nature genetics

PERSPECTIVE

https://doi.org/10.1038/s41588-020-0698-y

Treating medical data as a durable asset

Amalio Telenti[1]✉ and Xiaoqian Jiang[2]✉

Access to medical data is central for conducting research on genomics. However, to tap these metadata (observable traits and phenotypes, diagnoses and medication, and labels), researchers must grapple with the complex and sensitive nature of the information. In this Perspective, we argue that, at this exciting time for genomics and artificial intelligence, several critical aspects of data generation, infrastructure and management are pillars of a modern data ecosystem. Many risks to privacy and many obstacles to medical research can be eliminated or mitigated by new secure data analytics. Finally, we discuss the potential consequences of medical data exiting the institutions and being managed by individuals. These shifts in data ownership have the potential for profound disruption and opportunity across many fields.

Medical data are used by academia and medical institutions in research to improve human health. Medical and health data are also considered a valuable commodity by insurance companies, technology giants and countries. In reality, the access to and usability of medical data—as represented by the new field of medical data sciences—progresses slowly. These current limitations substantially affect the field of population and clinical genomics, which relies on the quality of data and on access to large numbers of individuals. In fact, genome analysis alone has limited value, but these data can be linked to electronic medical record (EMR) data to create research value. Even in the best-supported population initiatives, access to metadata (for example, phenotypes and demographics) is absent (for example, gnomAD, https://gnomad.broadinstitute.org/) or limited (for example, UK Biobank, https://www.ukbiobank.ac.uk/), and in many cases the data represent only a snapshot of a person in time. These limitations are present even though medical institutions generate massive amounts of qualitative and quantitative data[1]. The integration of multiple modalities involves data sharing and analysis, which are strongly connected with other aspects of data protection and liability issues.

A lack of unified nomenclature exemplifies the different cultures and technical worlds that use medical data: the field of population genomics refers to metadata (including observable traits and phenotypes); the field of medicine uses the term diagnosis; and the field of machine learning uses labels. At a deeper level, understanding of the requirements for effective and responsible use (minimum necessary) of medical information remains incomplete. Progress is unequal: there is a developing corpus of regulations and requirements aiming to uphold the ethical use and protection of the data. There is also a large capacity for generating human sequence data and other types of large-scale data that are highly digitized, such as medical images. Progress in artificial intelligence (machine and deep learning) has been remarkable in terms of computational power and algorithms. The number of deep-learning applications in genomics is increasing[2]. Bottlenecks exist at the very basic level of data management, including standardization, capture, storage and retrieval. In addition, there is limited awareness and a lack of implementation of emerging technologies for data protection and secure analytics. Themes of genomics, bioinformatics, genomic medicine, ethics, data sharing, privacy and community engagement are particularly relevant to the Electronic Medical Records and Genomics Network (eMERGE; https://emerge-network.org/).

In this Perspective, we do not offer a detailed review of the current status of medical data sciences. Instead, we specifically focus on the management of healthcare research data, with an emphasis on relevance to genomics research. We highlight areas that are pillars supporting the use of data for analytics and emergent technologies supporting secure analytics. Finally, we describe the current trends of shifting the ownership of medical data to the end user or patient. Effectively, these trends result in medical data leaving the institutions and entering the public arena, thus creating a new balance that may facilitate research but also may void some existing guardrails for data protection. The various evolving concepts in data usage and analytics discussed herein are represented in Fig. 1.

Notably, an overarching theme in this Perspective—as indicated in the title—is the concept of medical data as an asset. An asset is defined in the accounting and finance conceptual frameworks[3] as "a resource controlled by the entity as a result of past events and from which future economic benefits are expected to flow to the entity." This concept is important because several components of this definition—'controlling entity', 'result from past events' and 'future benefits'—are integral to the present perspective.

Data management and data infrastructure

EMRs are complex. They include unstructured (for example, free text), structured (for example, insurance codes and claims) and processed data (for example, laboratory quantitative results) (Fig. 2). They do not include easy access to the raw data, whether they be sequence data represented by FASTQ files that are not attached to a genome sequencing report or raw data from imaging or other medical instruments. Altogether, a 'classical' EMR may comprise 3.5 gigabytes of data, whereas experimental deep-phenotyping medical exams including high-content imaging and genomics[4] may generate more than 120 gigabytes of data for a single individual (including approximately 120 gigabytes for genomes and 6 gigabytes for exomes[5]). The diversity of records and the fragmentation of data across platforms pose challenges in downstream analysis. Therefore, modernizing the very basic infrastructures for data acquisition, storage and management is a priority. A sane practice would call for standards for data lakes (repositories storing large sets of data in native format), databases and application programming interfaces (APIs) to support diverse uses.

Health systems are expected to generate more than 2,000 exabytes of data per year (one exabyte is one billion gigabytes), includ-

[1]Department of Integrative Structural and Computational Biology, Scripps Research Institute, La Jolla, CA, USA. [2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA. ✉e-mail: atelenti@scripps.edu; xiaoqian.jiang@uth.tmc.edu

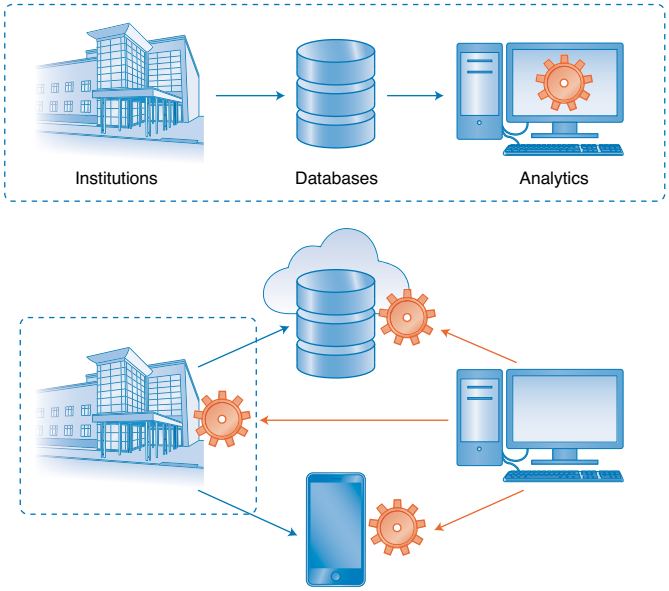

Fig. 1 | Evolving concepts in data usage and analytics. Traditionally, research has been conducted within institutional confinement (dotted line) in compliance with policies and regulations (top). Currently, data flows, infrastructures and analytics are changing (bottom), with models that contemplate data moving outside the physical limits of the institution (dotted line), owing to cloud computing, with the possibility of bringing users' data to their personal devices. Increasingly, analytics (orange cogwheels) is brought to the data where they reside and even deployed 'at the edge', on personal devices. End users or patients may or may not share results with the institution.

ing the generation of approximately 25 petabytes of genomic data annually worldwide by 2030 (ref. [6]). Hospital information technology (IT) systems may be ill prepared to manage the large volumes of data currently generated by imaging and genomics. The adoption of cloud services remains questioned by some in the medical field. A cloud-storage service must sign a business-associate agreement with the medical institution and must be Health Insurance Portability and Accountability Act of 1996 (HIPAA) compliant. Several commercial cloud offerings include encryption of data in transit and at rest and enable two-step authentication.

Data access and retrieval are generally viewed from the standpoint of institutional needs, such as for billing purposes, whereas less attention is given to the end user or patient. We have argued in the past that medical-record documents should be easily and intuitively searchable and retrievable so that patients can query for any term and find their related health documents or images quickly[7]. This concept, reflecting how easily people interact with tools such as Google, suggests that access to EMR files should also be built by using consumer-centric technologies. In genomics, the concept can also be extended to clinical or biomedical researchers who could benefit from search engines to rapidly query genome variants—an approach that we have used to support the searchability of the non-coding genome (https://omni.telentilab.com/).

Data analytics

An EMR stores patients' demographics, medical history and diagnosis, laboratory results and images, interventions, medications and outcomes. Data can be turned into more valuable resources (that is, information) through processing and analysis. Many components of the EMR require extensive normalization and coding for downstream use (Fig. 2). Because medical records are increasingly being acquired and stored in digital form, the EMR is the focus of the attention of machine learners[8], who either create or adapt from other fields an expanding number of algorithms. Various protocols and toolkits tailor natural language processing (NLP) for clinical text (for example, Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP), Word2Vec and Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (bioBERT)). NLP extracts knowledge of narratives to generate word or sentence vectors that can be understood by the learning algorithms. Popular EMR learning algorithms include autoencoders for nonlinear dimensionality reduction, convolutional neural networks for image analysis and deep attention models for processing clinical notes.

Transfer learning is particularly useful in medicine because of the current paucity of ground-truth clinical data for training[9]. Transfer learning is the idea of reusing model structure (or so-called stored knowledge) to address one problem in handling a different question with some relationship to the original problem, for example, fine-tuning bioBERT, a new method of pretraining language representations.

New algorithms may improve classification accuracy (the ratio of the number of correct predictions to the total number of input samples) by a few, but impactful, percentage points. Many clinical practitioners take point improvements in model performance at face value and underestimate the effects of small gains on prediction accuracy. When Google Translate pivoted from using statistical models to neural networks, the average accuracy of translation between English and several languages increased from 76% to 83% (highlighted in ref. [10]). This 7% improvement caused the machine translation to approach the accuracy of human translation. Similarly, for Google's voice-recognition software, increasing the accuracy rate from 90% to 95%, the threshold for human accuracy in voice recognition, took 2.5 years (highlighted in ref. [10]).

Despite the progress in data analytics, better algorithms may not substitute for the absence of large amounts of labeled, quality data, domain knowledge or solid data infrastructures. Specific to phenome-wide association studies (for example, https://phewascatalog.org/) and to the investigation of rare diseases (for example, http://www.orphadata.org/), access to clinical metadata remains the main bottleneck in the field.

Data portability, ownership and markets

Healthcare data are a unique asset. They can be copied and disseminated quickly, unlike traditional physical assets, which can be controlled by one entity at a time. They are unique (nobody is the same, and no combination is the same), additive (more data enable better understanding of the characteristics and strong statistical power in hypothesis testing), non-depletive (they are not a consumable resource) and replicable (as a digital resource that can be copied to numerous parties quickly). These properties also create unique challenges in controlling data assets.

Data portability is the users' right to control the free movement of their data between alternative service providers, thus encouraging competition within free markets. It is also a critical component in the European Union's General Data Protection Regulation (GDPR)[11]. Data portability in the US context, however, implies weakened protection of health data because of the limited scope of 'covered entities', which are responsible for the security and privacy of personal health information[12]. The disruption of data portability is further aggravated because of the interests of data 'controllers' (for example, direct-to-consumer genomic-testing companies, genealogy websites and insurance companies), the formidable complexity of healthcare IT infrastructure and the lack of universal information exchange protocols. Digital giants, such as Apple, are trying to support data portability by connecting miscellaneous healthcare providers[13] to allow patients to download their medical

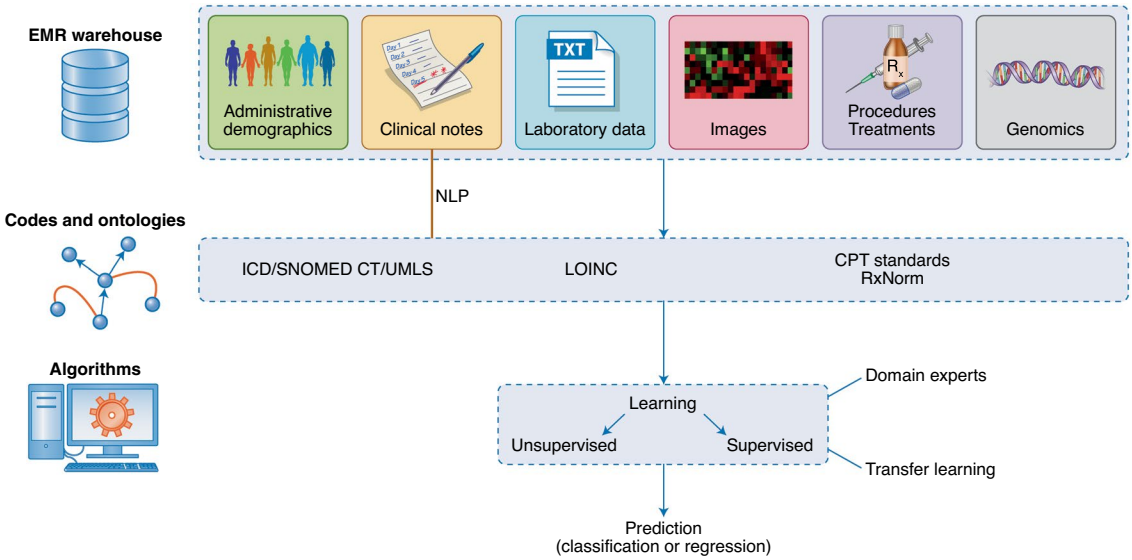

Fig. 2 | Structure, management and analysis of EMRs. Many components of EMRs require extensive normalization and coding for downstream use. International Classification of Diseases (ICD) code diagnoses, symptoms and procedures, Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) and United Medical Language System (UMLS) are standard ontologies for medical terminologies. Logical Observation Identifiers Names and Codes (LOINC) is used for laboratory measurements. Current Procedural Terminology (CPT) provides classification codes for treatment procedures. RxNorm provides standardized nomenclature for clinical drugs. Clinical notes require NLP.

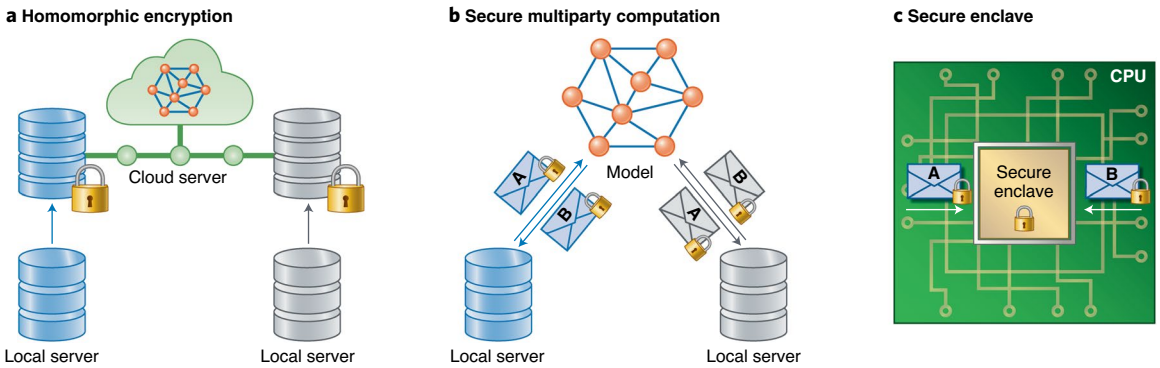

Fig. 3 | Operations on encrypted data. Illustration of different mechanisms of three cryptographic frameworks for encrypted operations. **a**, Homomorphic encryption: analyzing encrypted data with public or private algorithms. **b**, Secure multiparty computation: securely evaluating protocols across multiple data holders. **c**, Secure enclave: analyzing data with hardware-assisted encryption.

records through HealthKit. A major issue for data portability is the adoption of standards, and some active efforts, such as Health Level Seven International (HL7) and Fast Healthcare Interoperability Resources (FHIR), are available for exchanging electronic health records (EHRs) and Digital Imaging and Communications in Medicine (DICOM) to communicate medical imaging information. Importantly, we do not discuss the associated liability in managing data; liability risks in the field of genomics have been addressed in a recent review[14]. We also do not discuss that the ethical issues are profound when patient data become an asset and trained machine-learning models are used without consent[15,16]. There is a potential risk of release of private information, because adversaries could use the models to reverse-engineer and disclose sensitive information of participants involved in the training dataset[17]. We believe that an informed-consent mechanism or educational approach should be used to inform the general population on these implicit ethical issues related to privacy.

Initiatives such as that of Apple HealthKit leave the door wide open for third-party developers to create apps that focus on health. More generally, more than 45,000 healthcare apps were available as of the third quarter of 2019 (ref. [18]). The confluence of medical records with a wider set of health measurements creates a coexistence of EMR and EHR. These terms are used interchangeably; however, whereas medical-related attributes are part of health, not all health attributes are medical. The health record is a larger concept than the medical record, because it includes aspects such as activity, behavioral patterns and diet preference. The distinction also affects data ownership, discussed in the next section.

Health records are generated by individuals and belong to individuals as an asset. Medical records are shared between providers and patients to provide the necessary care and support the well-being of all society. Researchers, insurance and pharmaceutical companies are keen to access personal health data to gain a deeper understanding of diagnostics, disease development and potential preventive or

Table 1 | High-level comparison of different secure frameworks: performance and assumptions

	Homomorphic encryption	Secure multiparty computation	Secure enclave
Memory overhead	High	High	Low
Communication overhead	Medium	High	Low
Computation overhead	Medium/High	Medium	Low
Provable guarantee	Yes	Yes	No
Encrypted data storage	Yes	Yes	Yes
Offline operation	Yes	No	Yes
Requirement for special hardware	No	No	Yes
Virtual memory limitation	No	No	Yes
Summary	Most appropriate for secure outsourcing when data owners want to delegate the computation to a third party	Enables secure evaluation of a circuit (for example, function) on encrypted genomics data from different sources; it is most applicable to two or more collaborators who want to conduct a common study using combined data without revealing private information	Based on a hardware feature of certain central processing units; it is itself a mini-computer with encrypted memory and boots separately from the main device, thus making hacking difficult, because the primary operating system cannot see the decryption keys

treatment options. These demands create a large market in which data brokers (for example, Zenome, CoverUS, LunaDNA, Doc.ai, Medicalchain, ProofWork and Nebula Genomics) connect isolated silos to collect, integrate and monetize data for their customers. Commercial enterprises in genomics (for example, 23andMe, AncestryDNA, MyHeritageDNA and others) build value by using customers' data—users should attentively read the terms and conditions and privacy policies to assess each company's use of genetic data and phenotypic information. This eruption of the entrepreneurial sector could be seen as an opportunity, because it marshals a different type of incentives for the participation of individuals who contribute data. The healthcare data analytics market is expected to grow to US$47.7 billion by 2024. Despite the perceived economic value, health and medical data are often not managed as an asset.

Data protection and novel technologies for risk mitigation

The key for broad access to medical data relies on solving critical issues of privacy protection. Traditional approaches of data-protection technology, such as Advanced Encryption Standard (AES)[19], are implemented in cloud storage, thus ensuring strong encryption in transit and at rest. However, these methods do not protect against insider attacks or hackers who gain access to the remote server running analysis on decrypted data. Because encrypted data must be decrypted for computation (in untrusted servers), there is a large exposure of vulnerable attack surfaces.

A new, recently emerged category of encryption frameworks called encrypted operations allows operations to be performed on encrypted data without exposing their content. The calculated results are returned in an encrypted format for decryption, thereby ensuring zero information leakage during the entire life cycle of the data, from communication to storage to computation. Because these frameworks are relatively new and sometimes theoretical, they are not as widely known as the 'encryption at rest' frameworks. However, substantial progress has been made in encryption operations in recent years, and they have begun to demonstrate practical usability. The basic architecture of these three different frameworks is illustrated in Fig. 3.

Because of the novelty of these approaches and the great interest in the field of genomics (for example, GenoPri Consortium, https://www.genopri.org/; and iDASH workshop, http://www.humangenomeprivacy.org/), we summarize some of their key characteristics (Table 1). Homomorphic encryption is often deployed when users want to access the analytic models built by others using private data or offload computational tasks to cost-effective third-party computation services[20,21]. Secure multiparty computation has been applied for enabling pharmacological collaboration[22], genome-wide association analysis[23] and privacy-preserving distributed genomic tests for HIV treatment[24]. The major difference is that the secure, multiparty computation framework converts data into secrets distributed among multiple entities to support secure collaboration, whereas the homomorphic encryption framework outsources encrypted data that a single entity could store and run unforeseen algorithms upon. The secure enclave model[25] is highly flexible and can accommodate many tasks, including whole-genome-variant searching[26]. However, it does not offer the same level of security protection as homomorphic encryption or secure multiparty computation, which can be mathematically proven to safeguard information. These frameworks can also be used in combination to provide the best efficiency for certain tasks (for example, MedCo[27], a system that combines homomorphic encryption and secure multiparty computation to protect private data sharing while facilitating medical research on pathologies and COVID-19 (ref. [28]) across several hospitals).

Algorithms that go to the data and smart contracts

The traditional one-institution, one-data-center model is changing. Many hospitals are moving their data to the cloud (for example, the University of California, San Francisco recently established a research cloud on Amazon Web Service), thus allowing third-party algorithms to be run on the data without moving the data to the algorithm developers. When most providers host their EMR data in the cloud, the concept of physical isolation is blurred, although logical compartmentalization remains to satisfy institutional policies and regulations. On the end-user or patient side, edge devices, such as smartphones, are becoming increasingly powerful. Personal health information, such as exercise, vital statistics and physiological data, now resides on personal devices, and many apps are running directly on these data to analyze personal health without communicating with a central server. Variant call format (VCF) genome sequence files can feasibly reside on a smartphone. Bringing algorithms to data (for example, 'edge AI', in which AI algorithms are processed locally on a hardware device) has a dual benefit of distributing the computation and protecting privacy.

The last component of the emerging field of portable medical and health data is the ability to create smart contracts with third parties. The value of EHRs or EMRs is realized through active use rather

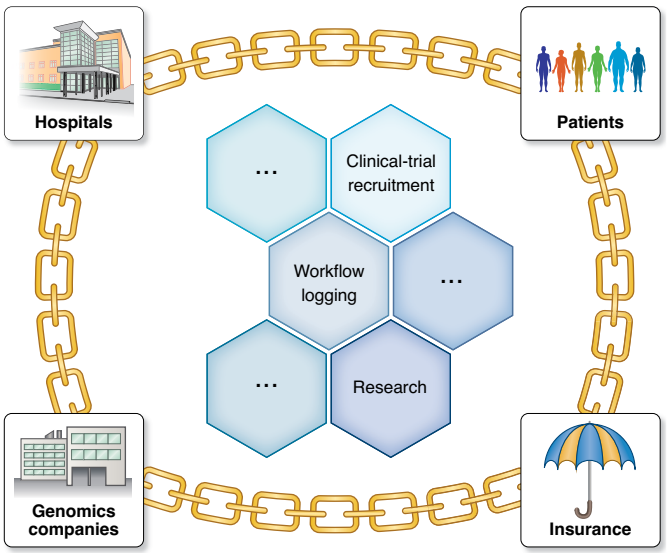

Fig. 4 | Blockchain serves as an immutable ledger to connect different users. A blockchain can create and maintain digital references of the medical data assets (for example, prospective study participants) as well as digital property certificates for assigning access rights to the corresponding medical data assets.

than depositing them in storage. In that sense, they are digitized liquid assets of individuals. The emergence of blockchain technology provides personal vaults to store digital information in an immutable distributed ledger, which creates an ecosystem for information exchange and utilization. Of note, immutability does not mean privacy, and the nature of the liquidity requires high-level consideration of protection of sensitive information. The blockchain technology also expands the traditional trust-based contractual system to executable protocols (via smart contract) in a decentralized and distributed environment (Fig. 4).

These new features broaden the horizon for novel applications and potential market space for personal health data. For example, we have recently participated in research[29] using property-rights blockchain to match patients to clinical trials. We envision innovative blockchain-based data marketplaces coupled with personal medical data vaults as a new model for managing the transfer, provenance and processing of individual health information. However, despite this excitement, lasting adoption of blockchain technologies in genomics remains to be seen.

The combination of data and algorithms moving to edge devices has profound implications for the healthcare industry, because individuals will be in full control of their own data, and medical service providers may no longer act as the central data custodians for all personal medical information. Improved privacy protection also brings challenges in data sharing, because the big-population model still requires the abilities of information exchange and synthesis to ensure robustness. The recent surge in federated learning (that is, learning a shared prediction model while keeping all training data on edge devices) has shed light on such challenges, which ensure privacy by design in constructing global models.

Conclusions

We wrote this Perspective with the goal of familiarizing the biomedical research community in general, particularly the genomics and genetics community, with the progress, challenges and opportunities in the use of EMR data. Progress has included the creation of improved data infrastructures that support access to the heterogeneous content of clinical records. Challenges include the implementation and acceptance of novel enabling encryption technologies that allow for interinstitutional collaboration and research without compromising data security and privacy. Opportunities include extension of the health space to the private domain (personal edge devices, apps and social media) with the potential to secure ownership of data. The overarching principle is that data are a durable asset: they have an intrinsic value that extends beyond the original purpose of why and when they were collected.

Received: 30 March 2020; Accepted: 21 August 2020;
Published online: 14 September 2020

References

1. Telenti, A. Machine learning to decode genomics. *Clin. Chem.* **66**, 45–47 (2020).
2. Zou, J. et al. A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
3. Morgan, E. & Prowle, M. (eds.) *Financial Management and Control in Higher Education* (Taylor & Francis, 2004).
4. Shomorony, I. et al. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med.* **12**, 7 (2020).
5. Krumm, N. & Hoffman, N. Practical estimation of cloud storage costs for clinical genomic data. *Pr. Lab Med* **21**, e00168 (2020).
6. Banks, M. A. Sizing up big data. *Nat. Med.* **26**, 5–6 (2020).
7. Telenti, A., Steinhubl, S. R. & Topol, E. J. Rethinking the medical record. *Lancet* **391**, 1013 (2018).
8. Harerimana, G., Kim, J. W., Yoo, H. & Jang, B. Deep learning for electronic health records analytics. *IEEE Access* **7**, 101245–101259 (2019).
9. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2019).
10. Telenti, A. Council post: do we need more data or more science in data science? *Forbes* (20 February 2020).
11. Hert, P. D. et al. The right to data portability in the GDPR: towards user-centric interoperability of digital services. *Comput. Law Secur. Rev.* **34**, 193–203 (2018).
12. Forcier, M. B., Gallois, H., Mullan, S. & Joly, Y. Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers? *J. Law Biosci.* **6**, 317–335 (2019).
13. Institutions that support health records on iPhone and iPod touch. *Apple.com* https://support.apple.com/en-us/HT208647 (2020).
14. Marchant, G., Barnes, M., Evans, J. P., LeRoy, B. & Wolf, S. M. From genetics to genomics: facing the liability implications in clinical care. *J. Law Med. Ethics* **48**, 11–43 (2020).
15. Ienca, M. et al. Considerations for ethics review of big data health research: a scoping review. *PLoS ONE* **13**, e0204937 (2018).
16. Goodman, K., Zandi, D., Reis, A. & Vayena, E. Balancing risks and benefits of artificial intelligence in the health sector. *Bull. World Health Organ.* **98**, 230–230A (2020).
17. Pan, X., Zhang, M., Ji, S. & Yang, M. Privacy risks of general-purpose language models. in *2020 IEEE Symposium on Security and Privacy* 1314–1331 (IEEE, 2020).
18. Number of mHealth apps available in the Apple App Store from 1st quarter 2015 to 1st quarter 2020. *Statista.com* https://www.statista.com/statistics/779910/health-apps-available-ios-worldwide (2020).
19. Hathaway, L. *National Policy on the Use of the Advanced Encryption Standard (AES) to Protect National Security Systems and National Security Information*. CNSS Policy 15, Fact Sheet 1 (National Security Agency, 2003); https://csrc.nist.gov/csrc/media/projects/cryptographic-module-validation-program/documents/cnss15fs.pdf
20. Jiang, X., Kim, M., Lauter, K. & Song, Y. Secure outsourced matrix computation and application to neural networks. *Conf. Comput Commun. Secur* **2018**, 1209–1222 (2018).
21. Kim, M. & Lauter, K. Private genome analysis through homomorphic encryption. *BMC Med. Inform. Decis. Mak.* **15**, S3 (2015). (Suppl. 5).
22. Hie, B., Cho, H. & Berger, B. Realizing private and practical pharmacological collaboration. *Science* **362**, 347–350 (2018).
23. Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
24. McLaren, P. J. et al. Privacy-preserving genomic testing in the clinic: a model using HIV treatment. *Genet. Med.* **18**, 814–822 (2016).
25. Chen, F. et al. PRINCESS: privacy-protecting rare disease international network collaboration via encryption through software guard extensions. *Bioinformatics* **33**, 871–878 (2017).

26. Kockan, C. et al. Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nat. Meth.* **17**, 295–301 (2020).
27. Raisaro, J. L. et al. MedCo: enabling secure and privacy-preserving exploration of distributed clinical and genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **16**, 1328–1341 (2019).
28. Raisaro, J.L. et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. *JAMA* https://doi.org/10.1093/jamia/ocaa172 (2020).
29. Bergeron, J. et al. Simulating patient matching to clinical trials using a property rights blockchain. *Digit. Med.* **6**, 44–52 (2020).

Acknowledgements

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of Vir Biotechnology or any other associated individual, employee, agency, organization or company.

Author contributions

A.T. and X.J. conceived this work and contributed equally to its writing.

Competing interests

A.T. an employee of Vir Biotechnology.

Additional information

Correspondence should be addressed to A.T. or X.J.