nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-020-0186-1

Secure, privacy-preserving and federated machine learning in medical imaging

Georgios A. Kaissis[1,2,3], Marcus R. Makowski[1], Daniel Rückert[2] and Rickmer F. Braren[1]✉

The broad application of artificial intelligence techniques in medicine is currently hindered by limited dataset availability for algorithm training and validation, due to the absence of standardized electronic medical records, and strict legal and ethical requirements to protect patient privacy. In medical imaging, harmonized data exchange formats such as Digital Imaging and Communication in Medicine and electronic data storage are the standard, partially addressing the first issue, but the requirements for privacy preservation are equally strict. To prevent patient privacy compromise while promoting scientific research on large datasets that aims to improve patient care, the implementation of technical solutions to simultaneously address the demands for data protection and utilization is mandatory. Here we present an overview of current and next-generation methods for federated, secure and privacy-preserving artificial intelligence with a focus on medical imaging applications, alongside potential attack vectors and future prospects in medical imaging and beyond.

Artificial intelligence (AI) methods have the potential to revolutionize the domain of medicine, as witnessed, for example, in medical imaging, where the application of computer vision techniques, traditional machine learning[1,2] and—more recently—deep neural networks have achieved remarkable successes. This progress can be ascribed to the release of large, curated corpora of images (ImageNet[3] perhaps being the best known), giving rise to performant pre-trained algorithms that facilitate transfer learning and led to increasing publications both in oncology—with applications in tumour detection[4,5], genomic characterization[6,7], tumour subtyping[8,9], grading prediction[10], outcome risk assessment[11] or risk of relapse quantification[12]—and non-oncologic applications, such as chest X-ray analysis[13] and retinal fundus imaging[14].

To allow medical imaging AI applications to offer clinical decision support suitable for precision medicine implementations, even larger amounts of imaging and clinical data will be required. Large cross-sectional population studies based solely on volunteer participation, such as the UK Biobank[15], cannot fill this gap. Even the largest current imaging studies in the field[4,5], demonstrating better-than-human performance in their respective tasks, include considerably less data than, for example, ImageNet[3], or the amount of data used to train algorithmic agents in the games of Go or StarCraft[16,17], or autonomous vehicles[18]. Furthermore, such datasets often stem from relatively few institutions, geographic regions or patient demographics, and might therefore contain unquantifiable bias due to their incompleteness with respect to co-variables such as comorbidities, ethnicity, gender and so on[19].

However, considering that the sum of the world's patient databases probably contains enough data to answer many significant questions, it becomes clear that the inability to access and leverage this data poses a significant barrier to AI applications in this field.

The lack of standardized, electronic patient records is one reason. Electronic patient data management is expensive[20], and hospitals in underprivileged regions might be unable to afford participation in studies requiring it, potentially perpetuating the aforementioned issues of bias and fairness. In the medical imaging field, electronic data management is the standard: Digital Imaging and Communications in Medicine (DICOM)[21] is the universally adopted imaging data format, and electronic file storage is the near-global standard of care. Even where non-digital formats are still in use, the archival nature of, for instance, film radiography allows post hoc digitization, seen, for example, in the CBIS-DDSM dataset[22], consisting of digitized film breast radiographs. Digital imaging data, easily shareable, permanently storable and remotely accessible in the cloud has driven the aforementioned successes of medical imaging AI.

The second issue representing a stark deterrent from multi-institutional/multi-national AI trials[23] is the rigorous regulation of patient data and the requirements for its protection. Both the United States Health Insurance Portability and Accountability Act (HIPAA)[24] and the European General Data Protection Regulation (GDPR)[25] mandate strict rules regarding the storage and exchange of personally identifiable data and data concerning health, requiring authentication, authorization, accountability and—with GDPR—AI interpretability, sparking considerations on data handling, ownership and AI governance[26,27]. Ethical, moral and scientific guidelines (soft law[28]) also prescribe respect towards privacy—that is, the ability to retain full control and secrecy about one's personal information. The term privacy is used in this article to encapsulate both the intention to keep data protected from unintended leakage and from deliberate disclosure attempts (that is, synonymous with 'confidentiality').

AI in medical imaging is a multifaceted field of patients, hospitals, research institutions, algorithm developers, diagnostic equipment vendors, industry and lawmakers. Its high complexity and resulting lack of transparency with respect to stakeholder motives and data usage patterns, alongside the facilitated data sharing enabled by electronic imaging data storage, threaten to diminish the importance of individual privacy and relax the grip on personal data in the name of, at best, scientific development and, at worst, financial interests. The field of secure and privacy-preserving AI offers techniques to help bridge the gap between personal data protection and data utilization for research and clinical routine. Here, we present an overview of current and emerging techniques for privacy

[1]Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Technical University of Munich, Munich, Germany. [2]Department of Computing, Imperial College London, London, UK. [3]OpenMined. ✉e-mail: rbraren@tum.de

Table 1 | Glossary of terms encountered in the article alongside conceptual examples

Method	Description	Example
Attack vectors		
Attacks against the dataset		
Re-identification attack	Determining an individual's identity despite anonymization based on other information present in the dataset.	Exploiting similarities to other datasets in which the same individual is contained (linkage).
Dataset reconstruction attack	Deriving an individual's characteristics from the results of computations performed on a dataset without having access to the dataset itself (synonyms: feature re-derivation, attribute inference).	Using multiple aggregate statistics to derive data points corresponding to a single individual.
Tracing attack	Determining whether an individual is present in the dataset or not without necessarily determining their exact identity (synonym: membership inference).	Exploiting repeated, slightly varying dataset queries to 'distil' individual information (set differencing).
Attacks against the algorithm		
Adversarial attack	Manipulation of the input to an algorithm with the goal of altering it, most often in a way that makes the manipulation of the input data impossible to detect by humans.	Compromising the computation result by introducing malicious training examples (model poisoning).
Model-inversion/reconstruction attack	Derivation of information about the dataset stored within the algorithm's weights by observing the algorithm's behaviour.	Using generative algorithms to recreate parts of the training data based on algorithm parameters.
Secure and private AI terminology		
Secure by default implementation (synonym private by design)	Systems that have been designed from the ground up with privacy in mind and at best require no specialized data handling.	—
Anonymization	Removal of personally identifiable information from a dataset.	Removing information related to age, gender and so on.
Pseudonymization	Replacement of personally identifiable information in a dataset with a dummy/synthetic entry with separate storage of the linkage record (look-up table).	Replacing names with randomly generated text.
Secure AI	Techniques concerned with protecting the AI algorithms.	Algorithm encryption.
Privacy-preserving AI	Techniques for protecting the input and output data.	Data encryption, decentralized storage.
Federated machine learning	Machine learning system relying on distributing the algorithm to where the data is instead of gathering the data where the algorithm is (decentralized/distributed computation).	Training of algorithms on hospital computer systems instead of on cloud servers.
Differential privacy	Modification or perturbation of a dataset to obfuscate individual data points while retaining the ability of interaction with a data within a certain scope (privacy budget) and of statistical analysis. Can also be applied to algorithms.	Random shuffling of data to remove the association between individuals and their data entries.
Homomorphic encryption	Cryptographic technique that preserves the ability to perform mathematical operations on data as if it was unencrypted (plain text).	Performing neural network computations on encrypted data without first decrypting it.
Secure (multi-party) computation	Collection of techniques and protocols enabling two or more parties to split up data among them to perform joint computations in a way that prevents any single party from gaining knowledge of the data but preserving the computational result.	Determining which patients two hospitals have in common without revealing their respective patient list (private set intersection).
Hardware security implementation	Collection of techniques whereby specialized computer hardware provides guarantees of privacy or security.	Secure storage or processing enclaves in mobile phones or computers.

preservation with a focus on their applications in medical imaging, discuss their benefits, drawbacks and technical implementations, as well as potential weaknesses and points of attack aimed at compromising privacy. We conclude with an outlook on the current and future developments in the field of medical imaging and beyond, alongside their potential implications.

Definitions and attack vectors

A glossary of the terms presented throughout the article can be found in Table 1, and a visual overview of the field can be found in Fig. 1.

Optimal privacy preservation requires implementations that are secure by default (synonymously privacy by design[29]). Such systems

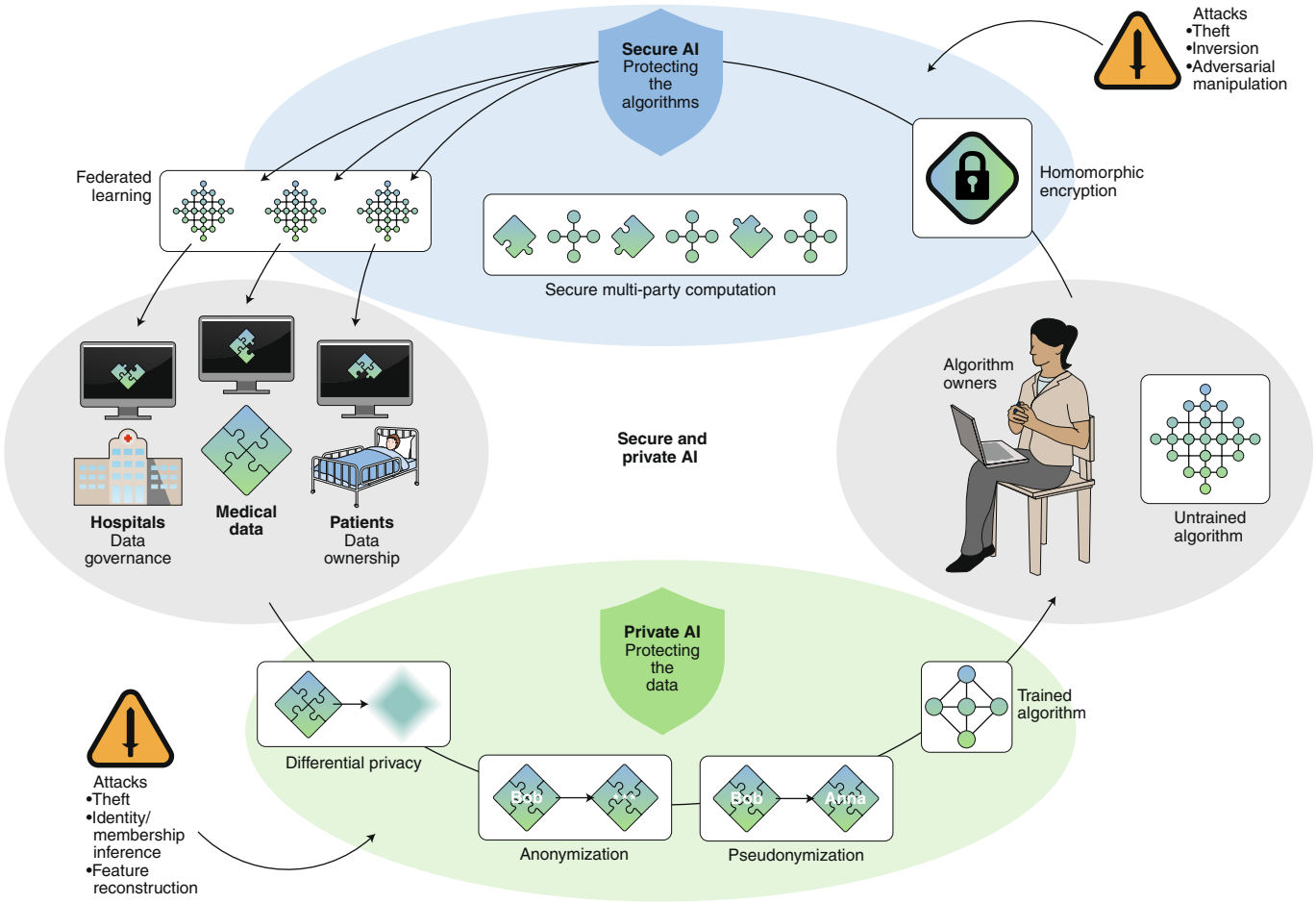

Fig. 1 | Secure and private AI. Schematic overview of the relationships and interactions between data, algorithms, actors and techniques in the field of secure and private AI.

should require minimal or no data transfer and provide theoretical and/or technical guarantees of privacy.

The term secure AI is used for methods concerned with safeguarding algorithms, and the term privacy-preserving AI for systems allowing data processing without revealing the data itself. Their combination aims to guarantee sovereignty over the input data and the algorithms, integrity of the computational process and its results, and to offer trustworthy and transparently auditable technical implementations (structured transparency). Such systems must resist attacks against the dataset[30], for example identity or membership inference/tracing[31] (determining whether an individual is present in a given dataset) and feature/attribute re-derivation/re-identification[30] (extraction of characteristics of an individual from within the dataset, for example by linkage attacks[32]). They must also withstand attacks on the algorithm or the computational process—for instance, modification of algorithm parameters (for example, by poisoning[33])—or derivation of information about the dataset from them (model-inversion/reconstruction[34]). Finally, they must protect the data and the algorithms from theft both in storage and when transmitted over networks (asset/integrity protection).

Anonymization, pseudonymization and the risks of re-identification

Anonymization (the removal of private data from a record) and pseudonymization (replacement of sensitive entries with artificially generated ones while still allowing re-attribution using a look-up table)—collectively de-identification—are currently the most widely used privacy preservation techniques for medical datasets. In medical imaging, anonymization requires removing all pertinent DICOM metadata entries (for example, patient name, gender and so on). For pseudonymization, the true entries are replaced by synthetic data (see overview of techniques in ref. [35]), and the look-up table safe-kept separately. The main benefit of both approaches is simplicity. Anonymization software is built into most clinical data archiving systems, rendering it the easiest method in practice. Pseudonymization poses additional difficulties since it requires data manipulation, not just data deletion, and safekeeping of the look-up tables for reversing the process. The latter can be problematic in the setting of insecure storage, risking data theft[36]. Furthermore, technical errors can render the protection ineffective and potentially (for example, in case of retaining institution names), an entire dataset identifiable. Moreover, there is substantial discourse regarding the definition of 'sufficient/reasonable' de-identification[37] related to the objective/technical difficulty of reversing the process. Different points of view exist in different jurisdictions[38], complicating the establishment of international standards. Also, de-identification techniques are usually employed as a preparation to data transfer or sharing. This presents issues in case the patient withdraws their consent, since it uncouples data governance from data ownership (impeding the right to be forgotten, GDPR article 17), or if the legislation changes. Lastly, requirements towards the de-identification process vary according to the type of imaging dataset: a radiograph of a leg is harder to link back to an individual than a computed tomography scan of their head, where the contours of the face can be reconstructed directly from the image. Such re-identification

attacks[39] have been shown to yield high success rates both with tabular data[40,41] (such as patient records) and medical imaging data[42]. As a consequence, datasets more prone to identification must be processed more rigorously, for instance by removal of the face or skull region from the images (defacing/skull stripping). This complicates data handling, increasing the probability of errors and constitutes a manipulation of the imaging data, which, at worst, represents an adversarial update to the algorithm[43], reducing its performance and robustness. Ultimately, even such processing might not be sufficient for the full de-identification of datasets[44]. Re-identified patient records are a lucrative target for health insurance companies wishing to reduce their financial risk by discriminating against individuals with certain illnesses. It has been reported that large-scale re-identification attacks and the sale of re-identified medical records have become a business model for data-mining companies[45]. De-identification by naive anonymization or pseudonymization alone must therefore be viewed as a technically insufficient measure against identity inference.

Decentralized data and federated machine learning

The concept of federated machine learning began gathering significant attention around the year 2015[46]. It belongs to a class of decentralized/distributed systems that rely on the principle of remote execution—that is, distributing copies of a machine learning algorithm to the sites or devices where the data is kept (nodes), performing training iterations locally, and returning the results of the computation (for example, updated neural network weights) to a central repository to update the main algorithm. Its main benefit is the ability of the data to remain with its owner (retention of sovereignty), while still enabling the training of algorithms on the data. The federation topology is flexible (model sharing among the nodes and aggregation at a later time (peer to peer/gossip strategy[47]) or full decentralization, combined, for example, with contribution tracking/audit trails using blockchains[48]). Continuous online availability is not required since training can be performed offline and results returned later. Thus, federated learning approaches have arguably become the most widely used next-generation privacy preservation technique, both in industry[49] and medical AI applications[50].

While federated learning is flexible and resolves data governance and ownership issues, it does not itself guarantee security and privacy unless combined with other methods described below. A lack of encryption can allow attackers to steal personally identifiable data directly from the nodes or interfere with the communication process. This communication requirement can be burdensome for large machine learning models or data volumes. The decentralized nature of the data complicates data curation to ascertain the integrity and quality of the results. Technical research must be performed to determine the optimal method for updating the central model state (distributed optimization, federated averaging). In case the local algorithms are not encrypted, or the updates aren't securely aggregated, data can leak or algorithms can be tampered with[51], reconstructed or stolen (parameter inference), which is unacceptable from the viewpoint of intellectual property, patent restrictions or asset protection. Moreover, neural networks represent a form of memory mechanism, with compressed representations of the training data stored within their weights (unintended memorization). It is therefore possible to reconstruct parts of the training data from the algorithm weights themselves on a decentralized node[52–54]. Such model inversion or reconstruction attacks can cause catastrophic data leakage: it has been shown that images can be reconstructed with impressive accuracy and detail[55], allowing visualization of the original training data. Federated learning thus offers an infrastructural approach to privacy and security, but further measures, highlighted below, are required to expand its privacy-preserving scope.

Differential privacy

Data-perturbation-based privacy approaches operate on the premise that the systematic randomized modification of a dataset or algorithm can reduce information about the single individual while retaining the capability of statistical reasoning about the dataset. The approach of retaining the global statistical distribution of a dataset while reducing individually recognizable information is termed differential privacy[56] (DP). Intuitively, a dataset is differentially private if an outside observer is unable to infer whether a specific individual was used for obtaining a result from the dataset. For example, a causal relationship between obesity and cardiac disease can be inferred without knowing the body mass index of the individual patients. DP thus offers resistance to re-identification attacks such as linkage or set differencing within a certain scope of interaction with the dataset (privacy budget[56]). DP can be applied to the input data (local DP), the computation results (global DP) or the algorithm. Implementations range from simple random shuffling of the input data[57] to the introduction of noise to the dataset (Gaussian DP[58] with the benefit of better interpretability). DP can also be applied to algorithm updates during training, for instance in neural networks via differentially private stochastic gradient descent[59] or private aggregation of teacher ensembles[60], or during inference time. Local DP ensures privacy at the source of the data, putting the data owner in control and is thus well suited to healthcare applications[61], for instance for federated learning applications in which health data are being collected by smartphones or wearable devices. DP applications to imaging are being actively explored[62].

Among the challenges associated with DP, the main is the perturbation of the dataset itself. Data manipulation can degrade the data, which in an area with access to relatively little data, such as medical imaging research, may prove deleterious to algorithm performance. The technique also poses challenges with respect to plausibility testing, explaining the process to patients—that is, data legibility (human–data interaction[63])—regarding algorithm development and implementation, and escalates the requirement for statistical expertise to ascertain data representativeness[64]. Most importantly, the specifics of implementing DP in imaging data are unclear. Tabular data can be easily shuffled, but the perturbation of images can have unpredictable effects, with research demonstrating this type of manipulation (for example, adversarial noise) both as an attack against algorithms[65] and a regularization mechanism leading to increased robustness[66] and resilience against inversion attacks. Thus, further research is required before the widespread application of DP in medical imaging.

Homomorphic encryption

A conceptually simple, albeit technically challenging approach to data or algorithm fortification is cryptography, widely recognized as a gold standard for information security. Current cryptographic algorithms cannot be cracked by brute force[67]. Encryption is easily explained to and trusted by patients and practitioners. It can be applied both to the algorithm and to the data allowing secure, joint computation.

Homomorphic encryption (HE) is an encryption scheme that allows computation on encrypted data as if it was unencrypted (plain text). Homomorphism is a mathematical concept whereby structure is preserved throughout a computation. Since only certain mathematical operations, such as addition and multiplication, are homomorphic, the application of HE to neural networks requires the operations defined within the algorithm to conform to these limitations and thus standard encryption algorithms like the advanced encryption standard (AES)[68] cannot be used. Several implementations of HE algorithms[69] with varying levels of efficiency exist, and the application of HE represents an efficiency–security trade-off, with computational performance currently the most notable issue. Nevertheless, HE has successfully been applied

to convolutional neural networks[70], and its benefits demonstrated in a 'machine learning as a service' scenario[71], whereby data is sent over the network to be processed on an off-site server (cloud computing). It can also be used in federated learning scenarios (with or without additional DP[61]) to securely aggregate encrypted algorithm updates[72].

Secure multi-party computation

Secure computation can be extended to multiple parties—secure multi-party computation (SMPC)[73]—whereby processing is performed on encrypted data shares, split among them in a way that no single party can retrieve the entire data on their own. The computation result can be announced without any party ever having seen the data itself, which can be recovered only by consensus. A conceptual example for SMPC is a ballot, where the result needs to be known, but the individual voter's preference does not. For a technical description of SMPC, we refer to ref. [74]. The research interest in SMPC has recently risen, since it allows for 'secret sharing' in semi-trusted and low-trust environments. Notably, SMPC has been used in the setting of genetic sequencing and diagnostics without revealing the patient's genome[75]. In the domain of medical imaging, SMPC can be employed to perform analyses on datasets completely in the encrypted domain and without otherwise perturbing the data. It can thus help to increase the effective amount of available data without revealing individual identities or risking information leakage. It can also enable the ethically responsible provision of machine learning services while rendering the commercial use of the data itself impossible, or at least under the control of the individual, and subject to legal regulation after appropriate ethical debate, similar to the debate about organ donation (single-use accountability). For example, machine-learning-assisted medical image analysis services can be provided under the guarantee of data protection from malicious use in case of theft or from unwarranted financial exploitation[76]. As long as the data and the algorithms are encrypted, they remain unusable unless permission is granted by both parties, yielding a shared governance model. The notable limitations of SMPC are the requirements for continuous data transfer between parties (communication overhead) and for their continuous online availability. The reliability/redundancy and scalability to more than a small number of parties is a concern for SMPC applications[77], and computational considerations are a concern beyond small algorithm sizes, with efficient SMPC implementations of state-of-the-art neural network algorithms currently under active development[78].

Secure hardware implementations

Encryption provides a theoretical/mathematical privacy guarantee. However, privacy guarantees on the hardware level also exist, for example, in the form of secure processors or enclaves implemented in mobile devices[79]. They can assure data and algorithm privacy, for example, in federated learning workflows, even in the case of operating system kernel breaches. Due to the rising significance of hardware-level deep learning implementations (for example, tensor processing units[80] or machine-learning-specific instruction sets[81]), it is likely that such system-based privacy guarantees (trusted execution environments) built into edge hardware such as mobile phones will become more prevalent.

Outlook

Medical imaging has arguably witnessed among the largest advances in AI applications due to the concurrent developments in computer vision. However, the issues of security and privacy are not limited to medical imaging[82], as seen for example in the 2019/2020 SARS-CoV2-pandemic, which sparked worldwide concern about the implications of setting political, ethical and legal precedents by large-scale automatic contact tracing and movement tracking, creating a demand for their safe and privacy-protecting technical implementation[83]. All AI applications including sensitive data unfold in a complex, multi-stakeholder tension field of conflicting interests. The unregulated use of private data is likely to be more widespread than assumed, and cases of misuse—especially out of financial interest—will probably increase further. Yet the techniques presented here offer an opportunity to prevent stakeholder interactions from becoming a zero-sum game.

We believe that the widespread adoption of secure and private AI will require targeted multi-disciplinary research and investment in the following areas. (1) Decentralized data storage and federated learning systems, replacing the current paradigm of data sharing and centralized storage, have the greatest potential to enable privacy-preserving cross-institutional research in a breadth of biomedical disciplines in the near future[84,85], with results in medical imaging[50,86] and genomics[87] recently demonstrated. (2) To counteract the drawbacks of the individual techniques already presented, efficient cryptographic and privacy primitives, neural network operations[88] based, for example, on functional encryption[89], quantization[90] and optimization strategies[91], and encrypted transfer learning approaches[92] must be further developed. (3) The trade-offs between accuracy, interpretability, fairness, bias and privacy (privacy-utility trade-offs), need to be researched. In the field of radiology, for instance, interpretability in the encrypted setting is limited to the evaluation of trained algorithms on new images or inspection of the plain-text input data; however, intermediate outputs might be obfuscated and hard to interpret. Current research about interpretable private algorithms[93] can alleviate this issue. (4) Cryptographic expertise is required for the design and implementation of secure and efficient systems that not only resist (or at least reveal) errors due to technical implementation, but are also robust against semi-honest or dishonest participants/adversaries attempting to undermine the system[94]. (5) Deployed models must be monitored and potentially corrected for temporal instability (that is, statistical drift[95]), which can be difficult with encrypted data or algorithms. (6) Until fully secure and private solutions are the standard, research has to address the question of how the right to be forgotten (for example, GDPR) can be realized—for example, via machine unlearning[96] ('un-training' an algorithm when an individual withdraws consent). (7) The widespread implementation of secure and private AI will hinge on lowering the barrier to entry for researchers and developers by provision of accessible, open-source tools such as open-source extensions to deep learning frameworks, implementations of state-of-the-art algorithms and federated learning solutions, many of which have recently become available[97,98]. (8) The development of auditable and objectively trustworthy systems[99] (that is, not relying on subjective assertions—for example, by governments) will promote the universal acceptance of secure and private AI solutions by individuals and policymakers. (9) The technical ability offered by secure and private AI solutions to retain sovereignty over one's identity[100] and new techniques to quantify and track the added value of individual datasets with respect to algorithm performance will strengthen the notion of private data as a scarce and valuable resource within an evolving data economy[101] currently experiencing oversupply[102]. (10) Lastly, we view both the education of patients, physicians, researchers and policymakers, and the open scientific, public and political discourse about privacy, current risks and technical possibilities as paramount for reinforcing the cultural value of privacy and cultivating a sustainable attitude of trust and value-aligned cooperation both in science and society.

Received: 26 February 2020; Accepted: 7 May 2020;
Published online: 8 June 2020

References

1. Aerts, H. J. W. L. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
2. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
3. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comp. Vision* **115**, 211–252 (2015).
4. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
5. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
6. Pinker, K., Chin, J., Melsaether, A. N., Morris, E. A. & Moy, L. Precision medicine and radiogenomics in breast cancer: new approaches toward diagnosis and treatment. *Radiology* **287**, 732–747 (2018).
7. Lu, H. et al. A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. *Nat. Commun.* **10**, 764 (2019).
8. Kaissis, G. et al. A machine learning model for the prediction of survival and tumor subtype in pancreatic ductal adenocarcinoma from preoperative diffusion-weighted imaging. *Eur. Radiol. Exp.* **3**, 41–41 (2019).
9. Kaissis, G. et al. A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy. *PLoS ONE* **14**, e0218642 (2019).
10. Cui, E. et al. Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. *Eur. Radiol.* **30**, 2912–2921 (2020).
11. Varghese, B. et al. Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. *Sci. Rep.* **9**, 1570 (2019).
12. Elshafeey, N. et al. Multicenter study demonstrates radiomic features derived from magnetic resonance perfusion images identify pseudoprogression in glioblastoma. *Nat. Commun.* **10**, 3170 (2019).
13. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
14. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
15. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
16. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
17. Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
18. Fridman, L. et al. MIT advanced vehicle technology study: large-scale naturalistic driving study of driver behavior and interaction with automation. *IEEE Access* **7**, 102021–102038 (2019).
19. Obermeyer, Z. & Mullainathan, S. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proc. Conf. Fairness, Accountability, and Transparency* 89 (ACM, 2019).
20. Wang, S. J. et al. A cost-benefit analysis of electronic medical records in primary care. *Am. J. Med.* **114**, 397–403 (2003).
21. DICOM reference guide. *Health Dev.* **30**, 5–30 (2001).
22. Lee, R. S. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**, 170177 (2017).
23. Price, W. N. & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
24. HIPAA. *US Department of Health and Human Services* https://www.hhs.gov/hipaa/index.html (2020).
25. GDPR. *Intersoft Consulting* https://gdpr-info.eu (2016).
26. Cath, C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Trans. R. Soc. A* **376**, 20180080 (2018).
27. Theodorou, A. & Dignum, V. Towards ethical and socio-legal governance in AI. *Nat. Mach. Intell.* **2**, 10–12 (2020).
28. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).
29. Cavoukian, A. *Privacy by Design* (Information and Privacy Commissioner of Ontario, 2011).
30. Dwork, C., Smith, A., Steinke, T. & Ullman, J. Exposed! A survey of attacks on private data. *Annu. Rev. Stat. Appl.* **4**, 61–84 (2017).
31. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership inference attacks against machine learning models. In *Proc. 38th IEEE Symp. Security and Privacy* https://doi.org/10.1109/SP.2017.41 (IEEE, 2017).
32. Bindschaedler, V., Grubbs, P., Cash, D., Ristenpart, T. & Shmatikov, V. The tao of inference in privacy-protected databases. In *Proc. VLDB Endowment* **11**, 1715–1728 (ACM, 2018).
33. Kurita, K., Michel, P. & Neubig, G. Weight poisoning attacks on pre-trained models. Preprint at https://arxiv.org/abs/2004.06660 (2020).
34. Al-Rubaie, M. & Chang, J. M. Privacy preserving machine learning: threats and solutions. *IEEE Secur. Priv.* **17**, 49–58 (2019).
35. Surendra, H. & Mohan, H. S. A review of synthetic data generation methods for privacy preserving data publishing. *Int. J. Sci. Technol. Res.* **6**, 95–101 (2017).
36. Jiang, J. X. & Bai, G. Types of information compromised in breaches of protected health information. *Ann. Intern. Med.* **172**, 159–160 (2019).
37. Taylor, M. J. & Wilson, J. Reasonable expectations of privacy and disclosure of health data. *Med. Law Rev.* **27**, 432–460 (2019).
38. *General Data Protection Regulation: NHS European Office Position Paper* (NHS Confederation, 2012).
39. El Emam, K., Jonker, E., Arbuckle, L. & Malin, B. A systematic review of re-identification attacks on health data. *PLoS ONE* **6**, e28071 (2011).
40. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symp. Security and Privacy* 111–125 (IEEE, 2008).
41. de Montjoye, Y. A., Radaelli, L., Singh, V. K. & Pentland, A. S. Identity and privacy. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* **347**, 536–539 (2015).
42. Schwarz, C. G. et al. Identification of anonymous MRI research participants with face-recognition software. *New Engl. J. Med.* **381**, 1684–1686 (2019).
43. Ma, X. et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* https://doi.org/10.1016/j.patcog.2020.107332 (2020).
44. Abramian, D. & Eklund, A. Refacing: reconstructing anonymized facial features using GANs. In *2019 IEEE 16th International Symp. Biomedical Imaging* https://doi.org/10.1109/ISBI.2019.8759515 (IEEE, 2019).
45. Tanner, A. *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records* (Beacon, 2017).
46. Konečný, J., McMahan, B. & Ramage, D. Federated optimization: distributed optimization beyond the datacenter. Preprint at https://arxiv.org/abs/1511.03575 (2015).
47. Hu, C., Jiang, J. & Wang, Z. Decentralized federated learning: a segmented gossip approach. Preprint at https://arxiv.org/abs/1908.07782 (2019).
48. Passerat-Palmbach, J. et al. A blockchain-orchestrated federated learning architecture for healthcare consortia. Preprint at https://arxiv.org/abs/1910.12603 (2019).
49. Konečný, J. et al. Federated learning: strategies for improving communication efficiency. Preprint https://arxiv.org/abs/1610.05492 (2016).
50. Rieke, N. et al. The future of digital health with federated learning. Preprint at https://arxiv.org/abs/2003.08119 (2020).
51. Tomsett, R., Chan, K. & Chakraborty, S. Model poisoning attacks against distributed machine learning systems. *Proc. SPIE* **11006**, 110061D (2019).
52. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. & Song, D. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proc. 28th USENIX Security Symp.* 267–284 (USENIX Association, 2019).
53. Zhang, Y. et al. The secret revealer: generative model-inversion attacks against deep neural networks. Preprint at https://arxiv.org/abs/1911.07135 (2019).
54. Hitaj, B., Ateniese, G. & Perez-Cruz, F. Deep models under the GAN: information leakage from collaborative deep learning. In *Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security* 603–618 (ACM, 2017).
55. Fredrikson, M., Jha, S. & Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. 22nd ACM SIGSAC Conf. Computer and Communications Security* 1322–1333 (ACM, 2015).
56. Roth, A. & Dwork, C. The algorithmic foundations of differential privacy. *Found. Trends Theoretical Comp. Sci.* **9**, 211–407 (2013).
57. Cheu, A., Smith, A., Ullman, J., Zeber, D. & Zhilyaev, M. Distributed differential privacy via shuffling. In *Annual Int. Conf. Theory and Applications of Cryptographic Techniques* 375–403 (Springer, 2018).
58. Dong, J., Roth, A. & Su, W. J. Gaussian differential privacy. Preprint at https://arxiv.org/abs/1905.02383 (2019).
59. Rajkumar, A. & Agarwal, S. A differentially private stochastic gradient descent algorithm for multiparty classification. In *Proc. Fifteenth Int. Conf. Artificial Intelligence and Statistics* **22**, 933–941 (PMLR, 2012).
60. Papernot, N. et al. Scalable private learning with PATE. In *Proc. 6th Int. Conf. Learning Representations* (ICLR, 2018).
61. Kim, J. W., Jang, B. & Yoo, H. Privacy-preserving aggregation of personal health data streams. *PLoS ONE* **13**, e0207639 (2018).
62. Mireshghallah, F. et al. A principled approach to learning stochastic representations for privacy in deep neural inference. Preprint at https://arxiv.org/abs/2003.12154 (2020).
63. Mortier, R., Haddadi, H., Henderson, T., McAuley, D. & Crowcroft, J. Human-data interaction: the human face of the data-driven society. Preprint at https://arxiv.org/abs/1412.6159 (2014).

64. Garfinkel, S. L., Abowd, J. M. & Powazek, S. Issues encountered deploying differential privacy. In *Proc. 2018 Workshop on Privacy in the Electronic Society* 133–137 (ACM, 2018).
65. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint at https://arxiv.org/abs/1412.6572 (2014).
66. You, Z., Ye, J., Li, K., Xu, Z. & Wang, P. Adversarial noise layer: regularize neural network by adding noise. In *2019 IEEE Int. Conf. Image Processing* https://doi.org/10.1109/ICIP.2019.8803055 (IEEE, 2019).
67. Schneier, B. & Sutherland, P. *Applied Cryptography: Protocols, Algorithms, and Source Code in C* 157–158 (Wiley, 1995).
68. Daemen, J. & Rijmen, V. *The Design of Rijndael: AES - The Advanced Encryption Standard* (Springer, 2013).
69. Acar, A., Aksu, H., Selcuk Uluagac, A. & Conti, M. A survey on homomorphic encryption schemes: theory and implementation. *ACM Comput. Surv.* **51**, 79 (2018).
70. Hesamifard, E., Takabi, H. & Ghasemi, M. CryptoDL: deep neural networks over encrypted data. Preprint at https://arxiv.org/abs/1711.05189 (2017).
71. Dowlin, N. et al. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy. In *Proc. 33rd Int. Conf. Machine Learning* Vol. 48 201–210 (PMLR, 2016).
72. Li, X., Chen, D., Li, C. & Wang, L. Secure data aggregation with fully homomorphic encryption in large-scale wireless sensor networks. *Sensors* **15**, 15952–15973 (2015).
73. Zhao, C. et al. Secure multi-party computation: theory, practice and applications. *Inform. Sci.* **476**, 357–372 (2019).
74. Evans, D., Kolesnikov, V. & Rosulek, M. *A Pragmatic Introduction to Secure Multi-Party Computation* (NOW, 2018).
75. Jagadeesh, K. A., Wu, D. J., Birgmeier, J. A., Boneh, D. & Bejerano, G. Deriving genomic diagnoses without revealing patient genomes. *Science* **357**, 692–695 (2017).
76. Helm, T. Patient data from GP surgeries sold to US companies. *The Guardian* https://www.theguardian.com/politics/2019/dec/07/nhs-medical-data-sales-american-pharma-lack-transparency (2019).
77. Tkachenko, O., Weinert, C., Schneider, T. & Hamacher, K. Large-scale privacy-preserving statistical computations for distributed genome-wide association studies. In *Proc. 2018 on Asia Conf. Computer and Communications Security* 221–235 (2018).
78. Kumar, N. et al. CrypTFlow: secure tensorflow inference. In *Proc. 41st IEEE Symp. Security and Privacy* (IEEE, 2020).
79. Secure enclave overview. *Apple Platform Security* https://support.apple.com/guide/security/secure-enclave-overview-sec59b0b31ff/web (2020).
80. Cloud TPU. *Google* https://cloud.google.com/tpu/ (2020).
81. Chen, A. Y. et al. An instruction set architecture for machine learning. *ACM Trans. Comput. Syst.* **36**, 9 (2019).
82. Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. Secure and robust machine learning for healthcare: a survey. Preprint at https://arxiv.org/abs/2001.08103 (2020).
83. Pandemic data challenges. *Nat. Mach. Intell.* **2**, 193 (2020).
84. Son, J. et al. Privacy-preserving electrocardiogram monitoring for intelligent arrhythmia detection. *Sensors* **17**, 1360 (2017).
85. Mudgal, K. S. & Das, N. The ethical adoption of artificial intelligence in radiology. *BJR Open* **2**, 20190020 (2020).
86. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In *Proc. 10th Int. Workshop on Machine Learning in Medical Imaging* 133–141 (Springer, 2019).
87. Grishin, D., Obbad, K. & Church, G. M. Data privacy in the age of personal genomics. *Nat. Biotechnol.* **37**, 1115–1117 (2019).
88. Takabi, D., Podschwadt, R., Druce, J., Wu, C. & Procopio, K. Privacy preserving neural network inference on encrypted data with GPUs. Preprint at https://arxiv.org/abs/1911.11377 (2019).
89. Ryffel, T., Dufour-Sans, E., Gay, R., Bach, F. & Pointcheval, D. Partially encrypted machine learning using functional encryption. In *Proc. 33rd Conf. Neural Information Processing Systems* (NeurIPS, 2019).
90. Chou, E. et al. Faster CryptoNets: leveraging sparsity for real-world encrypted inference Preprint at https://arxiv.org/abs/1811.09953 (2018).
91. Dathathri, R. et al. CHET: an optimizing compiler for fully-homomorphic neural-network inferencing. In *Proc. 40th ACM SIGPLAN Conf. Programming Language Design and Implementation* 142–156 (ACM, 2019).
92. Salem, M., Taheri, S. & Yuan, J.-S. Utilizing transfer learning and homomorphic encryption in a privacy preserving and secure biometric recognition system. *Computers* **8**, 3 (2019).
93. Harder, F., Bauer, M. & Park, M. Interpretable and differentially private predictions. In *Proc. Thirty-Fourth AAAI Conf. Artificial Intelligence* (AAAI, 2020).
94. Xu, Z., Li, C. & Jegelka, S. Robust GANs against dishonest adversaries. Preprint at https://arxiv.org/abs/1802.09700 (2018).
95. Nelson, K. et al. Evaluating model drift in machine learning algorithms. In *2015 IEEE Symp. Computational Intelligence for Security and Defense Applications* https://doi.org/10.1109/CISDA.2015.7208643 (IEEE, 2015).
96. Bourtoule, L. et al. Machine unlearning. Preprint at https://arxiv.org/abs/1912.03817 (2019).
97. Ryffel, T. et al. A generic framework for privacy preserving deep learning. Preprint at https://arxiv.org/abs/1811.04017 (2018).
98. Dahl, M. et al. Private machine learning in TensorFlow using secure computation. Privacy Preserving Machine Learning, NeurIPS 2018 Workshop, Montréal, December 8, 2018. Available at: https://arxiv.org/abs/1810.08130 (2018).
99. Brundage, M. et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. Preprint at https://arxiv.org/abs/2004.07213 (2020).
100. Tobin, A. & Reed, D. *The Inevitable Rise of Self-Sovereign Identity* (Sovrin Foundation, 2016).
101. Ghorbani, A. & Zou, J. Data Shapley: equitable valuation of data for machine learning. In *Proc. 36th Int. Conf. Machine Learning* (PMLR, 2019).
102. Elvy, S.-A. Paying for privacy and the personal data economy. *Colum. L. Rev.* **117**, 1369 (2017).

Acknowledgements

We thank A. Trask, J. Passerat-Palmbach and the OpenMined project members for their support and critical appraisal, and B. Farkas for creating the article's illustration.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to R.F.B.