J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Columbus, OH, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Technical Note: Software

Generating unique IDs from patient identification data using security models

Emad A. Mohammed[1,2,3], Jonathan C. Slack[2,3], Christopher T. Naugler[2,3]

[1]Department of Electrical and Computer Engineering, Schulich School of Engineering, University of Calgary, Departments of [2]Pathology and Laboratory Medicine and [3]Family Medicine, University of Calgary, Calgary Laboratory Services, Calgary, AB, Canada

E-mail: *Dr. Christopher T. Naugler - christopher.naugler@cls.ab.ca
*Corresponding author

Received: 12 October 2016 Accepted: 25 November 2016 Published: 30 December 2016

Abstract

Background: The use of electronic health records (EHRs) has continued to increase within healthcare systems in the developed and developing nations. EHRs allow for increased patient safety, grant patients easier access to their medical records, and offer a wealth of data to researchers. However, various bioethical, financial, logistical, and information security considerations must be addressed while transitioning to an EHR system. The need to encrypt private patient information for data sharing is one of the foremost challenges faced by health information technology. **Method:** We describe the usage of the message digest-5 (MD5) and secure hashing algorithm (SHA) as methods for encrypting electronic medical data. In particular, we present an application of the MD5 and SHA-1 algorithms in encrypting a composite message from private patient information. **Results:** The results show that the composite message can be used to create a unique one-way encrypted ID per patient record that can be used for data sharing. **Conclusion:** The described software tool can be used to share patient EMRs between practitioners without revealing patients identifiable data.

Key words: Electronic medical record security, message digest, patient private information encryption, secure hashing algorithm, security models

Access this article online

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.197203

Quick Response Code:

INTRODUCTION

The use of electronic health records (EHRs) has expanded in response to the burgeoning financial, administrative, and technological demands associated with modern health care. The first generation of EHRs was restricted to simple electronic medical records (EMRs), digital copies of paper charts limited to a single physician or hospital, which were stored in centralized in-house databases.[1] Over time, these simple EMRs progressed into EHRs that combined multiple EMRs with patient information such as allergies, prescriptions, contact information, and laboratory information.[1-3] The World Health Organization has defined the ideal EHR as one's entire health care record, which is continually updated over the course of their life, by all healthcare providers in all contexts.[1]

Multiple benefits of EHR adoption have been identified and include increased administrative efficiency, improved patient safety, decreased costs, easier data collection for research, more complete documentation, and increased ability of patient to access their healthcare information.[2,4,5] Multiple instances of increases in administrative efficiencies

This article may be cited as:
Mohammed EA, Slack JC, Naugler CT. Generating unique IDs from patient identification data using security models. J Pathol Inform 2016;7:55.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2016/7/1/55/197203

have been noted with the use of EHRs. For example, an EHR system with barcode readers was observed to take 97% less time to locate patient records when compared to using paper charts.[5] In addition, an EHR in Uganda was associated with a 91% reduction in costs.[5] EHRs increase patient safety by reducing duplicitous laboratory tests; documenting allergic reactions and other elements of patient history; enhancing communication between care providers; providing integrated, point-of-care clinical decision-making tools; reducing inappropriate antibiotic use; and alerting the user to possible drug–drug interactions.[2-4,6-8] A recent cost–benefit analysis in Europe suggests that EHR usage allows researchers to identify and enroll patients in clinical trials faster, better determine research protocol feasibility, and provide data that can be analyzed to see if patient safety outcomes were met.[4] A small study of breast cancer patients showed that patient's anxiety was decreased when they had access to their own healthcare information.[9]

Despite the aforementioned benefits of EHRs, three key barriers have limited their adoption.[10] First, the large upfront cost has deterred both individual users and countries unable to afford it. Ninety-one percent of health centers without EHRs cited lack of capital as the most important barrier to adoption.[11] Upfront adoption costs were noted to range from $16000 to $36000 per physician.[12] In another study, 86.7% of Canadian hospital managers also cited financial resources as the main barrier to providing patients access to their own EHR.[13] The second main barrier identified was privacy and security concerns.[1,2,10,13-16] In a survey of 309 physicians who were nonusers of EHRs, 55.3% stated that privacy or security concerns were a barrier to EHR adoption.[14] Patients also reported privacy of their health information as a primary concern.[15] The third major barrier to widespread adoption of EHRs is lack of trained employees to create and maintain the system.[1,10,13]

The aforementioned obstacles have largely constrained the adoption of EHRs to physicians, hospitals, and countries wealthy enough to afford the initial investment, employee training, and software to address security and privacy concerns. Therefore, EHRs have the potential to serve as a source of health disparity, their usage reserved to those able to pay for them, and depriving those unable to afford the initial investment of the long-term savings and patient safety benefits.[6,7,11,16,17] Those not wealthy enough to pay will get second-tier access to their healthcare information and may subsequently have diminished safety. This represents a potential violation of the basic bioethical tenets of autonomy and beneficence.[18] EHRs that do not align with these fundamental bioethical principles are unlikely to be successful as evidenced by the failure of an attempt to create a nationwide EHR in the UK.[1,19]

As previously mentioned, security and privacy concerns are one of the fundamental barriers to the adoption of EHRs. Various initiatives to address security concerns pertaining to EHR have been undertaken. The ISO/TS 18308 standard defined the secure storage and communication of health information as a fundamental component of an EHR.[20] The International Medical Informatics Association was established to address these security and privacy issues and contributed in creating guidelines and educational training program to address the concerns of healthcare providers, managers, biomedical, health informatics specialists to the confidentiality, privacy, and security of patient data.[21] The Advanced Informatics in Medicine/Secure Environment for Information Systems in MEDicine project has taken into account the traditional and proved principles of healthcare data processing, the various regulations within the European Union, the enormous and subtle risks of healthcare information technology systems, the cost of changing existing technology, and the mandatory need for encryption software to keep patient information secure from different privacy violations during data sharing.[22]

Until such time that the universal adoption of high-level EHRs is a reality, there exists a need to handle pathology and laboratory data files in such a way that privacy is not breached. Data files with test results, patient names, sex, and birthdate are commonly generated but may lack a unique identifier that can be used to anonymize the data. For example, the Mosoriot Medical Record System, an EMR system developed for a primary care center in rural Kenya, required the creation of unique patient identifiers, as Kenya lacks the equivalent of a social insurance number.[23] In this paper, we describe an open-source tool to encrypt private patient information using MD5 and secure hashing algorithm (SHA)-1 implemented in R statistical software package (R digest package, Version 0.6.10, https://CRAN.R-project.org/package=digest).[24]

Cryptography is the process of storing and sending information in a secure manner that limits access to intended recipients.[25]

The basic goals of cryptography are as follows:[25]

- Confidentiality/privacy: Ensuring that only the intended receiver is able to read the message
- Data integrity: Ensuring that the message content received is not altered during sharing process
- Authentication: Identifying the intended recipients.

Message digest (MD) is a security model that generates a unique code for the purpose of providing a message authentication code.[26] MD5 and SHA[26] are one-way hashing functions (security models), which are easy to generate but are harder to reverse.

MESSAGE DIGEST 5 ALGORITHM

An MD is a cryptographic hash function encompassing a string of digits created by a one-way hashing function to protect the integrity of exchanged data. The original MD algorithm (MD1) was shortly followed by a modified version (MD2).[27] However, MD2 was soon found to be quite weak and shortly followed by MD3, which however was never released. MD3 was further developed and MD4[27] was released; however, it was unsatisfactory, but it provided the theoretical foundations for MD5 and SHA-0.[27] MD5 produces 128-bit MD from input messages of variable length. MD5 operates iteratively on all message subblocks as explained in the following:

Step 1: Preprocessing (Padding, Block Preparation, and Initialization)

A processed message is padded such that its length (in bits) is corresponding to 448 mod 512. Shorter messages are padded with the first bit set to "1" and all the rest set to zero. The message length is then appended to the original message in the remaining 64 bits to form a block of 512 bits. MD5 operates on two inputs: the input message block and the output hash from the previous MD. In the first step, the initial hash values are constants provided by the algorithm. The initial values for MD5 are provided into four 32-bit words. A four-word buffer is used to store those values which are then replaced by the output hash values after each step.

Step 2: Length Attaching

A 64-bit delineation of the length of the message before the padding is attached to the result of the previous step. The resulting message has a length that is exactly a multiple of 512 bits.

Step 3: Initialize Message Digest Buffer

A four-word buffer (A, B, C, D) is used to compute the MD. Here, each of A, B, C, D is a 32-bit register.

Step 4: Process Message in 16-Word Block

Four auxiliary functions are defined to process the three 32-bit words and produce as output one 32-bit word.

Step 5: Output

The MD output is the processed words, A, B, C, D, with the low-order byte of A and end with the high-order byte of D.

SECURE HASHING ALGORITHM

The SHA algorithm is a cryptography hash function and used in digital certificate and data integrity.[26] The MD output is calculated using the final padded message as "*n*" 512-bit blocks. The algorithm makes use of two 160-bit registers, each consisting of five 32-bit sub-registers. The basic SHA-1 algorithm is described as follows:

- Step 1: The algorithm starts by initializing the five sub-registers of the first 160-bit register
- Step 2: SHA-1 iterates through each of the 512-bit message blocks and updates the 160-bit register by binary manipulation of the message blocks. The SHA-1 algorithm copies the 160-bit register into the second register
- Step 3: This step involves a sequence of four rounds, each round takes as input the current value of register X and the blocks for that interval and operates upon them for 20 iterations
- Step 4: Once all four rounds of operations are completed, the second 160-bit register (A, B, C, D, E) is added to the first 160-bit register
- Step 5: Once the algorithm has processed all of the 512-bit blocks, the final output of X becomes the 160-bit MD.

IMPLEMENTATION AND VALIDATION

The validation data used were supplied by the Department of Pathology and Laboratory Medicine, University of Calgary and Calgary Laboratory Services, Calgary, AB, Canada. The validation data have 1,205,973 patient records, each of which has patient identification information, i.e., first name, middle name, last name, gender, and date of birth (DOB), in addition to clinical laboratory test results. We bind the patient's DOB, gender, and last name to form a composite identification field per record that is encrypted using the MD5 and SHA-1 algorithms.

We compare the uniqueness of the composite ID to the corresponding encrypted ID and the results show that the encrypted composite message can be used as a new patient ID to share patient EMRs among practitioners. However, faulty data entry may cause inconsistency in the encrypted IDs due to last name change from single to married names, gender change, in case of twin patients, and other data entry errors that may generate different composite ID for the same patient.

Availability

The encryption tool is freely available from the authors. The software can be accessed online through the following link: https://github.com/ClinicalLaboratory/Generating-Unique-IDs-from-Pateint-identification-Data-Using-Security-Models

Using the Software

To us, the encryption tool, R and RStudio, must be installed on the machine that has the patient record file. Place the provided R code file (UIDGen.R) in the folder where the data file is located. Open RStudio and then open the downloaded R code file. Change the path to your file as outlined in code, press Ctrl + A to select all the code, and finally, press Ctrl + Enter to run the code. The execution time may vary depending on your file size, the encryption algorithm selected, and the processing

platform. Both algorithms, i.e., MD5 and SHA-1, are called in the R code file and their output are attached to the original composite message and the user is free to choose the encrypted message that is most suitable to use for patient record sharing.

Validation Results

These encryption algorithms were applied to the validation dataset of 1,205,973 patient records. As expected, both algorithms resulted in no duplicated identifiers for different patients. Furthermore, in all instances when the same patient had multiple records, both algorithms always generated a single unique identifier for that patient.

DISCUSSION AND CONCLUSION

Designing a secure EHR sharing environment has attracted a lot of attention within healthcare industry and academic community. However, this extensively mandates the need for security models to assure the privacy of patient identification information.

A hash function receives a variable length message and produces a fixed-length digested message as its output. It is estimated that the efforts of coming up with two messages having the same MD are on the order of 2^{64} computations and that the difficulty of coming up with any message having a given MD is on the order of 2^{128} operations.[26,27]

The SHA-1 algorithm is used the Digital Signature Algorithm for digital signatures. The SHA-1 algorithm belongs to a set of cryptographic hash functions similar to the MD family. However, the main difference between the SHA-1 and the MD family is the more frequent use of input bits during the hash function in the SHA-1 algorithm than in MD4 or MD5. This fact results in SHA-1 being more secured compared to MD4 or MD5 but at the expense of slower execution.[26]

A major barrier to the adoption of EHRs in developing countries has been the perception that they are not secure.[1,23] However, when adequate policies and technologies are implemented, EHRs have several security advantages over paper records.[28] A trail of who has accessed the record can easily be created, and partial access can be controlled on a need to know basis.[28] This is often extremely important in developing countries as patients may face severe financial, social, and psychological ramifications if their private health information is disclosed. One such example is the significant stigma patients face if their HIV status is revealed to their community.[29] This software can be used as a cost-effective method of generating encrypted patient identifiers from data sets in limited-resource settings. EMRs in resource-limited settings may use spreadsheet or access-based datasets, and our software tool could be used to easily generate anonymized patient identifiers in these settings.[23,29,30] Similarly, other applications of this software could be to anonymize data sets assembled from manual chart reviews or historical data sets.

The open-source software presented in this paper can be used solve identity (private patient information) encryption concerns in many different settings amongst them are as follows:

- EMRs in resource-limited settings that are stored or created as spreadsheets.[31-33]
- Clinical data sets assembled from manual chart reviews.[34,35]
- Historical data sets created before modern EMRs.[33]

In these settings, the presented software tool can be both time and cost effective to encrypt the extracted data and/or EMRs. Moreover, the tool does not require trained personnel to use, which is not the case in many modern EMR systems.

Message Digest 5 and Secure Hashing Algorithm-1 Limitations

It is observed that if the input size is increased the program became slower performing the SHA-1 than the MD5 algorithm. The SHA-1 algorithm is claimed to be secure because it is practically infeasible to compute the message corresponding to a given MD.[26] Furthermore, it is extremely improbable to detect two messages hashing to the same value.[26] Both algorithms performed as expected, with the SHA-1 being slightly slower but believed to be more secure than MD5. Therefore, we suggest that when computing capability or time is not a concern that SHA-1 may be better to use than MD5 as it may be more secure than MD5 due to the more bits used in the encrypted output message (ID).

The encryption algorithms will produce the same identifier for different patients due to data entry errors or in some situations where the personal data are the same for different patients. One suggested solution for these situations is to sort out the EHR records before encryption by name, gender, and DOB. This will group the identical data records next to each other and the encryption algorithm can check the similarity of the generated identifiers and create a count field for the replica of the type unsigned long integer that is composed of four bytes. The unsigned long integer can accommodate a count start from 0 to 4,294,967,295 that is big enough to accommodate any possible duplicates in the EHR data. This will assure one-to-one mapping between the patient personal data and the generated identifier even in the case of multiple patients with the same personal data. However, this solution is computationally expensive and may require distributed processing to handle the massive data due to several sort and count updates.

It worth noting that after the identifier is generated by the encryption algorithms, different physicians can have

access to the patients' medical records, for example, vital signs, and can track these records locally, i.e., between physicians; however, the original personal data will remain secure.

Acknowledgements

The authors would like to thank the Canadian Institutes of Health Research Foundation Scheme for their continuous support.

Financial Support and Sponsorship

This work is supported and funded by a Canadian Institutes of Health Research Foundation Scheme grant to CN.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. World Health Organization. Electronic Health Records: A Manual for Developing Countries. Manila: WHO; 2006.
2. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: An evaluation of information system impact and contingency factors. Int J Med Inform 2014;83:779-96.
3. Peters SG, Khan MA. Electronic health records: Current and future use. J Comp Eff Res 2014;3:515-22.
4. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, *et al.* Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. Contemp Clin Trials 2016;46:85-91.
5. Blaya JA, Fraser HS, Holt B. E-health technologies show promise in developing countries. Health Aff (Millwood) 2010;29:244-51.
6. Butler MJ, Harootunian G, Johnson WG. Are low income patients receiving the benefits of electronic health records? A statewide survey. Health Informatics J 2013;19:91-100.
7. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, *et al.* Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood) 2005;24:1103-17.
8. Thurston J. Meaningful use of electronic health records. J Nurse Pract 2010;10:510-3.
9. Wiljer D, Leonard KJ, Urowitz S, Apatu E, Massey C, Quartey NK, *et al.* The anxious wait: Assessing the impact of patient accessible EHRs for breast cancer patients. BMC Med Inform Decis Mak 2010;10:46.
10. Cripps H, Standing C. The implementation of electronic health records: A case study of bush computing the Ngaanyatjarra lands. Int J Med Inform 2011;80:841-8.
11. Shields AE, Shin P, Leu MG, Levy DE, Betancourt RM, Hawkins D, *et al.* Adoption of health information technology in community health centers: Results of a national survey. Health Aff (Millwood) 2007;26:1373-83.
12. Miller RH, Sim I. Physicians' use of electronic medical records: Barriers and solutions. Health Aff (Millwood) 2004;23:116-26.
13. Urowitz S, Wiljer D, Apatu E, Eysenbach G, Delenardo C, Harth T, *et al.* Is Canada ready for patient accessible electronic health records? A national scan. BMC Med Inform Decis Mak 2008;8:33.
14. Kaushal R, Bates DW, Jenter CA, Mills SA, Volk LA, Burdick E, *et al.* Imminent adopters of electronic health records in ambulatory care. Inform Prim Care 2009;17:7-15.
15. Mandl KD, Szolovits P, Kohane IS. Public standards and patients' control: How to keep electronic medical records accessible but private. BMJ 2001;322:283-7.
16. Willyard C. Focus on electronic health records. Electronic records pose dilemma in developing countries. Nat Med 2010;16:249.
17. Hing E, Burt CW. Are there patient disparities when electronic health records are adopted? J Health Care Poor Underserved 2009;20:473-88.
18. Meslin EM, Schwartz PH. How bioethics principles can aid design of electronic health records to accommodate patient granular control. J Gen Intern Med 2015;30 Suppl 1:S3-6.
19. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: Case study of HealthSpace. BMJ 2010;341:c5814.
20. Schloeffel P. Requirements for an Electronic Health Record Architecture. ISO TS 18308. Washington, DC 20036-4910, USA: International Organisation for Standardisation; 2002.
21. Haux R. Medical informatics: Past, present, future. Int J Med Inform 2010;79:599-610.
22. Katsikas SK. Health care management and information systems security: Awareness, training or education? Int J Med Inform 2000;60:129-35.
23. Rotich JK, Hannan TJ, Smith FE, Bii J, Odero WW, Vu N, *et al.* Installing and implementing a computer-based patient record system in Sub-Saharan Africa: The Mosoriot Medical Record System. J Am Med Inform Assoc 2003;10:295-303.
24. The R Project for Statistical Computing. Available from: http://www.r-project.org/. [Last accessed on 2015 Feb 20].
25. Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: 2010 IEEE 3rd International Conference on Cloud Computing. Washington, DC 20036-4910 USA: IEEE; 2010. p. 268-75.
26. Aggarwal S, Goyal N, Aggarwal K. A review of comparative study of MD5 and SHA security algorithm. Int J Comput Appl 2014;104:1-4.
27. Rodríguez-Henríquez F, Saqib NA, Pérez AD. Cryptographic Algorithms on Reconfigurable Hardware. New York, NY 10013, USA: Springer Science & Business Media; 2007.
28. Barrows RC Jr., Clayton PD. Privacy, confidentiality, and electronic medical records. J Am Med Inform Assoc 1996;3:139-48.
29. Fraser HS, Biondich P, Moodley D, Choi S, Mamlin BW, Szolovits P. Implementing electronic medical record systems in developing countries. Inform Prim Care 2005;13:83-95.
30. Kalogriopoulos NA, Baran J, Nimunkar AJ, Webster JG. Electronic medical record systems for developing countries: Review. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Washington, DC 20036-4910 USA: IEEE publications; 2009. p. 1730-3.
31. Ohuabunwa EC, Sun J, Jubanyik KJ, Wallis LA. Electronic medical records in low to middle income countries: The case of Khayelitsha Hospital, South Africa. Afr J Emerg Med 2016;6:38-43.
32. Millard PS, Bru J, Berger CA. Open-source point-of-care electronic medical records for use in resource-limited settings: Systematic review and questionnaire surveys. BMJ Open 2012;2. pii: E000690.
33. Jawhari B. Benefits and Challenges of Implementing an Electronic Medical Record System in an Urban Slum in Kenya. University of Alberta; 2016.
34. Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers M, Weinstein RA. Development of a clinical data warehouse for hospital infection control. J Am Med Inform Assoc 2003;10:454-62.
35. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health 2007;97:654-8.