patient's clinical condition that other data types cannot (53). However, de-identifying free text data is more complicated than de-identifying tabular data because there is no schema similar to what databases have and no well-defined structure of records and fields. With no controlled vocabulary, expressions in English or in any other natural language are intrinsically vague such that the same word (e.g., "may") can have multiple meanings in different contexts; thus, distinguishing HI from PII can be challenging.

Still it is feasible to manually de-identify clinical reports if the amount of text is limited and if experienced de-identification professionals are available and well trained. In the era of Big Data, the first premise is rarely applicable. Manual de-identification can be very expensive for institutions and would become infeasible as the number of studies relying on clinical reports increases. Data providers may have little or no incentive to perform manual de-identification. Although they may charge only a nominal fee for such a service, by law, providers cannot seek any profit by providing the data. To provide manual de-identification services, institutes would have to undertake huge burdens of operational and financial overhead as well as risk negative publicity if the process failed.

Types of automatic text de-identification systems

Fortunately, there are automatic clinical text de-identification tools whose sensitivity and accuracy increase continuously. PARAT (54) and De-ID™ (55) are two such tools that are commercially available. There are also other freely available software applications such as MIST (56), de-id (57), and NLM-Scrubber (58-60) developed at MITRE, MIT and NIH, respectively. Both MIST and de-id have been academic exercises without a long-term plan for improvement and maintenance.

NLM-Scrubber was developed with a long-term goal of providing a non-commercial solution to biomedical scientists and institutions that do not have necessary resources to undertake the clinical text de-identification process. As a government research organization, NIH has no profit motive. The project aims to provide US taxpayers the best patient privacy protection and allows them to benefit from rapid scientific advances.

Current de-identification systems use various methods to recognize PII in clinical data, but neither a survey (61-64) of current de-identification systems nor a survey of computational techniques and mathematics of de-identification is within the scope of this review. However, clinical data scientists need to know what to expect from a de-identification system, how it can be used, and most critically, what types of input they need to provide the system to receive the desired output.

Annotated health information

Some clinical text de-identification systems such as MIST use supervised machine learning (ML) methods and require a set of training data where each PII element is annotated manually. Such ML systems either come with an annotation tool or are capable of using outputs of existing annotation tools (65). Annotation is a precursor of de-identification performed by human experts (66); thus, an ML system learns from the annotations of human experts, and attempts to recognize PII elements in the non-training data and to replicate the human performance during the de-identification process. Other de-identification systems such as NLM-Scrubber operate without training data.

All health institutes that de-identify health records need to employ human experts to annotate a small subset of randomly selected health records from the data of the larger cohort that needs to be de-identified. This small set of annotated HI would serve as the gold standard to evaluate the performance of the automatic de-identification system of the institute (67,68). Without such an evaluation and verification, a health institute would not know if the output of the de-identification system is truly de-identified.

An ML system requires annotated training data that is usually much larger than the annotated gold standard data. Unfortunately annotated gold standard data cannot be reused for training purposes since the two need to be mutually exclusive; otherwise, the evaluation results would be misleading. The size of the training data depends on the learning ability of the system and on the complexity of the data that needs to be de-identified; thus, an institute has to train the system in iterative steps with increasing size of training data, until the size increase does not significantly improve the system's performance. Due to the open-ended nature of the training data production and the large size of the prerequisite training data, the overhead of creating training datasets may be overwhelming for health institutes that lack the necessary human resources to carry out the task.

Modes of de-identification

From the clinical data scientist point of view, an automatic clinical text de-identification application is a black box; that is, the application takes some input and produces de-identified data-the underlying mechanism of de-identification does not matter much as long as the application produces the desired output. To produce optimal results, the data scientist needs to know the various operation modes available for the de-identification system in hand.

An earlier study (69) distinguished eight modes of de-identification, to which we add a ninth, pseudonymization

TABLE 2. Modes of de-identification

a) Repository-wide batch de-identification,
b) On-demand cohort-specific de-identification,
c) On-demand de-identification of query results,
d) De-identification with patient and provider identifiers,
e) Pseudonymization,
f) Scientist involved de-identification,
g) Patient involved de-identification,
h) Physician involved de-identification,
i) Online de-identification by honest brokers.

(Table 2). These modes define how the user can operate a given de-identification system if the system provides a particular functionality. The first three change the mode of operation in terms of de-identification time and input. The next two alter the input and output modes, respectively. The following three modes involve different stakeholders as active participants in the de-identification operation, and the last mode moves de-identification to the cloud. Most of these modes can be combined to maximize the protection of patient privacy and the integrity of the de-identified data.

Repository-wide batch de-identification is the default mode of operation adopted by most (if not all) existing systems. For an institute, it is tempting to de-identify its entire repository at once and make the de-identified data available to researchers when requested. In contrast, the next two modes de-identify data on demand. The repository-wide batch mode makes the data available at the time of request without additional operational overhead. However, the data might have been de-identified using an older technology with a lower quality of de-identification; the de-identified data may be incomplete and/or incorrect if the source data has been updated since the de-identification occurred; and it may not contain some of the required demographic information necessary for the study.

In *on-demand cohort-specific de-identification*, the data of the cohort that researchers defined is de-identified on demand. Since modern de-identification systems are very fast, the delay between the data collection and de-identification would be insignificant. *On-demand de-identification of query results* requires the integration of the de-identification system into the EHR system. Results of the query can be de-identified on the fly before being displayed to researchers.

By augmenting the input mode of de-identification with patient and provider identifiers, the accuracy of results can be improved significantly (58). In the pseudonymization mode, the de-identified data replaces PII elements (e.g., "Fred Jones") with pseudonyms (e.g., "John Doe") instead of with a label of the corresponding PII element [e.g., "(Personal Name)"], so if the system fails to de-identify some PII elements, the user might not be aware of the failure as remaining PII elements blend in among other pseudonyms (70).

In the *scientist involved de-identification* mode, scientists actively participate in the de-identification, producing better de-identification results. If the scientist's active participation is ensured, the sensitivity of the de-identification system for recognizing PII elements can be increased manually. As a side effect of the increased sensitivity, some HI could be misidentified as PII. By reviewing the first batch of de-identified results, the scientist can identify a set of misidentified terms, which can then be input to the system, so those terms can be preserved during the second de-identification cycle. De-identification using this mode results in better protection of patient privacy and a more complete set of de-identified data with higher scientific value and data integrity.

Patient involved de-identification is hypothetical since no existing system currently offers patients to annotate their own records for de-identification purposes. In very rare occasions, the context of the narrative might inadvertently reveal the identity of the patient; e.g., "injured during his US championship match today" (71). In such cases, manual patient annotations would help improve de-identification results. Furthermore, as de-identified clinical reports become widely available to researchers, it is likely that patients would demand to be informed of which portions of their records are made available to researchers.

Physicians are occasionally required to cite the patient's full name and medical record number to link the record to that specific patient but it is a generally unnecessary and unadvisable practice. It would be best if medical students are trained to write anonymous clinical reports without patient identifiers so that these reports can be used for scientific research purposes in the future. Using *physician involved de-identification* mode, the system warns physicians whenever they use patient identifiers. If such identifiers are necessary for clinical care purposes, they can be automatically labeled and those labels then verified by the physician.

As big health data becomes widely available to clinical scientists, it will likely be accumulated and accessed at large centers such as state cancer registries, state universities, and government research centers, which can allocate the expertise and necessary resources to handle big data and provide services to other institutes nationwide. The *online de-identification* mode would enable scientists of smaller institutes to access de-identified data of much larger cohorts. Centers holding big health data can act as honest brokers, de-identify the data, develop proper data use agreements, and monitor compliance of users.

DISCUSSIONS

Protecting patient privacy requires various technical tools. It involves regulations for sharing, de-identifying, securely storing, transmitting and handling PHI. It involves privacy laws and legal agreements. It requires establishing rules for monitoring privacy leaks, determining actions when they occur, and handling de-identified clinical narrative reports. De-identification is one such indispensable instrument in this set of privacy tools.

Protecting patient privacy requires collaboration among all stakeholders, which include patients, PHI holding institutions, users of HI, developers of automatic de-identification tools, and regulatory and law enforcement government agencies. Each group has a different set of roles and responsibilities. For example, institutions should be held responsible to select the right tools, monitor the adequacy of these tools over time, and ensure the quality and content of de-identified data before presenting it to the user. They also are required to use these tools properly by supplying all necessary input to the de-identification system and utilizing all available modes of de-identification to maximize privacy protection. Institutions are also responsible to establish proper data use agreements.

Institutions and users of HI are equally responsible for ensuring that the requested and granted data comprise only the HI that is necessary for the study. Both regulatory agencies and institutions should empower patients to actively protect their privacy by monitoring their EHRs and let them know what portions of their data have been shared, with whom, and to what end. Institutions should demand from their users to provide study terms of interest to input to the de-identification process, so that the scientific integrity of the data can be preserved while privacy protection can be achieved at the highest level of sensitivity for de-identifying PHI.

As outlined above, the demand from institutions holding PHI is significant. Smaller institutions can be overwhelmed by the operational and financial overhead. There is little or no incentive structure for these institutions to take this challenge eagerly and share the data for secondary scientific use, particularly with scientists outside of those institutions. The entire scientific community including journal editors and the public, with the help of regulatory and grant providing agencies, should build incentive structures to support these institutions and make their contributions to the advancement of science visible.

In conclusion, Big Data makes the problem of patient privacy protection bigger and more difficult to attain; however, recent advances in computational de-identification help remedy the problem and enable scientists to access big health data by minimizing the risk to patient privacy. We have made great strides in developing both regulatory and technical privacy tools for the era of big data; however, this is still a work in progress. We reviewed the progress of patient privacy protection with a focus on the U.S. As seen in References, regulations have been continuously updated with numerous amendments. We did not discuss the European efforts but the regulations there are more in flux. In 2016, the European Parliament enacted the General Data Protection Regulation (GDPR), which will take effect in 2018 (38). GDPR provides patient privacy protection using a language similar to the Privacy Rule.

Thanks to the digital communication revolution, the world gets smaller every day. As everyone deserves to equally benefit from scientific advances, it is inevitable that any legal differences among nations including U.S., Europe, Canada, and Australia will soon be ironed out so that we all can collaborate to find cures to today's incurable diseases and improve the quality of life around the world.

Acknowledgements

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. The author is the principal investigator in NLM-Scrubber project at NIH. He receives royalties from University of Pittsburgh for his contribution to a de-identification project; the resulting product was acquired by a third party, which today is known as De-ID Data Corp. NLM's Ethics Office reviewed and approved his appointment.

REFERENCES

1. Hippocrates. Jusjurandum (The Oath). In: Jones WHS, editor. Loeb Classical Library. Reprint: Hippocrates Collected Works I. Hippocrates ed. Cambridge, MA: Harvard University Press; 1868.
2. Higgins GL. The history of confidentiality in medicine. Canadian Family Physician. 1989;35:921-14. PubMed PMID: PMC2280818.
3. Parent WA. Recent work on the concept of privacy. American Philosophical Quarterly. 1983;20(4):341-55.
4. Heins M. "The right to be let alone": privacy and anonymity at the US Supreme Court. Revue française d'études américaines. 2010 (1):54-72.
5. Coke E. Semayne's case. In: Court of King's Bench, editor. 5 Co Rep 91a, 77 Eng Rep 1941604.
6. U.S. Constitution Amendment IV–search and seizure (1791).
7. Warren SD, Brandeis LD. The right to privacy. Harvard Law Review. 1890;4(5):193-220.
8. Coleman AH. The patient's right to privacy. J Natl Med Assoc. 1961 Mar;53(2):207. PubMed PMID: 20894011. Pubmed Central PMCID: 2641881. eng.
9. Al-Fedaghi SS. The "right to be let alone" and private information. In: Chen C-S, Filipe J, Seruca I, Cordeiro J, editors. Enterprise Information Systems VII. Dordrecht: Springer Netherlands; 2006. p. 157-66.
10. Yamamoto R. [Management system of personal data protection in the health care field]. Rinsho Byori. 2014 Nov;62(11):1129-34. PubMed PMID: 27509734. Epub 2014/11/01. jpn.
11. United Nations. Universal declaration of human rights. Paris 1948.

12. Thomson JJ. The right to privacy. Philosophy & Public Affairs. 1975;4(4):295-314.
13. Feinberg W. Recent developments in the law of privacy. Columbia Law Review. 1948;48(5):713-31.
14. Veal WR. Torts–right of privacy. Louisiana Law Review. 1949;9(4):17.
15. Thompson IE. The nature of confidentiality. Journal of Medical Ethics. 1979;5(2):57-64. PubMed PMID: PMC1154714.
16. Scanlon T. Thomson on privacy. Philosophy & Public Affairs. 1975;4(4):315-22.
17. Richards NM, Solove DJ. Privacy's Other Path: Recovering the law of confidentiality. Georgetown Law Journal. 2007;96:123-82.
18. Beardsley E. Privacy: Autonomy and selective disclosure. Nomos XIII: Privacy. 1971:56-70.
19. American Medical Association. Code of medical ethics. 2001.
20. Directorate-General for Research and Innovation. Ethics for researchers, facilitating research excellence in FP7. Luxembourg: European Commission; 2013.
21. Bernasek A. Should tax bills be public information? The New York Times. 2010 2/13/2010;Sect. Your Taxes.
22. 45 CFR §46.102. Definitions. Basic HHS policy for protection of human research subjects. Deparment of Health and Human Services; 2017.
23. Sebelius K. 45 CFR Parts 160 and 164. Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the health information technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; other modifications to the HIPAA rules; final rule In: Office of the Secretary of the Deparment of Health and Human Services, editor. Federal Register, Volume 78, No 172013. p. 5566-702.
24. Administrative simplifications: definitions, 42 U.S. Code §1320d, (1996).
25. 45 CFR §160.103 Definitions. [65 FR 82798, Dec. 28, 2000, as amended at 67 FR 38019, May 31, 2002; 67 FR 53266, Aug. 14, 2002; 68 FR 8374, Feb. 20, 2003; 71 FR 8424, Feb. 16, 2006; 76 FR 40495, July 8, 2011; 77 FR 1589, Jan. 10, 2012; 78 FR 5687, Jan. 25, 2013]: Department of Health and Human Services; 2013.
26. Family educational and privacy rights, 20 U.S. Code §1232g, (2010).
27. National Institutes of Health. Certificates of confidentiality (CoC) [7/12/2017]. Available from: https://humansubjects.nih.gov/coc/index, https://humansubjects.nih.gov/coc/faqs.
28. Research and investigations generally, 42 U.S. Code §241, (2016).
29. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191(1996).
30. Office of Civil Rights. The HIPAA privacy rule 2015. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.
31. 45 CFR §164.502 Uses and disclosures of protected health information: general rules. [65 FR 82802, Dec. 28, 2000, as amended at 67 FR 53267, Aug. 14, 2002; 78 FR 5696, Jan. 25, 2013]: Department of Health and Human Services; 2013.
32. 45 CFR §164.508 Uses and disclosures for which an authorization is required. [67 FR 53268, Aug. 14, 2002, as amended at 78 FR 5699, Jan. 25, 2013]: Department of Health and Human Services; 2013.
33. 45 CFR §164.510 Uses and disclosures requiring an opportunity for the individual to agree or to object. [65 FR 82802, Dec. 28, 2000, as amended at 67 FR 53270, Aug. 14, 2002; 78 FR 5699, Jan. 25, 2013]: Department of Health and Human Services; 2013.
34. 45 CFR §164.512 Uses and disclosures for which an authorization or opportunity to agree or object is not required. [65 FR 82802, Dec. 28, 2000, as amended at 67 FR 53270, Aug. 14, 2002; 78 FR 5699, Jan. 25, 2013; 78 FR 34266, June 7, 2013; 81 FR 395, Jan. 6, 2016]: Department of Health and Human Services; 2016.
35. 45 CFR §164.514 Other requirements relating to uses and disclosures of protected health information. [65 FR 82802, Dec. 28, 2000, as amended at 67 FR 53270, Aug. 14, 2002; 78 FR 5700, Jan. 25, 2013; 78 FR 34266, June 7, 2013]: Department of Health and Human Services; 2013.
36. 45 CFR §46.116. General requirements for informed consent. [56 FR 28012, 28022, June 18, 1991, as amended at 70 FR 36328, June 23, 2005]: Department of Health and Human Services; 2005.
37. 21 CFR §50.20 General requirements for informed consent. [46 FR 8951, Jan. 27, 1981, as amended at 64 FR 10942, Mar. 8, 1999]: Department of Health and Human Services; 1999.
38. The European Parliament and the Council of the European Union. General Data Protection Regulation. 2012/0011 (COD). Brussels Council of the European Union; 2016.
39. Office of Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with Health Insurance Portability and Accountability Act (HIPAA) privacy rule. In: U.S. Department of Health and Human Services, editor. 2012.
40. Privacy Analytics. When is it appropriate to use safe harbor? 2015 [7/9/2017]. Available from: https://privacy-analytics.com/de-id-university/blog/using-safe-harbor-de-identification/.
41. Shankland S. Google begins blurring faces in Street View. c|net. 5/13/2008.
42. Blake H. Google's EU warning over Street View privacy. The Telegraph. 2010 2/26/2010.
43. Miller CC, O'Brien KJ. Germany’s complicated relationship with Google Street View. The New York Times. 4/23/2013.
44. Johnston C. Google Street View's beefed-up privacy blurs cow's face. The Guardian. 9/18/2016.
45. Bischoff-Grethe A, Ozyurt IB, Busa E, Quinn BT, Fennema-Notestine C, Clark CP, et al. A technique for the deidentification of structural brain MR images. Hum Brain Mapp. 2007 Sep;28(9):892-903. PubMed PMID: 17295313. Pubmed Central PMCID: 2408762. Epub 2007/02/14. eng.
46. Gonzalez DR, Carpenter T, van Hemert JI, Wardlaw J. An open source toolkit for medical imaging de-identification. Eur Radiol. 2010 Aug;20(8):1896-904. PubMed PMID: ISI:000279656400012. English.
47. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008 Aug 29;4(8):e1000167. PubMed PMID: 18769715. Pubmed Central PMCID: 2516199. Epub 2008/09/05. eng.
48. Sweeney L, Abu A, Winn J. Identifying participants in the Personal Genome Project by name. Data Privacy Lab, IQSS, Harvard University. [White paper]. In press 2013.
49. van 't Hoff E. The data explosion along the care cycle. NVKVV 16de Colloquim ICT en gezondheidszorg; Affligem, Belgium: Dell Healthcare; 2012.
50. Wu W, Ding H. Big data solutions for healthcare. [Presentation]. In press 2013.
51. Datamark Inc. Unstructured data in electronic health record (EHR) systems: challenges and solutions. 2013.
52. Rhinehart C. The impact of cognitive computing on healthcare. IBM Watson Health; 2015.
53. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):54-64. PubMed PMID: PMC2274868.
54. Privacy Analytics. PARAT maintenance and support information [updated 6/25/20147/9/2017]. Available from: http://knowledgebase.privacy-analytics.com/index.php?/article/AA-00335/25/PARAT/General/PARAT-Maintenance-and-Support-Information.html.
55. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004 Feb;121(2):176-86. PubMed PMID: 14983930. Epub 2004/02/27. eng.
56. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assn. 2007 September 1, 2007;14(5):564–73.

57. Neamatullah I, Douglass M, Lehman L-w, Reisner A, Villarroel M, Long W, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008;8(1):32. PubMed PMID: doi:10.1186/1472-6947-8-32.
58. Kayaalp M, Browne AC, Callaghan FM, Dodd ZA, Divita G, Ozturk S, et al. The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. J Am Med Inform Assn. 2014;21(3):423-31.
59. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. Clinical text de-identification research. A report to the Board of Scientific Counselors. U.S. National Library of Medicine, National Institutes of Health, Communications LHNCfB; 2013 September 2013. Report No 2013-001.
60. Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of address, date, and alphanumeric identifiers in narrative clinical reports. Proc AMIA Annu Symp. 2014:767-76.
61. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assn. 2007 September 1, 2007;14(5):550-63.
62. Meystre S, Friedlin F, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology. 2010;10(1):70. PubMed PMID: doi:10.1186/1471-2288-10-70.
63. Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care. 2012 Jul;50 Suppl:S82-101. PubMed PMID: 22692265. Epub 2012/06/22. eng.
64. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;58, Supplement:S11-S9.
65. MITRE. Annotations [7/9/2017]. Available from: http://mist-deid.sourceforge.net/current_docs/html/annotation_intro.html.
66. Stubbs A, Uzuner O. De-identification of medical records through annotation. In: Ide N, Pustejovsky J, editors. Handbook of Linguistic Annotation. Dordrecht, The Netherlands: Springer; 2017. p. 1433-59.
67. Browne AC, Kayaalp M, Dodd ZA, Sagan P, McDonald CJ. The challenges of creating a gold standard for de-identification research. Proc AMIA Annu Symp. 2014:353-8.
68. Kayaalp M, Browne AC, Sagan P, McGee T, McDonald CJ. Challenges and insights in using HIPAA privacy rule for clinical text annotation. Proc AMIA Annu Symp. 2015:707-16.
69. Kayaalp M. Modes of de-identification. Proc AMIA Annu Symp; Forthcoming; an advance copy available at https://lhncbc.nlm.nih.gov/publication/pub95262017.
70. Carrell D, Malin B, Aberdeen J, Bayer S, Clark C, Wellner B, et al. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):342-8. PubMed PMID: 22771529. Pubmed Central PMCID: 3638183. Epub 2012/07/10. eng.
71. Kayaalp M, Sagan P, Jones JK, Browne AC, McDonald CJ. Guidelines for annotating personal identifiers in the clinical text repository of the National Institutes of Health. Available at https://scrubber.nlm.nih.gov/annotation/ 2016.