https://doi.org/10.1038/s42256-024-00880-0

A question of trust for AI research in medicine

Medical research is one of the most impactful areas for machine learning applications, but access to large and diverse health datasets is needed for models to be useful. Winning trust from patients by demonstrating that data are handled securely and effectively is key.

Before computerized medical notes, patients relied on a doctor's implied trustworthiness to keep their medical records secure and private. However, the old and famously poorly handwritten notes are no more, with clinicians now relying on electronic versions of medical history, clinical data and results such as medical imaging data or laboratory tests, which are typically collated as Electronic Health Records (EHRs). In the UK, EHRs are allowed by law to be shared between qualified clinicians, creating an efficient and private ecosystem in which patients can visit different clinicians with various expertise who often understand the reasons for the patient's appointment before they've even walked through the door. The inherent nature of this trust is not implicitly afforded by patients to medical research initiatives however, and data sharing in this area is subject to a rigorous set of rules and criteria. Here arrive the speed bumps of progress: how can medical research make the most of advances in machine learning given the need for access to large amounts of patient genomic, diagnostic or imaging data?

In some cases, data can be anonymized and openly shared between researchers. Participants enrolled in studies that require medical data can opt for their data to be shared in openly available repositories, such as the Medical Segmentation Decathlon, a global collaborative effort sharing ten large medical imaging datasets, and the UK Biobank, a large-scale genetic, lifestyle and medical data repository for researchers. Research directives such as these have been highly beneficial, as data gathered from diverse sources with various characteristics can be used by many research communities. However, in many other cases research health data cannot be shared at all — for instance, when anonymization is difficult or when participants opt out of data sharing. To overcome these limitations and facilitate advances in medical applications of machine learning, federated learning schemes can be adopted, where each institute involved trains a model on their local data and shares only the model weight updates with others, so that datasets do not need to leave institutes' data centres. However, on its own, federated learning offers no guarantees that data are kept secure, and studies have shown that adversarial agents can reconstruct the original data from the model updates[1]. As Bak et al.[2] pointed out in a recent Comment, the use of federated learning does not relieve the data holders from their data protection responsibilities and additional privacy-preserving tools are needed.

A concern is that model performance is degraded by the addition of privacy-preserving tools, and it is generally accepted that there is a trade-off between accuracy and privacy. In an Article in this issue, Ziller et al. test this principle. They evaluate the use of differential privacy in machine learning research for medical imaging applications to counter the risk that training data can be reverse engineered from access to the trained model. Differential privacy is a common technique that involves adding a small amount of noise to the data, which reduces the effect that an individual patient's data has on the output of a machine learning model. The trade-off between privacy and accuracy can be controlled by the amount of noise added, a parameter known as a privacy budget. Ziller et al. provide empirical proof that what was originally considered an ineffective and overly generous privacy budget can be used to keep patient data effectively private in realistic attack scenarios while retaining the prediction accuracy of machine learning models. Based on these findings, the authors advocate for the consistent use of differential privacy, offering a path to homogenize and encourage the sharing of trained models.

Another noteworthy tool that is expected to help in protecting patient privacy is the use of synthetic data to either partially or wholly replace traditional medical data when training machine learning models. Generative artificial intelligence algorithms are now commonplace and can be used to curate or bolster existing medical databases with generated content. The advantage is that synthetic datasets do not contain any real personal information, thereby protecting original patient data from being shared. Instead, data are generated from scratch or by statistically modelling the original data. Although the use of synthetic data may boost the predictive accuracy of artificially intelligent algorithms for medical tasks, there are drawbacks. A potential concern is that for a training dataset with a ratio of synthetic to original data that weighs too heavily towards the synthetic side, bias, overfitting and generalization issues may arise, which hinders the development of artificial intelligence systems. Moreover, possible adversarial scenarios to reverse engineer the original data and mitigation strategies need to be carefully examined[3,4].

Rigorous testing of privacy-preserving tools in medical machine learning applications in real-world settings, as demonstrated by Ziller et al., should be strongly encouraged. Unambiguous evidence that data sharing methods can withstand privacy attacks, and that machine learning models are both secure and effective, will provide important incentives for patients to give consent to their health data being used for medical research. Retaining this willingness is essential for the success of machine learning applications, which depends on the availability of large and diverse real-world training datasets.

Published online: 24 July 2024

References

1. Geiping, J., Bauermeister, H., Dröge, H. & Moeller, M. In *Proc. 33rd International Conference on Advances in Neural Information Processing Systems* 16937–16947 (2020).
2. Bak, M. et al. Federated learning is not a cure-all for data ethics. *Nat. Mach. Intell.* **6**, 370–372 (2024).
3. Bak, M., Madai, V. I., Celi, L. A., Williamson, D. F. K. & Mahmood, F. Federated learning is not a cure-all for data ethics. *Nat. Mach. Intell.* **6**, 370–372 (2024).
4. Giuffrè, M. & Shung, D. L. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digit. Med.* **6**, 186 (2023).