Soc Indic Res
https://doi.org/10.1007/s11205-018-1837-z

Electronic Health Record Breaches as Social Indicators

Waldemar W. Koczkodaj[1] · **Mirosław Mazurek**[2] · **Dominik Strzałka**[2] · **Alicja Wolny-Dominiak**[3] · **Marc Woodbury-Smith**[4]

Accepted: 16 January 2018

Abstract This study presents compelling social indicators of such magnitude that they cannot be ignored. The statistical evidence shows that data breaches of electronic health records have taken place at an unprecedented scale. Currently, the number of individuals affected, as regulations of Health Insurance Portability and Accountability Act refers to us, has surpassed the half of the US population (some data is breached several times lowering the number of victims but increasing the possibility of being sold quickly). The data breaches are recorded and posted by Department of Health and Human Services.

Keywords Security breach · Office for Civil Rights · Department of Health and Human Services

1 Introduction

The main goal of this study is to make social science researchers and the public-at-large aware of issues regarding the privacy of their health-related data. The paper presents evidence that health data breaches have been taking place at an unprecedented level. Under HIPAA, a breach is defined as “the unauthorized acquisition, access, use or disclosure of protected health information (PHI) which compromises the security or privacy of such information”. The analyzed data show that the electronic health records (EHR) of at least 173 million entries have been breached since such data started being collected in October

✉ Waldemar W. Koczkodaj
wkoczkodaj@cs.laurentian.ca

[1] Computer Science, Laurentian University, 935 Ramsey Lake Rd, Sudbury, ON P3E 2C6, Canada

[2] Faculty of Electrical and Computer Engineering, Rzeszów University of Technology, Al. Powstańców Warszawy 12, 35-959 Rzeszow, Poland

[3] Department of Statistical and Mathematical Methods in Economics, University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland

[4] The Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

Published online: 03 February 2018

2009. The data beaches might have affected as many as one-third of the US population. Of note, the same EHRs might have been misappropriated by multiple perpetrators.

With the development of a market for stolen data and related hacking skills hospitals and other healthcare providers have become popular targets for hackers and cybercriminals. In June 2016 alone, more than 11 million healthcare records were exposed because of cyber-attacks. Indeed, the top three data security breaches were from the healthcare industry. According to a "Health Warning" report by the Intel Security McAfee Labs, cybercriminals are putting more time and resources into exploiting and monetizing health care data. Notably, the actions of perpetrators are becoming more and more aggressive (Nigrin 2014). Upon stealing medical records, perpetrators must analyze the data, cross-reference it with data from other sources before lucrative fraud, theft, extortion, or blackmail opportunities can be identified.

The US government has posted online the number of security data breaches recorded by the agency Health and Human Services (HHS Breach Portal 2017). Using data gathered by HHS in the USA, we have conducted statistical analysis to articulate these issues. Our empirical study shows how the confidentiality of EHRs is breached at a level that the computed social indicators deserve to be regarded as very problematic for the public. The "individuals affected", a term borrowed from the regulations of Health Insurance Portability and Accountability Act (HIPAA), have the right to know about the violations of privacy and the potentiality of having private health records sold. Healthcare records (in particular, medical) have to be protected. We argue that new laws should be passed and adequate penalties imposed.

2 Data Source, Definition of Terms and Methods

The security breach data have been collected by the Office for Civil Rights, Department of Health and Human Services (HHS) in the USA. The data collection is involuntary and regulated by Section 13402 of the Health Information Technology for Economic and Clinical Health (HITECH) Act which is a part of the American Recovery and Reinvestment Act of 2009 (ARRA). ARRA was enacted on February 17, 2009 by requiring HHS to issue interim final regulations within 180 days. Entities under the HIPPA of 1996 and their business associates are required to provide notification in the case of breaches of health data. HHS is requested to update its guidance specifying the technologies and methodologies that render protected health information unusable, unreadable, or indecipherable to unauthorized individuals.

Section 13402 of the Act regulates the breach notification process. It applies to HIPAA covered entities and their business associates that access, maintain, retain, modify, record, store, destroy, or otherwise hold, use, or disclose unsecured protected health information. The Act defines "covered entity", "business associate", and "protected health information" used in the HIPAA Administrative Simplification regulations (45 CFR parts 160, 162, and 164 HIPAA Rules) at §160.103. Under the HIPAA Rules, a covered entity is:

- a health plan,
- health care clearinghouse,
- or health care provider that transmits any health information electronically in connection with a covered transaction, such as submitting health care claims to a health plan.

A business associate, defined by the HIPAA Rules, is a person or service performing functions or activities on behalf of a covered entity. It involves the use or disclosure of individually identifiable health information. Business associates include third party administrators or pharmacy benefit managers involved in health plans, claims processing. Business associates may work in billing companies, transcription companies. They may also provide legal, actuarial, accounting, management, or administrative services for covered entities who require access to protected health data.

Section 13407(f)(3) stipulates that "unsecured personal health records" (PHR) are "identifiable health information" that is not protected through the use of a technology or methodology specified by the Secretary of HHS. Section 13402 of the Act requiring breach notification following the discovery of a breach of unsecured protected health information.

The HSS posts numerous data items that contain information about *the number of individuals affected (NIA) in one breach* and the total *number of breaches (NB)*. For our own analyses, we extracted relevant data from the HHS between October 2009 and April 2017 and using R prepared visualizations that more clearly articulate the issues we wish to illustrate. Frequency data are presented as bar charts with categories: location and breach type.

3 Results

Tables 1 and 2 show case counts (across top 10 breaches and those with more than one million stolen records) in successive years and covered entity types. Having this data it can be noted that since 2009 NIA is 173,398,820 and NB is 1863. It can be seen that the number of breaches (NB) is not rapidly growing but the amount of stolen data is growing rapidly.

Similarly, in a time series analysis of truncated HSS data (Fig. 1) a slow upward trend of NBs is observed. There are 19 top breaches with NIA higher than 1,000,000—the details of which are in Table 2. The chart in Fig. 1 shows NIA and the total NB in a month. The average time between health data breaches is approximately 1.5 day (since October 2009, the initial date of recording). In 1 month (November 2016), there were 38 data breaches for a total number of individuals affected of 776,797. The monthly average of NIA is 1,907,365 (nearly 2 million individuals).

The geolocation of top breaches is shown in Fig. 2. The map illustrates locations of the 19 largest security data breaches. In total, they account for nearly 142 million individuals affected. The top security data breaches are spread around USA without a clear hub or

Table 1 NIA and NB by years (only last 3 months in 2009 and the first 4 months of 2017 are recorded)

	Year	NIA	NB
1	2009	134,773	18
2	2010	5,932,276	199
3	2011	13,150,298	195
4	2012	2,808,042	201
5	2013	6,939,276	265
6	2014	12,682,073	289
7	2015	113,267,174	267
8	2016	16,655,952	328
9	2017	1,828,956	101
10	Total	173,398,820	1863

Table 2 19 Top breaches with the number of individuals affected higher than 1,000,000

State	Covered entity type	Breach submission date	Individuals affected
IN	Health plan	03/13/2015	78,800,000
WA	Health plan	03/17/2015	11,000,000
NY	Health plan	09/09/2015	10,000,000
VA	Business associate	11/04/2011	4,900,000
TN	Business associate	08/20/2014	4,500,000
CA	Healthcare provider	07/17/2015	4,500,000
IL	Healthcare provider	08/23/2013	4,029,530
IN	Business associate	07/23/2015	3,900,000
AZ	Healthcare provider	08/03/2016	3,620,000
NY	Business associate	08/09/2016	3,466,120
FL	Healthcare provider	03/04/2016	2,213,597
TX	Business associate	09/10/2014	2,000,000
NY	Business associate	04/14/2011	1,900,000
NJ	Business associate	02/11/2011	1,700,000
FL	Health plan	06/03/2010	1,220,000
MD	Health plan	05/20/2015	1,100,000
MT	Health plan	07/07/2014	1,062,509
FL	Healthcare provider	10/07/2011	1,055,489
TN	Health plan	11/01/2010	1,023,209
		Total	141,990,454

concentration. The total number of individuals affected in the 19 largest data breaches is 83% of the total number of individuals affected. Exploring these 83% to analyze how the NIA and the NB are distributed, the truncation criterion is the quartile on the order of 0.99 ($Q(0.99)$), which exactly excludes the observations from Table 2.

Figure 3 illustrates the scatterplot of the NIA for truncated data. The *outliers* are the data breaches with the NIA over $Q(0.99) = 382{,}494$. The red line shows the median $Q(0.5) = 2279$ of the truncated data and the orange line—the mean NIA equals 16,531. The scatterplot shows not only 19 outliers but also the evidence that there are many less significant breaches.

The distribution of the NIA for truncated data is strongly skewed to the right. It means that there are many data breaches with relatively lower number of NIA. Histograms in Fig. 4 illustrate the distribution of the NIA, their mean and median values as vertical dash lines.

The spatial analysis of truncated HSS data characterizes the most sensitive states. Choropleths in Figs. 5 and 6 show NIA and NB. Color saturation shows the value of NIA/NB. The lighter color means lower values.

Figure 7 (bar chart) shows the total NIA in one data breach group by the type of breach (category): Hacking/IT Incident (A); Improper Disposal (B); Loss (C); Other (D); Theft (E); Unauthorized Access/Disclosure (F); Unknown (G). This categorization has fundamental flaws. For instance, it shows that nearly 2,000,000 NIA are unknown, whereas ~ 6,800,000 NIA are categorized as “unauthorized”. But conversely, this implies that approx. 130,000,000 “hacking/IT incidents” may be authorized. What sound like a case

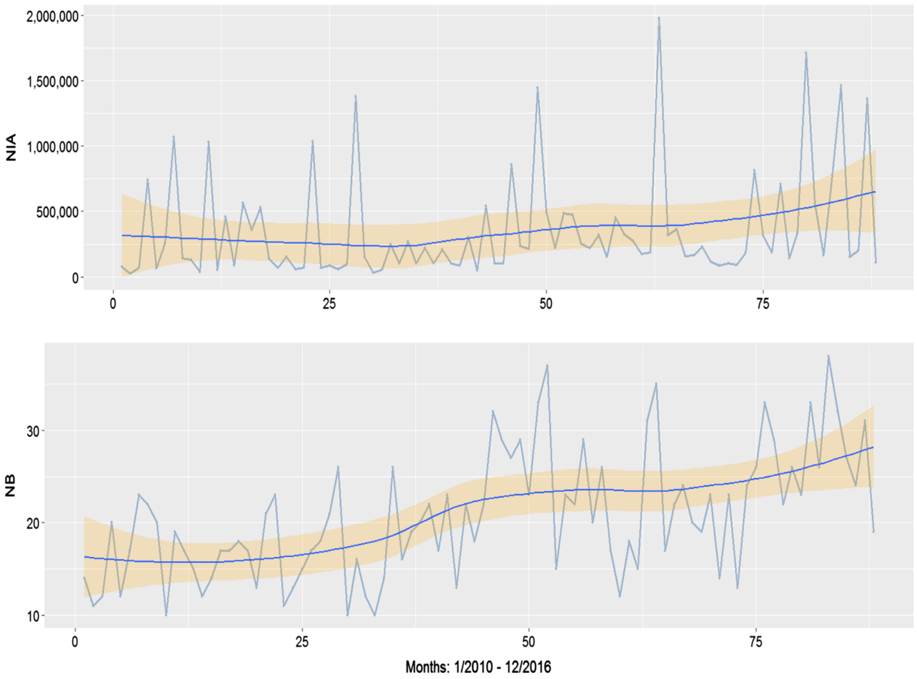

Fig. 1 Monthly time series of individuals affected and number of breaches

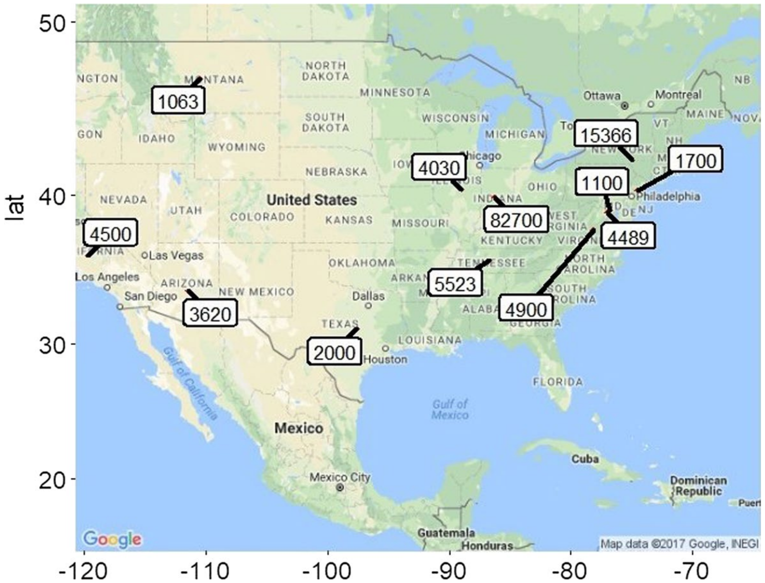

Fig. 2 Geolocation of total top health data breaches

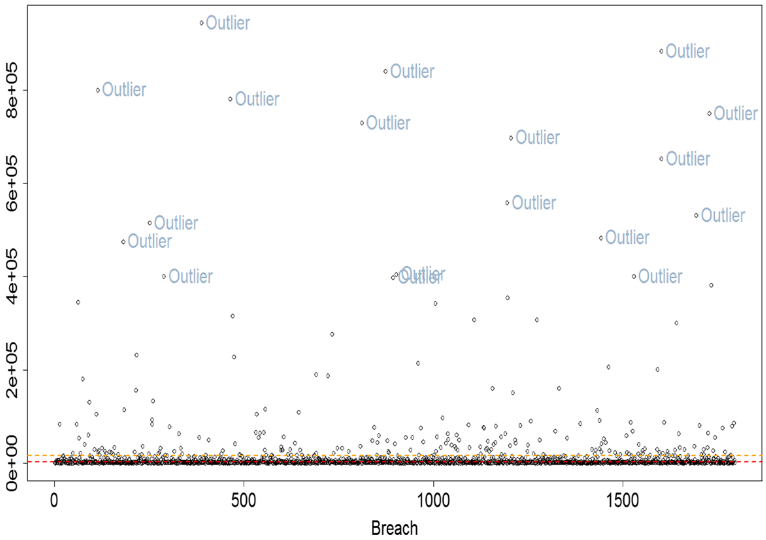

Fig. 3 Scatter plot of NIA for the largest data breaches (in 1000)

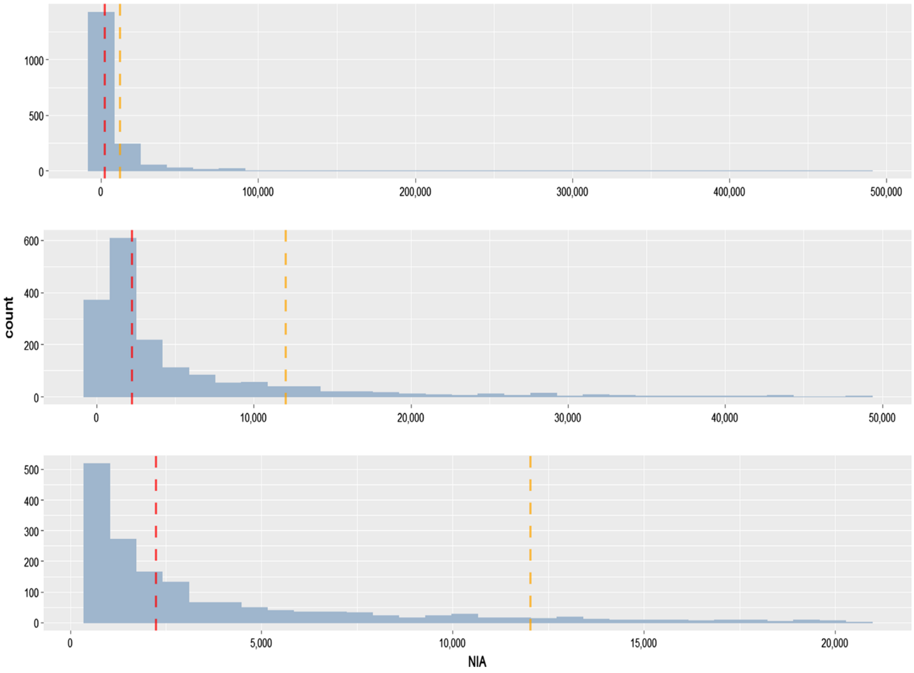

Fig. 4 Histograms of NIA for different truncation criteria

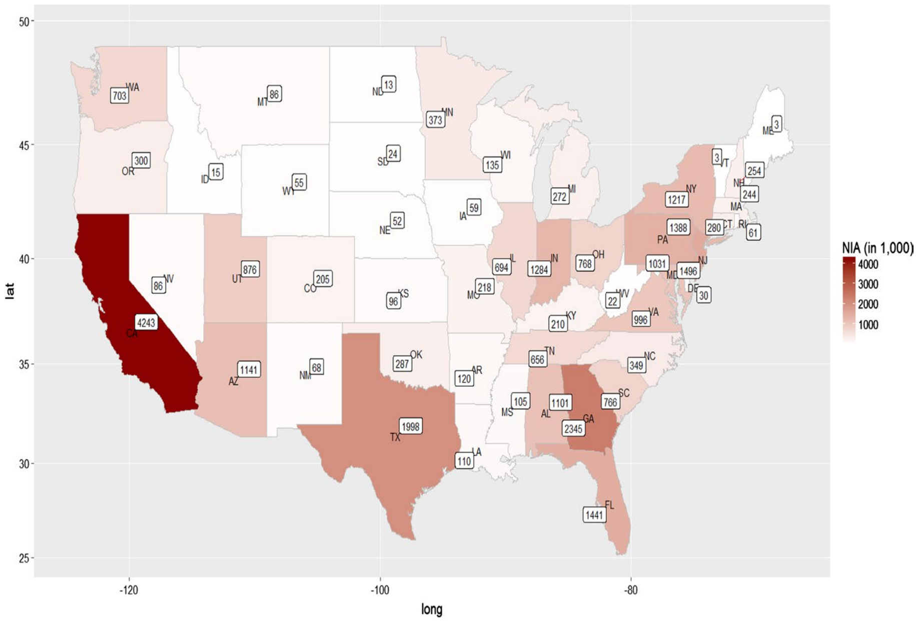

Fig. 5 Choropleth of number of individuals affected (in 1000)

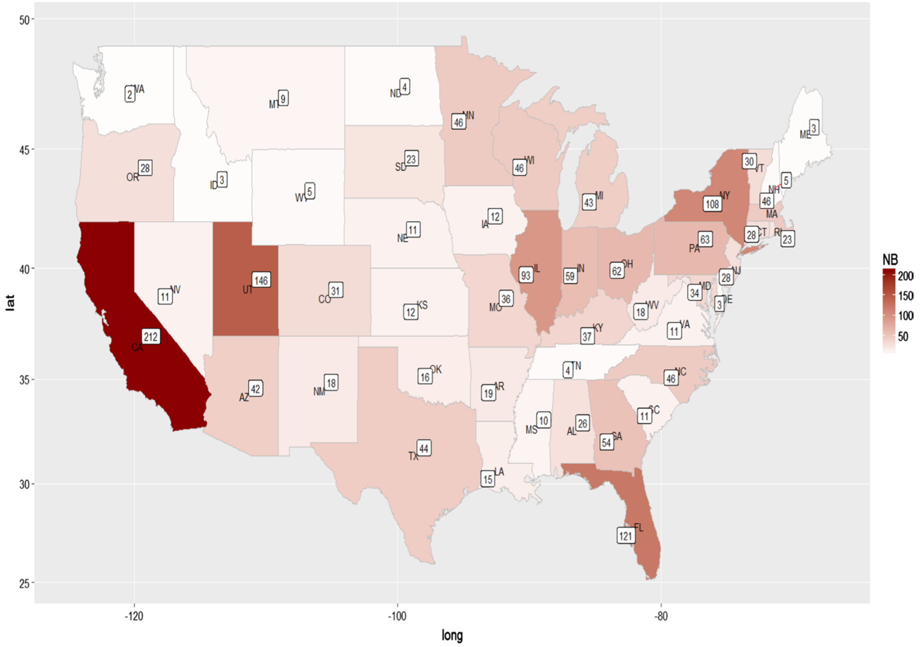

Fig. 6 Choropleth of number of breaches

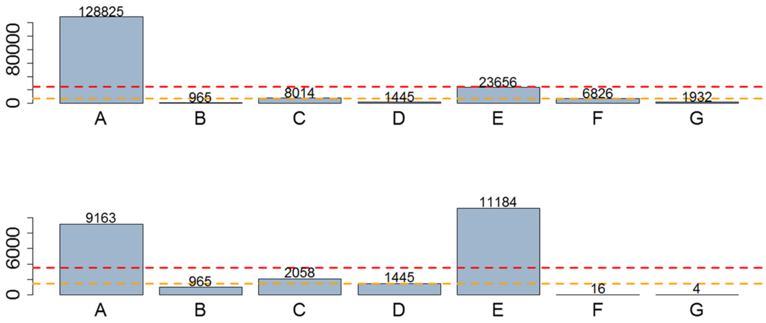

Fig. 7 The total number of individuals affected in main type of breach

of black humor, can also serve as an evidence that basic breach terminology has not been developed yet. Visualization of the analyzed data is provided for two cases: all data (top graph) and data without 19 top breaches with the number of individuals affected higher than 1,000,000 (bottom graph).

The bar chart in Fig. 8 shows the NB group by the type of breach: Hacking/IT Incident (A); Improper Disposal (B); Loss (C); Other (D); Theft (E); Unauthorized Access/Disclosure (F); Unknown (G). The highest numbers of NIA took place in the theft category; it represents 42.39% of NB (for all data). The lowest numbers of breaches took place in the Unknown category reaching only 0.66%.

4 Discussion

A steady increase in security breaches of data processing systems has been reported in numerous studies (Brennan et al. 1991), and our own data analyses shed light on the true extent of the problem. The need for patients to protect themselves and their families from harm, and for hospitals to make patient safety a priority is evident. According to Leape et al. (1991) and Amante et al. (2015) the statistics show that many hospitals are making headway in addressing errors, accidents, injuries and infections that kill or hurt patients,

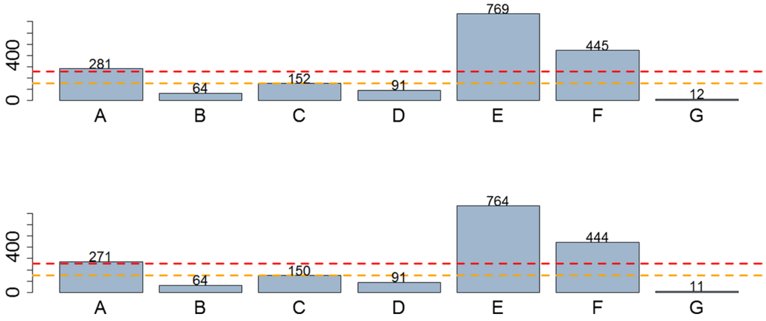

Fig. 8 The total number of breaches by the type of breach

but overall progress is not impressive. The trend recorded by our analyses suggests a far bigger strategy is required to avert a potential crisis.

If exists a possibility of having direct or indirect access to resources, there always exists enticement to breach these resources and make possible use of them for financial benefit. Medical data are not perishable, which makes them particularly valuable compared to financial data that can quickly become unusable after being stolen. For example, being aware of loss of such data, the customers should quickly change their credit card numbers—and usually they do this.

However, data security is a complex issue. Since the Internet has become the most convenient, fastest, and cheapest way to access data, EHR breaches is not only the matter of fast Internet growth but the problem of Internet security and, in general, IT systems security (Crotty and Mostaghimi 2014). Health data breaches can occur for different reasons. The software itself may be vulnerable, leading to computer hacking, and unauthorized access. Similarly, users of EHR records may create system vulnerability through failing to log off, or using 'easy' passwords. Covered entities and their business associates need to ensure they have a comprehensive data security plan, and are able to implement the necessary physical, administrative, and technical safeguards. To successfully respond to incidents, we need to minimize the number and severity of security incidents, assemble the core Computer Security Incident Response Team (CSIRT), define an incident response plan and contain the damage and minimize risks.

Healthcare executives must work closely with IT to come up with a strategy that takes the latest threats into account. An important issue to resolve is the effect of Internet security breach announcements on market value. Any information that leaks into the network poses a major threat to the capital markets, companies and may be a source of speculation on the stock markets. With the increase of the dynamics of networks interconnection, security issues become a critical point that needs to be considered. The widely adopted solution considers a mix of routers, switches, firewalls and virtual private networks (VPNs) together with the deployment of intrusion detection systems (IDSs) and vulnerability assessment tools. The network security assessment instrument is a comprehensive set of tools that can be used individually or collectively to ensure the security of network aware software applications and systems. As attacks upon critical network infrastructures increase in complexity and destructiveness, new methods are needed to aid security administrators in protecting their networks. There are many models for protecting IT systems and networks, but evidently they are not very effective.

The resolution of this lack of trust relating to the use of the Internet, particularly orientated towards its commercial use and on-line purchasing, requires website developers to create and maintain web applications that are robust and provide a certain degree of resilience to attack from outside threats. According to O'Connor (2011): "We found that half of states have no statutes addressing nondisclosure of personally identifiable health information generally held by public health agencies".

A properly secured processing system should use high-tech security tools to protect patient data. There are many technical solutions such as: data access monitoring, security event and information management (SIEM) systems, tokenization or very popular solutions—cloud security gateways. Adding extra layers of security makes it difficult for hackers to break into security systems, and can mitigate some of the effects of human error on your data security.

Our study calls for public awareness of potentially very dangerous consequences of the frivolous use of the information technology including taking sensitive data on notebooks, tablets, or USB memory sticks from hospital in the hope of working on these data at home although it is against regulations not only HIPAA but most hospitals. As the studies of Brennan et al. (1991) and Leape et al. (1991) have drastically changed irresponsible cover up attitude to remedial actions, our collaboration's dream is that this study may contribute to similar phenomena.

Beyond the bigger picture articulated in the preceding paragraphs, it is also important to bear in mind the impact of data breaches on the individual. For example, a lack of data security will impact on the well-being of users, and consequently their quality of life. The study of Senol-Durak and Durak (2011) clearly evidenced that trust in the Internet impacts on social outcome measures such as quality of life. There appears to be a common perception among the computer using community of a global lack of trust when using the Internet. It is vital that research is conducted to understand the role of the Internet on quality of life (QoL). Study presented in Lee et al. (2011) seeks to understand the role of the Internet. Specifically, it examines the question of whether Internet communication serves, like face-to-face interactions, to enhance quality of life or is it a threat to privacy. In Maggino and Faciotti (2017), the authors noted that: "What should be pointed out is that quality of life studies not only are focused on the present time but have also long term perspectives".

5 Conclusion and Future Actions

In this study, statistical evidence has been presented that show health data breaches occurring at an unprecedented level. Moreover, the data are from The Office for Civil Rights, Department of Health and Human Services (HSS) and therefore are likely credible. Preventing illegal breaches of EHR, currently taking place at such 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). A rating scale is required (mentioned in Koczkodaj et al. 2017) for data security for system users (rather than software developers) to assess this situation. Stiffer laws and penalties should be in place. And public awareness must be increased.

Perpetrators will never return or destroy stolen records. One million of stolen healthcare records may be sold for at much as $30,000 and the largest data breach is 78.8 million records (current and former members and employees of Anthem industry). A mass action is needed for "the law and order" to be restored.

6 Additional Information

The presented data analysis has been obtained by the open source R project for statistical computing. The details of notation and capabilities are described in Goksuluk et al. (2016), Bilgic and Susmann (2013), Bivand and Lewin-Koh (2017) and R Core Team (2016).

Acknowledgements We thank Grant O. Duncan, Team Lead, Business Intelligence, Integration and Development, Health Sciences North, (Sudbury, Ontario, Canada) and Martha O'Kennon from Albion College for reading the first draft.

References

Amante, D. J., Hogan, T. P., & Pagoto, S. L. (2015). Access to care and use of the internet to search for health information: Results from the US National Health interview survey. *Journal of Medical Internet Research, 17,* e106.

Bilgic, Y. K., & Susmann, H. (2013). rlme: An R package for rank-based estimation and prediction in random effects nested models. *The R Journal, 5*(2), 71–79.

Bivand, R., & Lewin-Koh, N. (2017). *Maptools: Tools for reading and handling spatial objects*. R package version 0.9-2.

Brennan, T. A., Leape, L. L., Laird, N. M., Hebert, L., Localio, A. R., Lawthers, A. G., et al. (1991). Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. *The New England Journal of Medicine, 324*(6), 370–376.

Crotty, B. H., & Mostaghimi, A. (2014). Confidentiality in the digital age. *BMJ-British Medical Journal, 348,* g2943.

Goksuluk, D., Korkmaz, S., Zararsiz, G., & Karaagaoglu, A. E. (2016). easyROC: An interactive web-tool for roc curve analysis using R language environment. *The R Journal*, *8*(2), 213–230.

HITECH—Health Information Technology for Economic and Clinical Health Act. (2009). *American Recovery and Reinvestment Act, Public Law* (pp. 111–115).

HHS Breach Portal. (2017). https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. Accessed May, 2017.

Koczkodaj, W. W., Kakiashvili, T., Szymańska, A., Montero-Marin, J., Araya, R., Garcia-Campayo, J., et al. (2017). How to reduce the number of rating scale items without predictability loss? *Scientometrics, 111*(2), 581–593.

Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., et al. (1991). The nature of adverse events in hospitalized patients: Results of the Harvard Medical Practice Study II. *The New England Journal of Medicine, 324,* 377–384.

Lee, P., Leung, L., Lo, V., Xiong, C., & Wu, T. (2011). Internet communication versus face-to-face interaction in quality of life. *Social Indicators Research, 100*(3), 375–389.

Maggino, F., & Facioni, C. (2017). Measuring stability and change: Methodological issues in quality of life studies. *Social Indicators Research, 130,* 161–187.

Nigrin, D. J. (2014). When 'hacktivists' target your hospital. *The New England Journal of Medicine, 371,* 393.

O'Connor, J. (2011). Informational privacy, public health, and state laws. *American Journal of Public Health, 101*(1845–1850), 2011.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Senol-Durak, E., & Durak, M. (2011). 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, 103*(1), 23–32.

UIC Health Informatics. (2017). *Why data security is the biggest concern of health care*. https://healthinformatics.uic.edu/resources/articles/why-data-security-is-the-biggest-concern-of-health-care/.