Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2025, 19(2), 73–81
https://doi.org/10.12913/22998624/195690
ISSN 2299-8624, License CC-BY 4.0

Received: 2024.09.20
Accepted: 2024.12.15
Published: 2025.01.01

Text mining analysis of over 392 million compromised healthcare records

Waldemar W. Koczkodaj[1], Marek Nowacki[2], Witold Pedrycz[3], Dominik Strzalka[4*]

[1] Laurentian University, 935 Ramsey Lake Rd, Sudbury, ON P3E 2C6, Canada
[2] WSB Merito University in Poznań, ul. Powstańców Wielkopolskich 5, 61-895 Poznań, Poland
[3] University of Alberta, 116 St & 85 Ave, Edmonton, AB T6G 2R3, Canada
[4] Rzeszów University of Technology, Al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland
* Corresponding author's e-mail: strzalka@prz.edu.pl

ABSTRACT

The U.S. Department of health and human services (HHS) and the Office for Civil Rights (OCR) enforce federal civil rights laws. This study analyzed the collected data on healthcare data breaches, which affected over 392 million records in the USA from 21 October 2009 until 19 April 2024, using text mining. Using Latent Dirichlet allocation (LDA) and the Elbow methods, five major topics for text mining analysis were established. The analysis allowed to identify key breach reasons for targeting effective remedial actions and increasing data security awareness.

Keywords: data security, text mining, AI, Latent Dirichlet allocation, elbow method, text coherence, perplexity measure.

INTRODUCTION

The U.S. Department of HHS and the OCR enforce federal civil rights laws. Health Insurance Portability and Accountability Act (HIPAA) prohibits providers and businesses (called *covered entities*) from disclosing protected information to anyone other than a patient and the patient's authorized representatives without their consent [1].

In this research study, an analysis of causes for compromising health records is done by text mining of health records. Such analysis augments what is presented in [2]. It is based on similar approaches with references to [3–7]. One of our goals is to bring the US healthcare data breaches to prevent similar problems in the healthcare system in Poland.

DATA SOURCE

The main reference for this research study and data analysis are the official reports provided by the US Department of Health and Human Services Office for Civil Rights [8], available as two official reports (https://ocrportal.hhs.gov/ocr/breach/breach report.jsf). These records currently contain: (i) more than 900 breach cases that are still under investigation (the last two years, Report 1), and (ii) more than 5,200 registered, checked, and investigated cases that are moved to the archives (starting from 2009, until 2022, Report 2). Records are stored in datasets with the following details, types of breaches and breaches location – see Table 1.

METHODS

We have analyzed 5,237 records (containing information about more than 392 milion of compromised healthcare records) posted by HHSOCR categorizing them as follows:

- 3,778 records (72.1%) related to healthcare providers,
- Business associates –768 (14.7%),
- Health plans – 676 (12.9%),
- Healthcare clearing house – 10 (0.2%).

Table 1. Details about analyzed records

Records details	Types of breaches	Location of breaches	Types of covered entities
Name of covered entity	Hacking/IT incident	Desktop computer	Health plan
[US] State	Improper disposal	Electronic medical record	Healthcare cleaning house
Covered entity type	Loss	Email	Healthcare provider
[number of] Individuals affected	Theft	Laptop	
Breach submission date	Unauthorized access/disclosure	Network server	
Type of Breach	Unknown or other reasons	Other portable electronic devices	
Location of breached information		Paper/films	
Business associate present, web [breach] description		Other media or platforms	

The above mentioned 5,237 different breaches (governmental recordings) account for 392,257,850 compromised healthcare records. The most data breaches occurred in: California – 549 (10.5%), followed by Texas – 425 (8.1%) and New York – 335 (6.4%). The fewest data breaches occurred in North Dakota – 10 (0.2%), South Dakota – 12 (0.2%) and Vermont – 14 (0.3%). In the surveyed sample, the most data were leaked in 2021 – 715 (13.7%), 2020 – 663 (12.7%) and 2022 – 531 (10.1%). The fewest data records come from 2024 – 7 (0.1%), 2009 – 18 (0.3%) and 2010 – 199 (3.8%).

LDA was used in the analysis of the descriptions. LDA is a method for topic modeling in natural language processing. It allows the automatic extraction of topics from a corpus of documents. In LDA, extracting words from topics is crucial for interpreting and understanding the model's output. These key words help identify and label topics, making categorizing and summarizing documents easier. They also aid in validating and tuning the data security model by assessing topic coherence. Additionally, extracted words provide actionable insights, guiding decision-making and making the results accessible for communication and reporting to non-technical stakeholders. To estimate the number of topics, the Elbow Method was used following [9]. Additionally, the text coherence and perplexity measures were calculated [10, 11, 12]. A typical description for perplexity reveals the degree of confusion or methods measure how 'uncertain' a model is about its prediction outcomes [13]. The coherence score measures how semantically similar the words are within a specific topic [14, 15]. Before LDA analysis, the data need to be preprocessed. We change upper case to lowercase, accents are removed, stop words and numbers are removed, and the entire text is normalized using Porter Stemmer process.

RESULTS

Latent Dirichlet allocation and the Elbow method were used to determine the number of topics for text mining analysis. They were used to corroborate findings. The sum of squared errors within clusters was calculated for each of 20 proposed clusters. Analysis showed that a sharp drop in the sum of squared errors (within clusters) values occurs after cluster 5, which indicates the number of topics equal to 5 as the optimal subset [9] (see Fig. 1 and a drop value from ~181k to 176k).

The coherency and perplexity (a measure of uncertainty) coefficients were calculated for the number of topics from 2 to 20 (Fig. 2). Lower perplexity values indicate better overall generalization performance of the model. The higher coherence values indicate more coherent and interpretable topics [10]. The lowest perplexity value with the highest coherence value occurs at the point of intersection of both [16], which corresponds to five topics (see Fig. 2, the intersection of blue and dark lines). It also indicates that the 5-topic variant is optimal in the sense of "sharp decline of the sum of squared errors" (elbow method – Fig. 1) and the lowest perplexity value with the highest coherence value (Fig. 2).

The analysis of the optimal number of topics using both methods indicated that the best fit of the analyzed text occurs for five topics. Therefore, Latent Dirichlet allocation was performed for this number of topics. Table 2 presents the characteristics of the identified topics. Individual

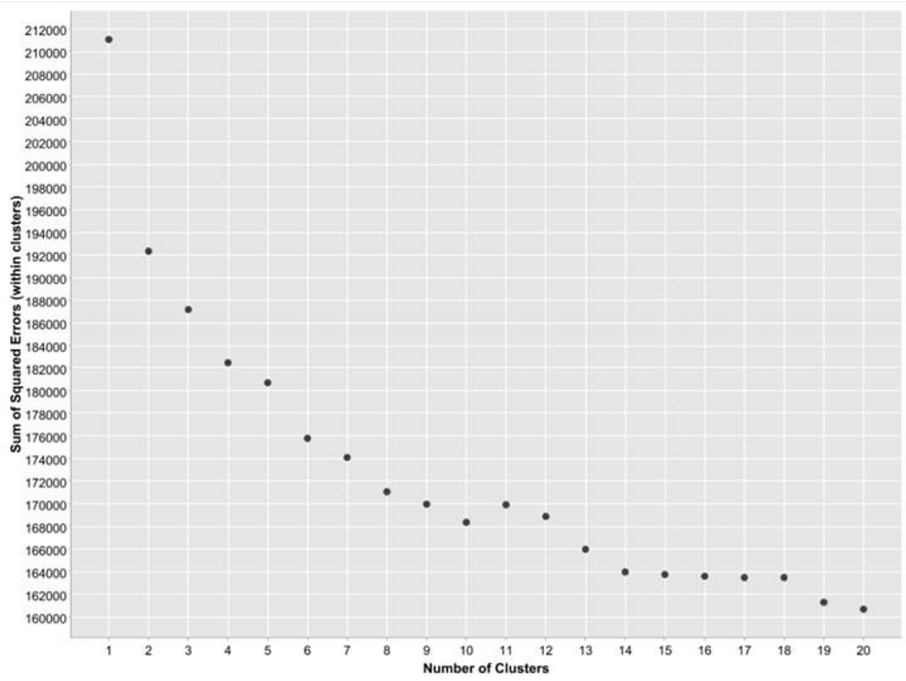

Figure 1. Number of topics: elbow method

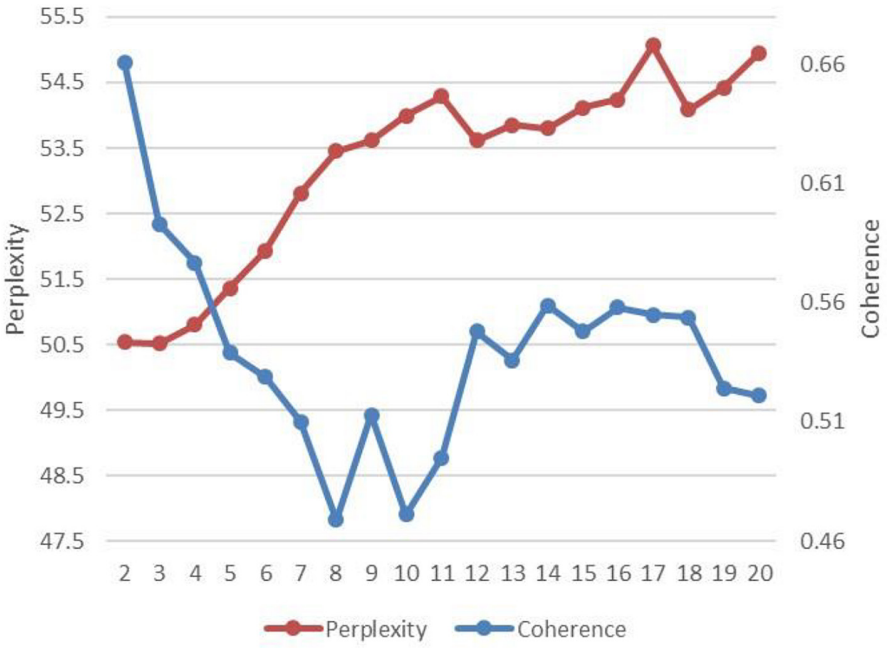

Figure 2. Number of topics: coherency and perplexity values

topics are named based on the most frequently occurring words ('Words' column).

Topic #1 – Cyberattacks and data breaches

The most numerous of the identified contained 45.25% of entries (n = 2,370).The most common words in this topic are: information, individual, affect, phi, health, provide, security, involve, protected, notify, name, include, report, number, hh, media, technical, cover, entity, and date. Incidents in this topic primarily involve ransomware attacks or other forms of cyberattacks that compromise personal data such as Social Security numbers, financial information, health insurance data, and other personal patient data. The descriptions highlight

Table 2. Topics obtained in the LDA analysis – the most common words in the description (numer of topics = 5, number of words in the topic = 20)

No.	Topic name	Words	*n*	%
1	Cyber attacks and data breaches	Information, individual, affect, phi, health, provide, security, involve, protected, notify, name, include, report, number, hh, media, technical, cover, entity, date	2,370	45.3
2	Accidental disclosure	Breach, information, individual, security, provide, computer, affect, ocr, implement, health, contain, notification, include, policy, laptop, entity, cover, number, media, steal	849	16.2
3	Improper data handling	Information, individual, email, health, affect, employee, implement, involve, report, security, entity, include, ephus, name, safeguard, notify, cover, phi, protected, hh	948	18.1
4	Detailed personal data breaches	Health, ocr, hipaa, information, security, rule, risk, investigation, plan, business, entity, report, associate, breach, agreement, ephi, privacy, include, review, action	192	3.67
5	Severe data violations	Breach, individual, phi, information, affect, provide, health, notification, ocr, patient, include, entity, employee, cover, name, number, implement, hh, media, obtain	878	16.8

remedial actions such as the introduction of additional security measures, staff training, and the offering of credit monitoring services to the victims. A typical description for this topic is: "Advanced Medical Practice Management, the business associate (BA), reported that it experienced a ransomware attack that compromised the protected health information (PHI) of 56,427 individuals. The PHI involved included names, Social Security numbers, financial information, driver's license and/or state identification numbers, dates of birth, passport numbers, electronic signature information, prescription information, medical record numbers, diagnoses, health insurance information, and other treatment information. BA notified HHS, affected individuals, and the media, and posted substitute notices on its website. In its mitigation efforts, the BA offered free credit monitoring services and implemented additional administrative, technical, and security safeguards to better protect its sensitive data. All staff were retrained." (Individuals Affected – 56,427).

Topic #2 – Accidental disclosure

It includes 16.21% of descriptions (n = 849). The most common words in it are: breach, information, individual, security, provide, computer, affect, OCR, implement, health, contain, notification, include, policy, laptop, entity, cover, number, media, and steal. Descriptions in this category include cases where patient information was accidentally disclosed to unauthorized individuals, such as sending documentation to the wrong recipient or mismanaging data in public places. Remedial actions include implementing improvements in data processing procedures, training staff, and enhancing administrative and technical security measures.

Topic #3 – Improper data handling

It includes 18.10% (n = 948) of descriptions. The most common words are information, individual, email, health, affect, employee, implement, involve, report, security, entity, include, ephus, name, safeguard, notify, cover, phi, protected, hh. Incidents in this topic focus on cases where patient data was inadvertently destroyed by external entities or due to poor management. This includes instances where data is destroyed or disposed of in ways that violate data protection regulations. In response to these incidents, institutions implement changes in data management policies and data security protections.

Topic #4 – Detailed personal data breaches

There are the fewest descriptions – 3.67% (n = 192). The most common words are health, OCR, HIPAA, information, security, rule, risk, investigation, plan, business, entity, report, associate, breach, agreement, ephi, privacy, include, review, and action. Descriptions in this topic focus on more detailed cases of personal data breaches, such as detailed medical information, test results, or detailed identification data. Remedial actions are similar to other categories but may also include more specialized procedures to protect specific data categories.

Topic #5 – severe data violations

It contains 16.77% (n = 878) entries. The most common words in this topic are breach, individual, phi, information, affect, provide, health, notification, ocr, patient, include, entity, employee, cover, name, number, implement, hh, media,

obtain. Descriptions in this topic pertain to particularly severe data breaches that involve large amounts of personal data or highly sensitive information. Remedial actions are comprehensive and may include thorough reviews and changes in IT infrastructure and data security strategies.

The analysis of differences in the proportions of Covered Entity Type occurrences in individual topics shows that topic #1 – Cyber Attacks and Data Breaches – has significantly more Business Associate breaches (60%) (Table 3). In topic #2 – Accidental Disclosure – there were significantly more Healthcare Providers (18.9%). In topic #3 – Improper Data Handling – there are significantly more breaches in Healthcare Providers (19%), and significantly fewer Business Associates (12.9%). In topic #4 – Detailed Personal Data Breaches – there are significantly more breaches related to Health Plans (8.6%). Finally, in #5 – Severe Data Violations – there were significantly more breaches related to the Health Plan (26.0%).

The analysis of differences in the frequency of topics indicates significant differences between individual years (Pearson Chi-square = 2,760.868, df = 24, $p < 0.01$) (Table 4). The topic "Cyberattacks and Data Breaches" occurred much less frequently than usual in 2014–15 (22.1%) and 2018–19 (22.8%), and in 2016–17 it almost did not occur (8.2%). We observe a very significant increase in the number of such breaches in 2020–21 (63.7%), and in 2022–24 this type of breach constitutes as much as 85.1% of cases.

Accidental Disclosures occurred most frequently in the initial years of the analysis: from 2009 to 2017, they accounted for approximately 1/3 of all breaches. A significant decline in this type of breach began in 2018–19 (16.1%), and in the following years, it virtually disappeared (< 1.0%). Improper Data Handling was very low in the initial years of the analysis – 2009–2017 (< 10.0%), while in 2018–2021 it accounted for approximately 1/3 of all breaches. However, in recent years the analysis has returned to a relatively low value (12.2%). Detailed Personal Data Breaches are a very small category of breaches (3.7%). A significant increase in these types of leaks occurred in 2012–13 (6.1%), 2014–15 (8.0%) and 2016–17 (5.1%). The last of the analyzed topics – Severe Data Violations – recorded a very strong growth in 2012–19 (with the highest share in 2016–17 – 41.4%), while in recent years of the analysis, it almost did not occur (< 2.0%).

All analyzes described above were performed in KNIME Analytics Platform 5.2. Differences in cross-tabulation proportions are verified using Pearson's Chi-square test using IBM SPSS Statistics 29.

DISCUSSION

Numerous studies have documented a consistent rise in data processing system security breaches, including recent analyses that highlight the severity of this issue [17–20]. It is clear that patients need to protect themselves and their families, and hospitals must prioritize patient safety. Statistics indicate that while many hospitals are making progress in reducing errors, accidents, injuries, and infections, the overall improvement remains insufficient [21, 22]. Presented findings suggest that a more extensive strategy is needed to prevent a looming crisis. Our analysis identified five key breach topics.

Table 3. Frequency of topics according to the Covered Entity Type (N = 5,237)

Covered Entity Type	Topics					Total
	#1	#2	#3	#4	#5	
	Cyber attacks and data breaches	Accidental disclosure	Improper data handling	Detailed personal data breaches	Severe data violations	
Not assigned (a)	40.00%	0.00%	40.00%	0.00%	20.00%	0.10%
Business associate (b)	60% ^{c,e}	10.0% ^{e}	12.9% ^{c,e}	3.0% ^{c}	14.1% ^{c}	14.70%
Health plan (c)	38.5% ^{b,e}	8.3% ^{e}	18.6% ^{b,e}	8.6% ^{b,e}	26% ^{b,e}	12.90%
Healthcare clearing house (d)	40.00%	10.00%	30.00%	0.00%	20.00%	0.20%
Healthcare provider (e)	43.5% ^{c}	18.9% ^{b,c}	19.0% ^{b}	2.9% ^{b}	15.6% ^{c}	72.10%
Row	45.30%	16.20%	18.10%	3.70%	16.80%	100.00%

Note: Pearson Chi-square = 217.002, df = 16, $p < 0.01$; Each subscript letter denotes a subset of Covered Entity Type categories whose column proportions are significantly different from each other at the 0.05 level.

Table 4. Frequency of topics according to the years

Years	Topics					Sum
	#1	#2	#3	#4	#5	
	Cyber attacks and data breaches	Accidental disclosure	Improper data handling	Detailed personal data breaches	Severe data violations	
2009–11 (a)	49.9% [b]	31.7% [b, c, d]	0.05% [b]	2.6% [e, f, g]	15.30%	8.00%
2012–13 (b)	42.0% [a]	28.3% [a, c, d]	1.4% [a, c]	6.1% [c, d]	22.2% [e]	9.50%
2014–15 (c)	22.1% [e]	30.7% [a, b, d]	4.1% [b, d]	8.0% [b]	35.10%	11.20%
2016–17 (d)	8.20%	9.6% [c]	9.6% [c]	5.1% [b, c]	41.40%	13.10%
2018–2019 (e)	22.8% [c]	37.3% [f]	37.3% [f]	2.5% [a, f, g]	21.2% [b]	16.80%
2020–21 (f)	64.7% [g]	30.7% [e]	30.7% [e]	2.9% [a, e, g]	1.2% [g]	26.30%
2022–24 (g)	85.1% [f]	12.2% [d]	12.2% [d]	0.9% [a, e, f ,d]	1.5% [f]	15.20%
Row	45.30%	16.20%	18.10%	3.70%	16.80%	100.00%

Note: Pearson Chi-square = 2,760.868, df = 24, $p < 0.01$; Each subscript letter denotes a subset of Covered Entity Type categories whose column proportions are significantly different from each other at the 0.05 level.

- Topic #1: Cyber Attacks and Data Breaches is the most prevalent (45.25%), involving ransomware and other cyberattacks that compromise sensitive data. These breaches have surged recently, emphasizing the need for enhanced cybersecurity measures and regular audits.
- Topic #2: Accidental Disclosure (16.21%) involves unintended data releases, which have decreased since 2018, indicating effective past interventions like improved data handling and staff training. However, ongoing vigilance is essential.
- Topic #3: Improper Data Handling (18.10%) includes incidents of data being destroyed or mismanaged. The rise in these breaches during 2018-2021 suggests a need for stricter data management policies.
- Topic #4: Detailed Personal Data Breaches (3.67%) involves breaches of highly sensitive information, requiring specialized remedial actions. These breaches are less common but can have severe consequences.
- Topic #5: Severe Data Violations (16.77%) includes large-scale breaches of highly sensitive data, necessitating comprehensive security overhauls.

Understanding the prevalence of these topics is vital for targeting remedial actions effectively [23, 24]. For example, the significant rise in cyberattacks highlights the urgency of investing in advanced cybersecurity infrastructure. The decline in accidental disclosures suggests that existing interventions are working, but ongoing efforts are needed to maintain these gains. The increase in improper data handling breaches calls for revisiting and strengthening data management practices. By focusing resources on the most common and severe breach types, organizations can better protect sensitive data and improve overall security [25]. Future research should continue monitoring these trends to adapt strategies accordingly.

The potential for direct or indirect access to resources creates a temptation to exploit these resources for financial gain [26]. Medical data, in particular, is highly valuable because it does not become obsolete as quickly as financial data, which can be rendered useless if, for instance, customers promptly change compromised credit card numbers [27].

Data security is a multifaceted issue. The Internet's role as the most convenient, fastest, and cheapest means of accessing data has made EHR breaches not just a matter of rapid Internet expansion but also an Internet and IT systems security challenge [28]. Health data breaches can result from various factors, including vulnerable software that leads to hacking and unauthorized access or user errors such as failing to log off or using weak passwords [29]. Entities handling these data and their business partners must have comprehensive security plans incorporating physical, administrative, and technical safeguards [30, 31]. Effective incident response requires minimizing the number and severity of security incidents, assembling a core Computer Security Incident Response Team (CSIRT), defining an incident response plan, and containing damage to mitigate risks [32].

Healthcare policy executives must work closely with IT departments to develop strategies that address the latest threats. One key issue is the impact of security breach announcements on

market value. Leaked information poses significant risks to capital markets and companies and can lead to stock market speculation [33]. Security concerns become increasingly critical as network interconnections become more complex [34]. A common solution involves using a mix of routers, switches, firewalls, VPNs, intrusion detection systems (IDSs), and vulnerability assessment tools to secure network-aware software applications and systems [35]. With the increasing complexity and destructiveness of attacks on critical network infrastructures, new methods are essential for aiding security administrators in protecting their networks [36]. Despite many models for protecting IT systems and networks, their effectiveness remains limited [37, 38].

To address the lack of trust in Internet use, especially for commercial purposes and online purchasing, web developers must create and maintain robust applications that can resist external threats [39, 40, 41]. O'Connor observed [42]: "We found that half of states have no statutes addressing nondisclosure of personally identifiable health information generally held by public health agencies."

A well-secured processing system should use advanced security tools to protect patient data. Various technical solutions, such as data access monitoring, security event and information management (SIEM) systems, tokenization, and cloud security gateways, add layers of security, making it harder for hackers to breach systems and reducing the impact of human error on data security [43]. Our study emphasizes the importance of public awareness regarding the dangerous consequences of careless use of information technology [44], such as taking sensitive data on notebooks, tablets, or USB drives from hospitals to work on at home, which violates regulations like HIPAA and those of most hospitals [45]. As earlier studies changed attitudes from irresponsible cover-ups to corrective actions, we hope this study will contribute to a similar transformation [46].

Beyond the broader implications, it is also crucial to consider the impact of data breaches on individuals. A lack of data security affects users' well-being and quality of life. Trust in the Internet significantly affects public health outcomes, such as quality of life [47]. There is a widespread perception among Internet users of a global lack of trust in using the Internet, for instance in [48] the case of harmful internet use was considered. Research must explore the Internet's role in quality of life (QoL).

CONCLUSIONS

The presented findings indicate considerable problems with confidentiality of health records. Information technology contributed to this problem and we call for action to develop better way of protecting our medial records.

The authors hope that the significance of the presented findings may attract the attention of proper Parliamentary Committee of Poland. We also hope that the Polish Parliament may pass laws for collecting data about healthcare data breaches similar to the US Department of Health and Human Services Office for Civil Rights, lowering the limit of 500 stolen records to 50 since the Polish population is approximately ten times lower than that of the USA.

Acknowledgments

Authors would like express deepest appreciation to adjunct Julia Johnson (Laurentian University, Professor Emerita) and Dr Tamar Kakiashvili, MD (of SudburyTherapy.com) for editorial enhancements.

REFERENCES

1. HSSOCR, (2024), https://ocrportal.hhs.gov/ocr/breach/breac report.jsf
2. Syed S., Spruit M. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 2017, 165–174, https://doi.org/10.1109/DSAA.2017. 61
3. Nam, J., Lee, H., Lee, S., Park, H. Literature review of complementary and alternative therapies: using text mining and analysis of trends in nursing research, BMC NURSING, 2024, 23, 526, https://doi.org/10.1186/s12912-024-02172-9
4. Limsomwong, P., Ingviya, T., Fumaneeshoat, O. Identifying cancer patients who received palliative care using the SPICT-LIS in medical records: a rulebased algorithm and text-mining technique, BMC Palliative Care. 2024, 23, 83, https://doi.org/10.1186/s12904-024-01419-1
5. Choi, S. Perceived challenges and emotional responses in the daily lives of older adults with disabilities: A Text Mining Study, Gerontology and Geriatric Medicina, 2024, 10, 23337214241237097, https://doi.org/10.1177/23337214241237097
6. Ji, M. and Mosaffa, M., and Ardestani-Jaafari, A.

and Li, J. and Peng, C. Integration of text-mining and telemedicine appointment optimization, Annals of Operations Research, 2023, https://doi.org/10.1007/s10479-023-05660-4

7. Boxley, C. and Fujimoto, M., and Ratwani, R.M. and Fong, A. A text mining approach to categorize patient safety event reports by medication error type, Scientific Reports, 2023, 13, 183154, https://doi.org/10.1038/s41598-023-45152-w
8. HSSOCR. 2024. https://ocrportal.hhs.gov/ocr/breach/breach report.jsf
9. Koczkodaj, W.W., Mazurek, M., Strzalka, D., Wolny-Dominiak, A., Woodbury-Smith, M. Electronic Health Record Breaches as Social Indicators, Social Indicators Research, 2019, 141(2), 861–871.
10. Hasan, M., Rahman, A., Karim, M.R., Khan, M.S.I., Islam, M.J. Normalized approach to find optimal number of topics in latent dirichlet allocation (LDA). In: Kaiser, M.S., Bandyopadhyay, A., Mahmud, M., Ray, K. (eds) Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, 2021, 1309. Springer, Singapore. https://doi.org/10.1007/978-981-33-4673-4_27
11. Chang, H.-Y. and Yang, Y.-H. and Lo, C.-L. and Huang, Y.-Y. Factors Considered Important by Healthcare Professionals for the Management of Using Complementary Therapy in Diabetes A Text-Mining Analysis, CIN-Computers Informatics Nursing, 2023, 41, 426–433, https://doi.org/10.1097/CIN.0000000000000977
12. Ahmad, P.N. and Shah, A. M. and Lee, K.Y. A Review on Electronic Health Record Text-Mining for Biomedical Name Entity Recognition in Healthcare Domain, Healthcare, 2023, 11, 1268, https://doi.org/10.3390/healthcare11091268
13. Javad P., Saeed A., Farhad F., Burton-Jones A. A systematic analysis of failures in protecting personal health data: A scoping review, International Journal of Information Management, 2024, 74, 102719.
14. Koczkodaj, W.W., Masiak, J., Mazurek, M., Strzalka, D., Zabrodskii, P.F. Massive health record breaches evidenced by the office for civil rights data, Iranian Journal of Public Health, 2019, 48(2), 278–288.
15. Gorgol, I. The use of complex networks tools to describe the current state of multidisciplinary research in Poland, Advances in Science and Technology-Research Journal, 2020, 14, 125–138, https://doi.org/10.12913/22998624/126970
16. Basil NN, Ambe S, Ekhator C, Fonkem E. Health records database and inherent security concerns: a review of the literature. Cureus. 2022 Oct 11; 14(10): e30168. https: //doi.org/10.7759/cureus.30168
17. Nemec Zlatolas, L., Welzer, T., Lhotska, L. Data breaches in healthcare: security mechanisms for attack mitigation. Cluster Comput. 2024. doi.org/10.1007/s10586-024-04507-2
18. Khan S., Khan HU., Nazir S. Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing. Scientific Reports 2022, 12(1), 22377.
19. Koczkodaj W.W., Kakiashvili T., Szymanska A., Montero-Marin J., Araya R., GarciaCampayo J., Rutkowski K., Strzalka, D. How to reduce the number of rating scale items without predictability loss? Scientometrics 2017, 111(2), 581–593, https://doi.org/10.1007/ s11192-017-2283-4
20. Kisilowski, M. Mathematical and Technical Quantitative Methods for Risk Assessment in Public Crisis Management, Advances in Science and Technology-Research Journal, 2023, 17, 215–225, https://doi.org/10.12913/22998624/162188
21. Khalid, H., Wade, V. Topic detection from conversational dialogue corpus with parallel dirichlet allocation model and elbow method, [in:] David C. Wyld et al. (Eds): ITCSE, NLCA, ICAIT, CAIML, ICDIPV, CRYPIS, WiMo – 2020 pp. 95-102. CS & IT – CSCP 2020, https://doi.org/10.5121/csit.2020.100508
22. Awrahman B.J., Aziz Fatah C., Hamaamin MY. A review of the role and challenges of big data in healthcare informatics and analytics. Computational Intelligence and Neuroscience 2022, 5317760.
23. Koczkodaj, W.W., Szybowski, J. and Wajch, E. Inconsistency indicator maps on groups for pairwise comparisons, Innternational Journal of Approximate Reasoning 2016, 69, 81–90.
24. Kucharski, D., Kajor, M., Grochala, D., Iwaniec, M., Iwaniec, J. Combining Spectral Analysis with Artificial Intelligence in Heart Sound Study, Advances in Science and Technology-Research Journal, 2019, 13, 112–118, https://doi.org/10.12913/22998624/108447
25. Mart´ınez AL., P´erez MG., Ruiz-Mart´ınez A. A Comprehensive Review of the Stateof-the-Art on Security and Privacy Issues in Healthcare. ACM Comput. Surv. 2023, 55, 12, Article 249 (December 2023), 38. https://doi.org/doi.org/10.1145/3571156
26. Trkman M., Popovic A., Trkman P. The roles of privacy concerns and trust in voluntary use of governmental proximity tracing applications, Government Information Quarterly, 2023, 40(1), 101787.
27. Almutairi, Y. and Alhazmi, B. and Munshi, A. Network Intrusion Detection Using Machine Learning Techniques, Advances in Science and Technology-Research Journal, 2022, 16, 193–206, https://doi.org/10.12913/22998624/149934
28. Al Zaabi, M., Alhashmi, S. M. Big data security and privacy in healthcare: A systematic review and future research directions. Information Development, (to be published), 2024. https://doi.org/10.1177/02666669241247781
29. Breve B., Desolda G., Deufemia V., Spano L. D. Detection And Mitigation Of Cyber attacks that

exploit human vuLnerabilitiES (DAMOCLES 2024). In Proceedings of the 2024 International Conference on Advanced Visual Interfaces (AVI '24). Association for Computing Machinery, New York, NY, USA, 2024, Article 125, 1–4. https://doi.org/10.1145/3656650.3660540

30. Shojaei, P., Vlahu-Gjorgievska, E., Chow, Y.-W. Security and privacy of technologies in health information systems: A systematic literature review. Computers, 2024, 13, 41. https: //doi.org/10.3390/computers13020041
31. Papanikolaou, Y. Foulds, J.R, Rubin, T.N., Tsoumakas; G. Dense Distributions from Sparse Samples: Improved Gibbs Sampling Parameter Estimators for LDA, Journal of Machine Learning Research, 2017, 18(62) 1–58.
32. Latulipe C., Mazumder S.F., Wilson R.K.W., et al. Security and privacy risks associated with adult patient portal accounts in US hospitals. JAMA Internal Medicine 2020, 180(6), 845–849.
33. Alotaibi Y.K., Federico F. The impact of health information technology on patient safety. Saudi Med. J. 2017, 38(12), 1173–1180. https://doi.org/10.15537/smj.2017.12.20631
34. Kruse C.S., Mileski M., Vijaykumar AG., Viswanathan S.V., Suskandla U., Chidambaram Y. Impact of electronic health records on long-term care facilities: Systematic review. JMIR Med Inform. 2017, 5(3), e35. Published 2017 Sep 29. https://doi.org/10.2196/medinform.7958
35. Steinke, J., Bolunmez, B., Fletcher, L., Wang, V., Tomassetti, A.J., Repchick, K.M., Zaccaro, S.J., Dalal, R.S., and Tetrick, L.E. Improving Cybersecurity Incident Response Team Effectiveness Using Teams-Based Research, IEEE Security & Privacy 2015, 13, 4, 20–29.
36. Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C. and Yao, H. Research on Topic Detection and Tracking for Online News Texts, IEEE Access, 2019, 7, 58407–58418.
37. Lawrence O. Gostin, James G. Hodge, Jr, and Ronald O. Valdiserri Informational Privacy and the Public's Health: The Model State Public Health Privacy Act, American Journal of Public Health 2001, 91, 1388–1392, https://doi.org/10.2105/AJPH.91.9.1388
38. O'Connor J., Matthews G. Informational privacy, public health, and state laws. Am J Public Health. 2011 Oct, 101(10), 1845–50. https://doi.org/10.2105/AJPH.2011.300206
39. James G. Hodge Jr, Jennifer L. Piatt, Erica N. White, and Lawrence O. Gostin. Public health legal protections in an era of artificial intelligence, American Journal of Public Health 2024, 114, 559–563, https://doi.org/10.2105/AJPH.2024.307619
40. Wartenberg D., Douglas W.T. Privacy versus public health: The Impact of Current Confidentiality Rules American Journal of Public Health 2010, 100, 407–412, https://doi.org/10.2105/AJPH.2009.166249
41. Abouelmehdi K., Beni-Hessane A., Khaloufi H. Big healthcare data: preserving security and privacy. Journal of big Data 2018, 5(1), 1–18.
42. Senol-Durak E., Durak M. The mediator roles of life satisfaction and self-esteem between the active components of psychological well-being and the cognitive symptoms of problematic internet use. Social Indicators Research, 2011, 103(1), 23–32.
43. van de Burgt, Britt W.M., Wasylewicz, Arthur T.M., Dullemond, B., Jessurun, Naomi T., Grouls, Rene J.E., Bouwman, R. Arthur, Korsten, Erik H. M., Egberts, Toine C.G., Development of a text mining algorithm for identifying adverse drug reactions in electronic health records, JAMIA OPEN, 2024, 7, ooae070, https://doi.org/10.1093/jamiaopen/ ooae070
44. Myers, J. Frieden T.R., Bherwani K.M., and. Henning K.J. Ethics in public health research, American Journal of Public Health 2008, 98, 793–801, https://doi.org/ 10.2105/AJPH.2006.107706
45. Ghafur, S., Van Dael, J. A retrospective analysis of cybersecurity threats impacting patient safety in the UK. BMJ Health & Care Informatics, 2019, 26(1), e100075.
46. Brennan T.A., Leape L.L., Laird N.M., Hebert L., Localio A.R., Lawthers A.G., Newhouse J.P., Weiler P.C., and Hiatt H.H. Incidence of adverse events and negligence in hospitalized patients. Results Of The Harvard Medical Practice Study I. The New England Journal of Medicine. 1991, 324, 370–376.
47. Sittig, D.F., Singh, H. A Socio-technical Approach to Preventing Health IT-related Patient Safety Hazards: The Need for Clinical and IT Collaboration. Journal of Biomedical Informatics, 2018, 78, 161–167.
48. Koczkodaj, W. W., Kowalczyk, A., Mazurek, M., Pedrycz, W., Redlarski, G., Rogalska, E., Strzalka, D., Szymanska, A., Wilinski, A., & Xue, O.S. Peer assessment as a method for measuring harmful internet use. MethodsX, 2023, 11, 102249. https://doi.org/10.1016/j.mex.2023.102249
49. Yilmaz, F. Evaluation of Working Conditions and Professional Independence Perceptions of Occupational Health and Safety Professionals, Advances in Science and TechnologyResearch Journal, 2021, 15, 118–125, https://doi.org/10.12913/22998624/142215
50. Koczkodaj, W.W., Szybowski, J. The limit of inconsistency reduction in Pairwise Comparisons, International Journal of Applied Mathematics and Computer Science, 2016, 26(3), 721–729. https://doi.org/10.1515/amcs-2016-0050