Evolution in the Research Pattern of Healthcare IT Security & Privacy

NKU | subramanis1@nku.edu

MBI 692 – RESEARCH PROJECT RePORT

Selva damodaran

2021

Contents

[Introduction 2](#_Toc70958250)

[Background 2](#_Toc70958251)

[What is Healthcare Privacy & Security? 2](#_Toc70958252)

[HIPAA and PHI 3](#_Toc70958253)

[Current State of Data Security & Privacy in Healthcare IT 4](#_Toc70958254)

[Researches surrounding the Privacy of Healthcare Data 5](#_Toc70958255)

[Related Works 6](#_Toc70958256)

[Methodology and Results 7](#_Toc70958257)

[Loading Data 7](#_Toc70958258)

[Exploratory Data Analysis 8](#_Toc70958259)

[LDA Model Development and Training 12](#_Toc70958260)

[Validation of LDA Model Results 14](#_Toc70958261)

[Conclusion 18](#_Toc70958262)

[References 19](#_Toc70958263)

# Introduction

Healthcare information systems are largely viewed as the single most important factor in improving US healthcare quality and reducing related costs. According to the Office of the National Coordinator for Health Information Technology, 95% of hospitals across the US have adopted certified EHR technology. Anecdotal evidence from recent years suggest that a lack of adequate security measures has resulted in numerous data breaches, leaving patients exposed to economic threats, mental anguish and possible social stigma (Health Privacy Project, 2007). A recent survey in the USA suggests that 75% of patients are concerned about health websites sharing information without their permission (2). Possibly, this patient perception is fueled by the fact that medical data disclosures are the second highest reported breach (3). In response to these increasing threats to health information and privacy, new regulations at both the state and the federal level have been proposed in the USA, e.g., Health Insurance Portability and Accountability Act (HIPAA).

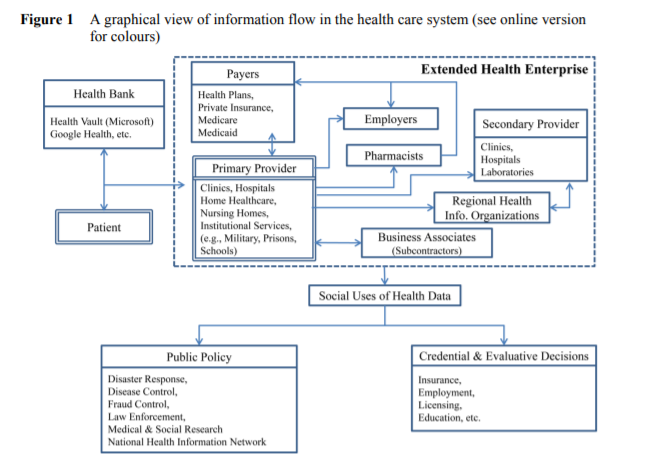
Over the past two decades, information security research has become a well-established area within the information systems discipline. Researchers have adopted several underlying theories from reference disciplines such as psychology and sociology to analyze information security risk management (4, 5, 6) and economic theories to characterize investment decisions and information governance (7, 8). Since then, there were several studies conducted in the preserving the security and privacy of patient data shared in Healthcare IT systems. In this paper, our aim is to study the trend of research in Healthcare IT Security and decipher a pattern that the EHR Privacy and Security has taken through these years. We first aim to explain the data / information flow in these EHR landscapes, then address the hole which creates the risk of security & privacy violations, and finally build a classification model to detect the pattern of researches conducted in EHR security.

# Background

## What is Healthcare Privacy & Security?

Privacy is viewed as a key governing principle of the patient–physician relationship. Patients are required to share information with their physicians to facilitate correct diagnosis and treatment, and to avoid adverse drug interactions. However, patients may refuse to divulge important information in cases of health problems such as psychiatric behavior and HIV, as their disclosure may lead to social stigma and discrimination (10). Over time, a patient’s medical record accumulates significant personal information including identification, history of medical diagnosis, digital renderings of medical images, treatments, medication history, dietary habits, sexual preference, genetic information, psychological profiles, employment history, income and physicians’ subjective assessments of personality and mental state (9).

Figure 1 shows a typical information flow in the healthcare sector. Patient health records serve a range of purposes apart from diagnosis and treatment provision. For example, information could be used to improve efficiency within the healthcare system, drive public policy development and administration, and in the conduct of medical research (12). A patient’s medical records are also shared with payer organizations (e.g., private insurance or Medicare/Medicaid) to justify payment of services rendered. Healthcare providers also use records to manage their operations and improve service quality.



## HIPAA and PHI

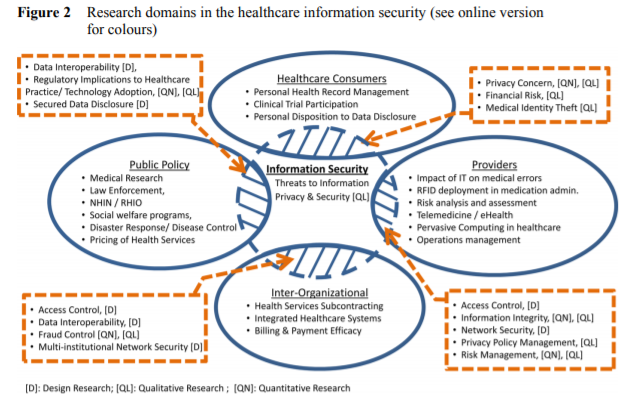
HIPAA has always been referred to as a complicated and overwhelming law that healthcare organizations have struggled to make sure that they are fully compliant with. That is why it is important to be reminded why protected health information is so valuable to criminals and hackers and why keeping it protected from them is so vital.

Protected health information, or PHI, which was defined by the HIPAA privacy rule, is any information within a person’s medical record that can identify them and is held by a covered entity. Under HIPAA and the Privacy Rule, there are 18 specific identifiers that must be handled with certain strict safeguards. Healthcare records are known to be one of the most valuable types of information that hackers look for. Most of the PHI that is compromised throughout the industry happens through hacking or IT incidents. That is because of the high value of PHI compared to other information that hackers may be able to find.

Another reason that medical records are extremely valuable to hackers is that there are many ways to use that data on the dark web. This information can be used to purchase prescriptions, receive treatment or make fake medical claims. These actions can cause long-term and widespread chaos for those whose information has been stolen. A breach of PHI can pose a real threat to patients and healthcare systems alike, so it's worth protecting.

## Current State of Data Security & Privacy in Healthcare IT

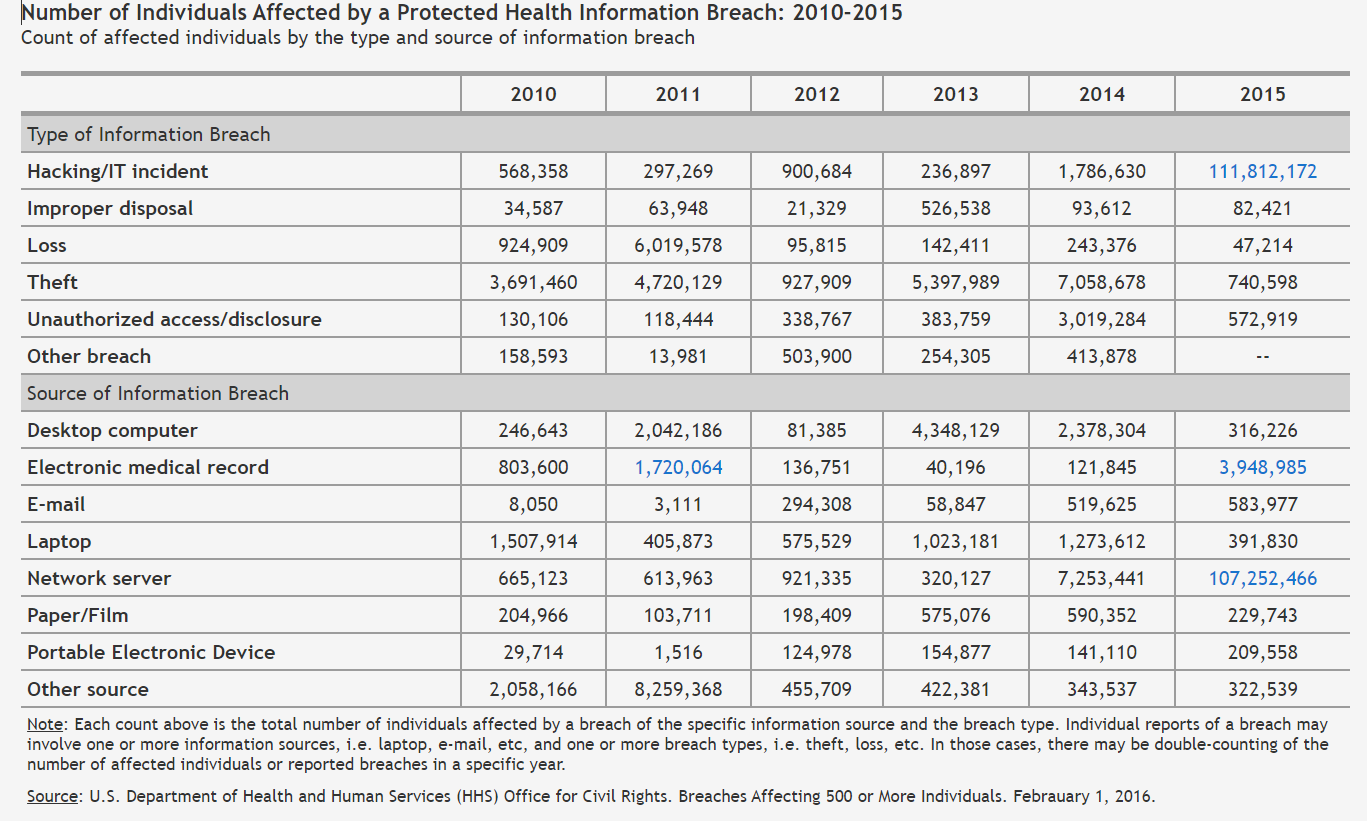
First, research on issues related to healthcare consumers, including personal health record management and web-based EHR systems, have raised a number of security-related topics including the drivers of privacy and security concerns among consumers, monetary impact of privacy and security breaches to consumers, and impact of medical identity theft on consumers’ well-being. Second, research focused on issues related to providers, such as the drivers of IT adoption, impact of IT on medical errors, telemedicine, pervasive computing and RFID adoption, also interacts with emerging security issues, for example, the design and development of access control systems, sustainability of information integrity, network security, privacy policy management and risk management. Similarly, research focusing on inter-organizational issues such as health services subcontracting, design and development of inter-organizational health networks, and EDI adoption gives rise to security and privacy research problems such as inter-organizational access control, data interoperability, multi-institutional network security and fraud control. Lastly, several information security and privacy research directions (e.g., development of data interoperability standards, regulatory implications of healthcare technology adoption and secured data disclosure mechanisms) have emerged in the public policy domain, particularly in areas such as medical research, development of national health information network, disaster response and pricing of health services.



Most individuals (84%) are confident that their medical records are safe from unauthorized viewing but have concerns (66%) when health information is electronically exchanged (Health IT Dashboards Quick Stats, 9). As compared to 2017, many of them are now confident about the privacy and protection of PHI records. This figure certainly shows a 15% increase from 2014. The recent report from ONCHIT in 2017 (10) shows that 1 in 3 individuals tracked health care charges and costs with a computer, smartphone, or other electronic means in the past 12 months. And about 4 in 10 individuals filled out paperwork related to their health care or made health care appointments online. Over one-third communicated with their provider online, while 3 in 10 individuals communicated with their health care provider via text message. Just over 4 in 10 individuals looked for a health care provider online.

With this increased use of health IT systems, the amount of PHI data flow over the internet is as well high. With this increased availability of PHI data on the web, hacking of this personal health information became easier. Each year Verizon releases data breach reports that tell the story of that year's worth of breaches that have occurred. Between the 2016 and 2019 reports, the number of data incidents and breaches increased by three times. The recent 2020 report shows that these numbers have continued to grow, now revealing a 71% increase in the number of breaches this year. With many of the challenges due to COVID-19 and a work-from-home environment, organizations need to be more aware than ever that the PHI they are responsible for is completely secure and protected.

Figure below shows the number of PHI data attacks encountered between 2010 and 2015.



## Researches surrounding the Privacy of Healthcare Data

Several organizations have already put in place technological solutions for maintaining and managing their patient’s privacy over wired and wireless networks. Numerous research studies have been conducted regarding the information protection of patients’ PHI data. Majority of the design research focuses on developing artefacts such as models, algorithms, prototypes and frameworks to solve specific problems related to data protection and governance in HER systems (11). Studies have also been performed aiming at a qualitative research, which involves examining a social phenomenon using a range of qualitative instruments/data such as interviews, documents, participants’ observation data, researcher’s observation and impression (12). In healthcare research, much of the qualitative research centers around the impact of HIPAA on healthcare practices (13). Lastly, researchers in healthcare information systems have adopted several quantitative methods including surveys, econometric analysis and statistical modelling in the areas of patients’ privacy concern, public policy, fraud control, risk management and impact of health IT on medical errors (14).

There are several threats to information privacy. Two of the major threats include the Organization Threat arising from inappropriate access of patient data by either internal agents abusing their privileges or external agents exploiting a vulnerability of the information systems and Systematic Threats rising from an agent in the information flow chain exploiting the disclosed data beyond its intended use. Organizational threats may be further categorized into multiple levels – Accidental Disclosure, Insider Curiosity, Insider Data Breach, Outsider Data Breach, etc. depending on the order of sophistication. Additionally, the emergence of cloud-based healthcare provider systems has transformed the business model greatly. Several researches have been conducted to identify the top security threats in healthcare systems. And numerous studies have been made to measure the effectiveness of security providing systems over healthcare data breaches. Results of these scientific research have been digitized and stored as news, scientific articles, and books. Yet there are multitude of events reported in relation to PHI security threat.

This paper focusses on identifying the various topics of research in the field of healthcare data security and privacy, studying the evolution of its research pattern over the last decade and predicting the trend of researches that will be performed in this field.

# Related Works

Several works of literature review have been performed surrounding the healthcare privacy concerns, malicious attacks and implementation of security providing systems to combat these data breaches. Ahmadi et al., reviewed and analyzed 60 research papers published between 2000 and 2016 regarding application of IoT in healthcare security. Their study employed a systematic review approach on the data that was collected (15). A similar study was conducted by Nazir et al., where again a systematic literature review of papers was performed. Researches published related to Mobile Computing in IoT for Healthcare were reviewed in their analysis (18). Mehraeen et al., performed a full text review of 210 articles published about security challenges in healthcare cloud computing technology and listed the information protection systems that are widely adopted to mitigate security incidents on healthcare applications (16). Ermakova et al., aimed to identify the state of research and determine potential areas of future work in the context of healthcare cloud computing (22). Ahmed et al., performed a question-based formalization to spot the major concerns in healthcare security, and had done a detailed review analysis of papers collected for their constructed search terms (17). Yao et al., studied the adoption and implementation of RFID in Healthcare through a systematic literature review of researches on RFID use in healthcare. They have followed a formal innovation-decision framework in their study (19). Wamba et al., highlighted the increased risk of RFID applications in healthcare and the potential issues that it could bring to the healthcare privacy (20). They developed a conceptual framework to perform an extant literature review in their study. Box and Pottas studied the gap between intent to use IT and actual compliance in Healthcare patient privacy by adopting an Appraisal Tendency Framework to review collected literature (21). Tandon et al., explored the application of blockchain in healthcare and then profiled & organized the intellectual capital in that domain. And integrated framework was developed during their research to illustrate thematic areas of future research (23). Despite numerous publications in the context of healthcare security and privacy there is no automated review of papers to decipher the current state of research so far in the domain. This paper aims to address this gap through a thematic analysis and identify the evolution pattern of research topics over the last decade in the healthcare privacy domain.

# Methodology and Results

NKUs Steely Library was considered as the source of information for articles related to healthcare security and privacy. Three keyword combination strings were used for this database search – Healthcare Security, Healthcare Privacy, Health IT. Papers searched were the ones published between Jan 2016 and March 2021. Results obtained for these terms were for the search performed in March 2021. These keywords were drawn from a review of prior studies (i.e. SLRs) in this field that used similar keywords, i.e. healthcare (or health\*), and IoT (or Internet of Things\*), and Data Security (or Data Privacy\*), etc. (19, 22, 23)

Table 1: Database Search Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Database** | **Keywords** | **Total hits appeared** | **Abstracts read** | **Full text downloaded** |
| NKU Steely Library | Healthcare Security | 217 | 217 | 195 |
| NKU Steely Library | Health IT | 138 | 138 | 132 |
| NKU Steely Library | Healthcare Privacy | 236 | 234 | 230 |

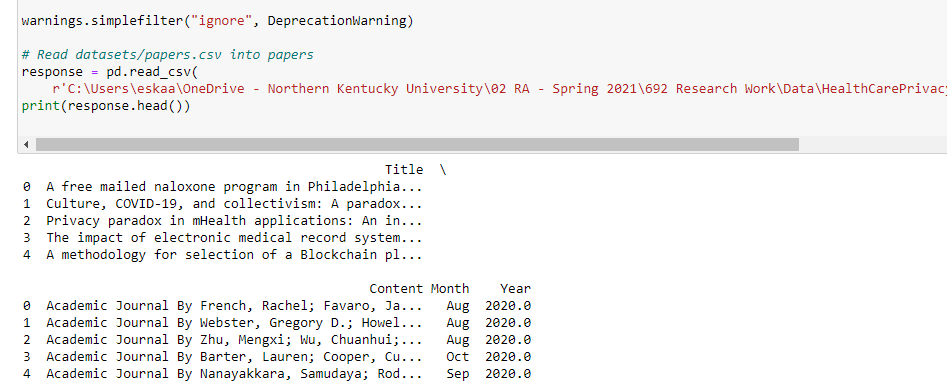
Note: Results include articles from multiple disciplines such as medicine, genomics, information science, banking etc. Multiple sources and document types were reflected in the search results including journals, trade magazines, books etc. Some papers appeared in the result list more than once and these were considered only once for further analysis.

Results of the search were sorted for “relevance” before performing the Topic Modelling. The topic mining of Research Paper titles was done through the implementation of LDA algorithm. As part of this LDA model development and validation, following steps were performed.

1. Data Cleaning, Data Preparation and Data Loading for analysis
2. Exploratory Data Analysis
3. LDA Model Development and Training
4. Validation of LDA Model results

## Loading Data

The dataset collected in the initial phase of this research was used in this step. The dataset comprised of 542 unique articles related to healthcare security and privacy collected from NKUs Steely Library. The CSV data file contains information on different Papers that were published between Jan 2016 and Mar 2021 (5 years and 3 months of data). These papers discuss a wide variety of topics on protecting the security and privacy of PHI & PII data in health care field.



## Exploratory Data Analysis

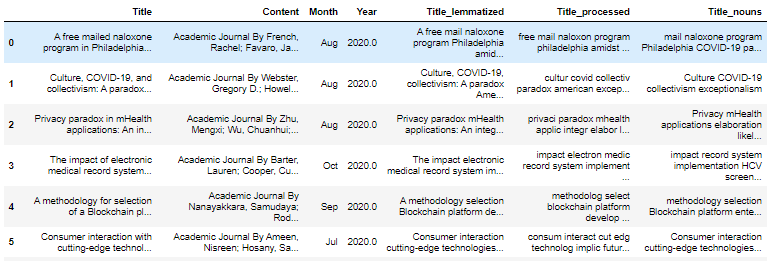
At first, a time series chart of the number of articles published year-wise during the last 5 years was plotted to get an understanding of how the papers were distributed.

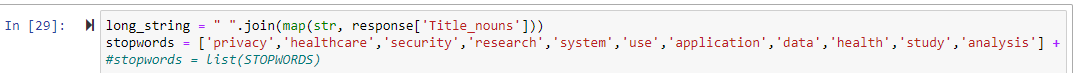
Chart, line chart

Description automatically generated

In this step, simple preprocessing of data was performed on the “Title” column of the dataset. As part of this stage, text analytics preprocessing steps such **as removal of punctuation, conversion to lowercase, lemmatization, extraction of nouns and stop word filtering** were performed.

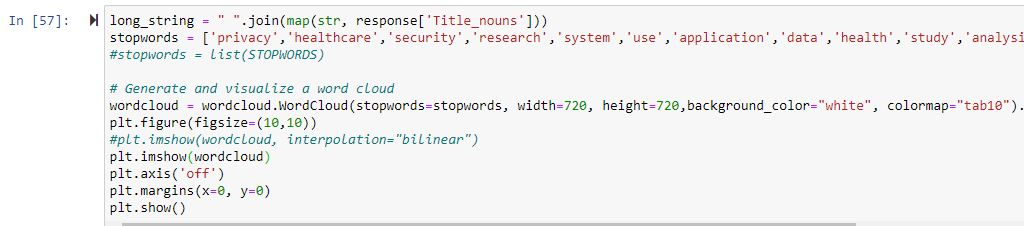






For the stop word filtering, a trial-and-error method was used to find the list of most repeating words that does not provide any meaning for the research. And these words were added to the standard set of English stop word filter available in the Word Cloud Python library. Once all these preprocessing steps were performed, a word cloud was produced to visualize the most frequently appearing words in the dataset.

This reveals that a lot of researchers were performed around studying the IoTs role in healthcare privacy, and a lot of survey-based studies were made on this domain. Several of the recent studies included in the dataset were about studying the effect of Covid-19 on healthcare security and privacy.





Based on this, a bar plot was created to understand the most frequently appearing words and the number of times they appeared. Based on this analysis we were able to get the following plot.

Chart, bar chart

Description automatically generated

## LDA Model Development and Training

The LDA topic model algorithm requires a document word matrix as the main input. In order to generate this document word matrix, we use Count Vectorizer python module in sklearn library. This Document-Word Matrix was produced through the following method.

Text

Description automatically generated

Once this matrix was generated, its sparsity index was studied to see how many cells in this matrix will be non-zero. Sparsity is nothing but the percentage of non-zero datapoints in the document-word matrix, that is vectorized\_data.

Graphical user interface, text, application

Description automatically generated

Then the LDA model was built using the SKLearn.decomposition Python Library through the LatentDirichletAllocation module. In the initial iteration, the number of topics (k-value) was set to 20. Thereafter the model with best number of topics and other hyper parameters was chosen through plotting the log-likelihood and perplexity factor.

Graphical user interface, text, application

Description automatically generated

|  |  |
| --- | --- |
| **Topic Number** | **Keywords** |
| Topic #0: | drive, wheel, anyone, bus, emergency, need, school, worth, department, vehicle |
| Topic #1: | disease, home, house, return, applications, white, intelligence, secure, money, prediction |
| Topic #2: | act, California, contact, trace, innovation, CCPA, protection, role, control, case |
| Topic #3: | story, consumers, year, company, technology, adoption, attitudes, adults, older, co |
| Topic #4: | 19, covid, technology, pandemic, report, protection, state, smart, IoT, framework |
| Topic #5: | year, name, bill, province, doctor, audit, specialist, Ohio, charge, stocks |
| Topic #6: | management, patient, centric, image, distributed, breach, framework, industry, authentication, implications |
| Topic #7: | federal, base, register, management, control, device, protection, risk, image, disease |
| Topic #8: | new, architecture, Ethiopia, legal, consumers, factors, practice, facilities, tracing, look |
| Topic #9: | new, IoT, information, case, plan, associate, poor, Australians, consumers, age |
| Topic #10: | state, power, family, community, project, world, area, baby, mother, point |
| Topic #11: | image, approach, trust, role, risk, attitudes, digital, mediation, applications, Ethiopia |
| Topic #12: | year, status, woman, information, nation, challenge, experience, cohort, treatment, drug |
| Topic #13: | systems, world, record, smart, preservation, IoT, marketers, rights, network, cluster |
| Topic #14: | survey, blockchain, design, address, Dubai, hospital, Indian, expat, Rahul, cricket |
| Topic #15: | information, policy, device, intelligence, trust, location, base, enterprise, house, mobile |
| Topic #16: | internet, things, online, utility, survey, life, technologies, trend, software, networks |
| Topic #17: | survey, systems, hospital, analytics, pattern, issue, core, patients, medical, policy |
| Topic #18: | IoT, education, strategy, experience, treatment, video, adult, emergency, comparison, industry |
| Topic #19: | network, law, service, effect, new, record, authentication, scheme, retrospective, mining |

## Validation of LDA Model Results

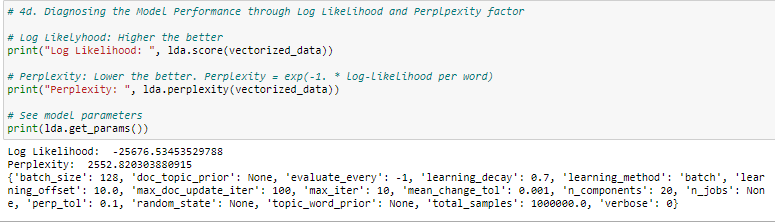
The very first step in evaluating the model performance is to study the log-likelihood and perplexity score. A model with higher log-likelihood and lower perplexity is considered to be good.

Perplexity Score is given by the formula,

Text

Description automatically generated

The measures of log-likelihood and perplexity for our model comes to be as follows.



Once this is found, our next interest is to find the best LDA model that would fit our analysis. To find this, GridSearch was performed. The most important tuning parameter for LDA models is n\_components (number of topics). The grid search constructs multiple LDA models for all possible combinations of param values in the param\_grid dictionary (which in our case contains only the n\_components hyperparameter).

Text

Description automatically generated

The best model was found to be the one with number of topics as 10.

Text

Description automatically generated

Plotting the log-likelihood scores against num\_topics, clearly shows number of topics = 5 has better scores. To tune this even further, a finer grid search for n\_topics between 5 & 10 can be done. But this is considered for future scope of work. Bottomline is that a lower number of distinct topics is reasonable for this dataset.

Chart, line chart

Description automatically generated

With this analysis, the 5 topics produced by our best fit LDA model was listed. Table below shows this information.

|  |  |  |
| --- | --- | --- |
| **Topic Number** | **Keywords** | **No. of Articles** |
| Topic #0: | year, information, act, California, new, protection, hospital, law, name, woman | 125 |
| Topic #1: | covid, 19, systems, survey, technology, network, state, disease, internet, authentication | 129 |
| Topic #2: | IoT, record, framework, applications, edge, preservation, service, blockchain, Dubai, trace | 100 |
| Topic #3: | federal, experience, home, register, program, women, education, naloxone, 19, covid | 91 |
| Topic #4: | perspective, base, world, role, consumers, management, technology, risk, new, age | 97 |

The number of documents in each topic was plotted for the best fit LDA model. The bar plot shows a maximum number of papers towards Topic 0 and Topic 1.

Chart, bar chart

Description automatically generated

Our next interest is to see the dominant topics in each document. To classify a document as belonging to a particular topic, a logical approach is to see which topic has the highest contribution to that document and assign it. In the table below, all major topics in a document are greened out and the most dominant topic are assigned in its own column. Table below shows a detailed look at the Topic classification for first 15 papers on our dataset. Likewise, the topic classification for all 542 papers in our dataset was made.

Table

Description automatically generated

Once all these are determined, our next interest would be to validate that the model we derived through grid search is in fact the best fit model. This can be checked through clustering analysis of topics derived in the best fit LDA. By measuring the intra-cluster and inter-cluster distances, we will be able to tell that the produced model is a best fit one. Intra-cluster topic distance is the measure of how close each topic within each cluster is to every other topic in its cluster. Inter-cluster distance is the measure of how close each cluster of topics is to other clusters. Models that produce relatively small intra-cluster distances and relatively large inter-cluster distances evaluate favorably because they appear to be doing a good job of grouping like topics with discrete characteristics.

Chart, bubble chart

Description automatically generated

And a K-means clustering of document-topic probability matrix shows the cluster of documents that share similar topic. This is shown in the plot below.

Chart, scatter chart

Description automatically generated

# Conclusion

Based on our research, it can be interpreted that several of the studies were performed in analyzing the safety and security of PHI data post COVID-19 situation. And a vast majority of papers were contributed towards studying the security and privacy of health care IT systems of California Hospitals. Equally distributed were the papers published on topics of new technology influence in the healthcare IT systems and the protection of PHI data with respect to IoT devices implementation in Health Care.

The future study of this research would continue with focus on development of a prediction model for predicting topics of unknown text input. And thereby optimizing the topic modelling algorithm that has been developed by passing various different datasets.

# References

1. Health report
2. Raman, A. (2007, November). Enforcing privacy through security in remote patient monitoring ecosystems. In 2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine (pp. 298-301). IEEE.
3. Appari, A., & Johnson, M. E. (2010). Information security and privacy in healthcare: current state of research. International journal of Internet and enterprise management, 6(4), 279-314.
4. Kripalani, S., LeFevre, F., Phillips, C. O., Williams, M. V., Basaviah, P., & Baker, D. W. (2007). Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. Jama, 297(8), 831-841.
5. Warkentin, M., Johnston, A. C., & Shropshire, J. (2011). The influence of the informal social learning environment on information privacy policy compliance efficacy and intention. European Journal of Information Systems, 20(3), 267-284.
6. Vaast, E. (2007). Danger is in the eye of the beholders: social representations of information systems security in healthcare. The Journal of Strategic Information Systems, 16(2), 130-152.
7. Munson, S. A., Cavusoglu, H., Frisch, L., & Fels, S. (2013). Sociotechnical challenges and progress in using social media for health. Journal of medical Internet research, 15(10), e226.
8. Liginlal, D., Sim, I., Khansa, L., & Fearn, P. (2009). Human Error and Privacy Breaches in Healthcare Organizations: Causes and Management Strategies. AMCIS 2009 Proceedings, 406.
9. Office of the National Coordinator for Health Information Technology. 'Individuals' Perceptions of the Privacy and Security of Medical Records and Health Information Exchange,' Health IT Quick-Stat #58. dashboard.healthit.gov/quickstats/pages/consumers-privacy-security-medical-record-information-exchange.php. June 2019. <https://dashboard.healthit.gov/quickstats/pages/consumers-privacy-security-medical-record-information-exchange.php>
10. Office of the National Coordinator for Health Information Technology. 'Individuals Use of Technology to Track Health Care Charges and Costs,' Health IT Quick-Stat #57. dashboard.healthit.gov/quickstats/pages/consumers-health-care-charges-costs-online.php. April 2018
11. Al Ameen, M., Liu, J., & Kwak, K. (2012). Security and privacy issues in wireless sensor networks for healthcare applications. Journal of medical systems, 36(1), 93-101.
12. <https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/#13compareldamodelperformancescores>
13. Ahmadi, H., Arji, G., Shahmoradi, L., Safdari, R., Nilashi, M., & Alizadeh, M. (2019). The application of internet of things in healthcare: a systematic literature review and classification. Universal Access in the Information Society, 18(4), 837-869.
14. Mehraeen, E., Ghazisaeedi, M., Farzi, J., & Mirshekari, S. (2017). Security challenges in healthcare cloud computing: a systematic review. Global Journal of Health Science, 9(3), 157-157.
15. Ahmed, A., Latif, R., Latif, S., Abbas, H., & Khan, F. A. (2018). Malicious insiders attack in IoT based multi-cloud e-healthcare environment: A systematic literature review. Multimedia Tools and Applications, 77(17), 21947-21965.
16. Nazir, S., Ali, Y., Ullah, N., & García-Magariño, I. (2019). Internet of things for healthcare using effects of mobile computing: a systematic literature review. Wireless Communications and Mobile Computing, 2019.
17. Yao, W., Chu, C. H., & Li, Z. (2012). The adoption and implementation of RFID technologies in healthcare: a literature review. Journal of medical systems, 36(6), 3507-3525.
18. Wamba, S. F., Anand, A., & Carter, L. (2013). A literature review of RFID-enabled healthcare applications and issues. International Journal of Information Management, 33(5), 875-891.
19. Box, D., & Pottas, D. (2013). Improving information security behaviour in the healthcare context. Procedia Technology, 9, 1093-1103.
20. Ermakova, T., Huenges, J., Erek, K., & Zarnekow, R. (2013). Cloud Computing in Healthcare–A literature review on current state of research.
21. Tandon, A., Dhir, A., Islam, N., & Mäntymäki, M. (2020). Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda. Computers in Industry, 122, 103290.
22. Behera, R. K., Bala, P. K., & Dhir, A. (2019). The emerging role of cognitive computing in healthcare: A systematic literature review. International journal of medical informatics, 129, 154-166.
23. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, 57-65.