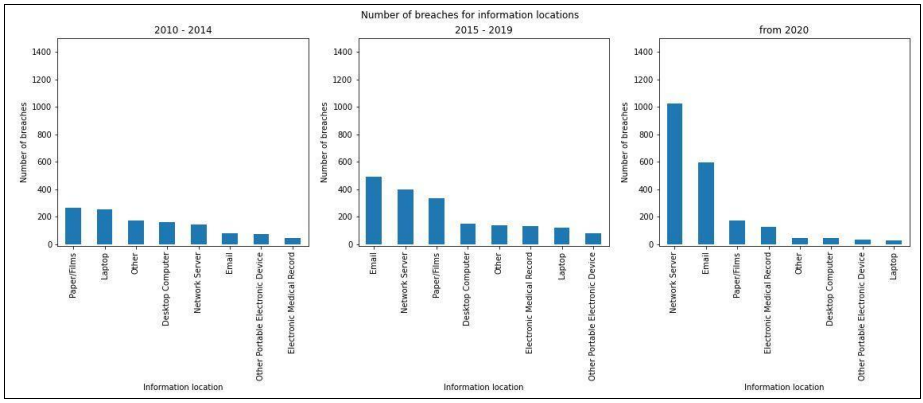

Figure 3. Number of breaches for information location

Correlation coefficient analysis

For each type of spill and each type of data location, a time series was created, being the number of reports of a given type of spill in consecutive months, as well as all reports in total, from the month in which the first spill was reported that is reported in the data, to the last month from which a spill was reported. Pearson's linear correlation coefficient was calculated for each pair of the resulting series. Two additional series were included in the analysis: the monthly change in the number of people employed in health care in thousands and the change in total profit from health and social care activities (both for the United States). For each coefficient, a statistical significance test was performed on the value of that coefficient with a significance coefficient of 0.01. The results are shown in Figure 4.

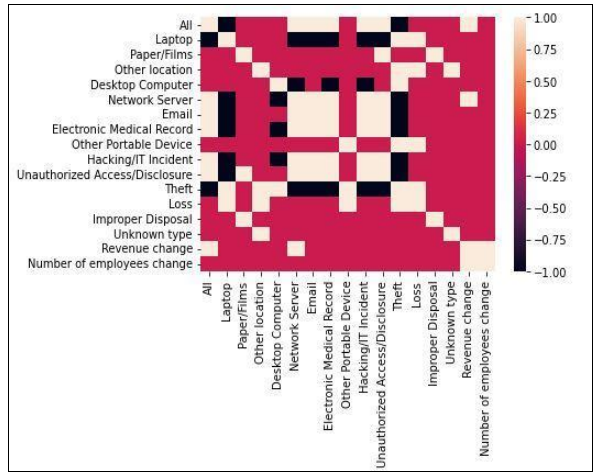

Figure 4. Map of the results of the Pearson correlation coefficient significance test (including the sign of the coefficient if the coefficient is significant)

Both negative and positive correlations were observed, as well as quite a few cases where the test showed no statistically significant correlation. It was observed that the total number of leaks per month is correlated with the monthly numbers of leaks: from a web server, from an email, from an EMR record, from a hacking/IT incident, from unauthorized access, as well as with the monthly change in total healthcare gain.

Models simulation

In the analyses, 12 (corresponding to 12 months) was used as the window length (number of observations in the window) for the moving averages. The autocorrelation analysis in the sub-series was performed for the 2010–2013 model, the 2014–2017 model, and the 2018–2021 model, respectively. Figure 5 shows the results for the 2018–2021 model.

Based on the ACF, the series for the leakage types were classified as series with non-random trends:

- 2010–2013: Other, Unauthorized Access/Disclosure.
- 2014–2017: Other, Hacking/IT Incidents.
- 2018–2021: Hacking/IT Incident.

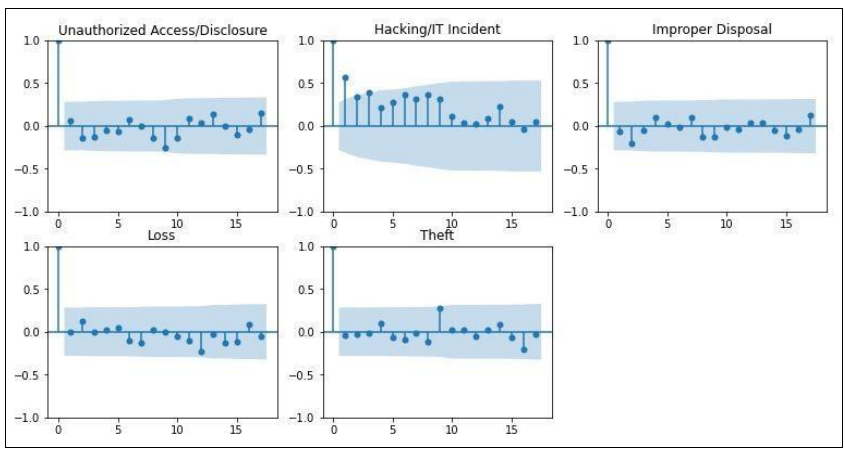

Figure 5. Autocorrelations for 2018–2021

The unspecified series for each modeled period were classified as noise. Each of the three models was simulated over a period of 48 months (4 years), including the 36 months at the values from which it was learned, and the 12 months immediately following that period (forecast).

Figure 6 shows graphs of the actual monthly number of leakages and the predictions of the individual models. The vertical lines separate the learning periods of the models.

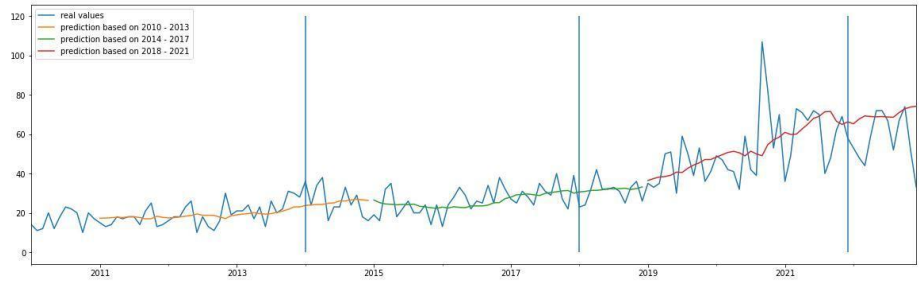

Figure 6. Autocorrelations for 2018–2021

Models validation

The first type of validation performed was a visual-informal validation, based on Figure 6. For the first two models (learned in the years periods: 2010–2013, and 2014–2017 respectively), a good fit to the learning data was observed in terms of mean value and trend, but the actual data has a much higher variability – the error changes its sign, but is rarely close to zero. The behavior of the model for the values in the months on which it performed the forecast was similar in each of the two cases to the behavior of the data used for learning – which should be read as a positive sign. For the most recent period, the model looks like a poor fit for the learning data, and the forecast is mostly overestimated. Perhaps this behavior is to some extent due to the anomaly seen in one of the months towards the end of 2020.

Table 2 presents the error measures of each of the three models, calculated for the months where each model performed forecasts (i.e. months not involved in learning).

Table 2. Calculated error measures for leakage number models

Learning range	2010–2013	2014–2017	2018–2021
Forecast period	2014	2018	2022
ME (Mean Error)	–0,77	1,16	12,13
MAE (Mean Absolute Error)	6,68	3,99	13,34
MPE (Mean Percentage Error)	6.13%	6,76%	28,03%
MAPE (Mean Absolute Percentage Error)	27,29%	14,07%	29,70%

The numerical error measures largely confirm the previously made observations in the graph. The first two models have MPEs in the order of a few percent, meaning that the forecast was not significantly under or overestimated overall. The difference is in the MAPE values, with the first model being wrong by an average of around 27.29 percent (excluding the sign), while the second model was only wrong by 14.07 percent (excluding the sign). However, the large ME and MPE clearly indicate that the model is more likely to predict values that are too large than too small.

Conclusion

In this paper, a multifaceted analysis of data on data leakage in healthcare-related institutions in the United States was performed, and a predictive model of the number of leakages was built based on these data. The data came from one of the websites belonging to the US Department of Health and Human Services. The data consisted of several thousand records from 2009 to 2023, where each record described one reported spill.

The analyses showed several clear patterns, as well as changes in the risk profile over the years. A clear change was observed regarding the types of leaks and the types of the location of the affected data. Back in the early years of the second decade of the 21st century, theft was the dominant category. Data leaked was mainly stored on laptops or in non-computerized form. Over the decade, hacking and other IT incidents have clearly become the dominant category of leaks, with data leaking mainly from web servers and emails. A significant positive correlation was observed between the counts for leak categories for which similar long-term trends were previously observed in the graph (e.g. for Email and Hacking/IT Incident, for Theft and Laptop, for Hacking/IT Incident, and Network Server).

A model was built to predict the monthly leakage rate for a selected point in time. In order to build it, the learning data was partitioned based on an autocorrelation function. The model uses a least-squares linear trend approximation and moving averages. For the temporally earliest and middle periods of the three modeled terms, the model fitted the learning data well, with the forecast trend close to that of the true data. Good results were recorded especially for the middle period, where the absolute forecast error averaged about 14 %, which was considered a satisfactory result given the large variance in the data. Worse results were obtained for the last period, where the leakage forecast tended to be significantly overestimated. This result can be interpreted to mean that towards the end of the learning period of the last model (i.e. around 2021), there was a change in the growth rate of the monthly number of leaks.

In this study, statistical evidence has been presented that shows health data breaches occurring at an unprecedented level. Preventing illegal breaches of EHR, currently taking place at such a level, is no longer possible by technology alone, and a wider discussion is needed, with relevant stakeholders involved, including patients and the public at large (patient public involvement, PPI)

References

2018 Data Breach Investigations Report. Retrieved from: https://www.verizon.com/business/resources/reports/DBIR_2018_Report.pdf (10.06.2022).

Angst, C.M., Block, E.S., D'Arcy, J., Kelley, K. (2017). When Do IT Security Investments Matter? Accounting for the Influence of Institutional Factors in the Context of Healthcare Data Breaches. *MIS Quarterly*, *41*(3), 1–XX.

Ayyagari, R. (2014). An Exploratory Analysis of Data Breaches from 2005–2011: Trends and Insights. *Journal of Information Privacy and Security*, 33–56. doi:10.1080/15536548.2012.10845654.

Beltran-Aroca, C.M., Girela-Lopez, E., Collazo-Chao, E., Montero-Pérez-Barquero, M., Muñoz-Villanueva, M.C. (1992). Confidentiality breaches in clinical practice: what happens in hospitals? *BMC Medical Ethics*, *17*.

Benefits of EHRs. Retrieved from: https://www.healthit.gov/topic/health-it-and-health-information--exchange-basics/benefits-ehrs (16.10.2022).

Boyd, K.M. (1992). HIV infection and AIDS: the ethics of medical confidentiality. *Journal of Medical Ethics*, *18*(4), 173–179.

Breach Notification Rule. Retrieved from: https://www.hhs.gov (6.10.2022).

Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information. Retrieved from: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (10.10.2022).

Kleinman, I., Baylis, F., Singer, P. (1997). Bioethics for clinicians: 8. Confidentiality. *Canadian Medical Association Journal*.

McLeod, A., Dolezel, D. (2018). Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decision Support Systems*, *108*, 57–68.

Schlackl, F., Link, N., Hoehle, H. (2022). Antecedents and consequences of data breaches: A systematic review. *Information & Management*, *59*(4), 103638. doi:10.1016/j.im.2022.103638.

Seh, A.H., Zarour, M., Alenezi, M., Sarkar, A.K., Agrawal, A., Kumar, R., Khan, R.A. (2020). Healthcare Data Breaches: Insights and Implications. *MDPI Healthcare*, *8*(2), 133. doi:10.3390/healthcare8020133.

What is a data breach and what do we have to do in case of a data breach? Retrieved from: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/what-data-breach-and-what-do-we-have-do-case-data-breach_en (6.10.2022).