review articles

Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care

Fadila Zerka, PhD[1,2]; Samir Barakat, MSc, PhD[1,2]; Sean Walsh, MSc, PhD[1,2]; Marta Bogowicz, PhD[1,3]; Ralph T. H. Leijenaar, MSc, PhD[1,2]; Arthur Jochems, PhD[1]; Benjamin Miraglio, PhD[2]; David Townend, LLB, MPhil, PhD[4]; and Philippe Lambin, MD, PhD[1]

abstract

Big data for health care is one of the potential solutions to deal with the numerous challenges of health care, such as rising cost, aging population, precision medicine, universal health coverage, and the increase of noncommunicable diseases. However, data centralization for big data raises privacy and regulatory concerns.

Covered topics include (1) an introduction to privacy of patient data and distributed learning as a potential solution to preserving these data, a description of the legal context for patient data research, and a definition of machine/deep learning concepts; (2) a presentation of the adopted review protocol; (3) a presentation of the search results; and (4) a discussion of the findings, limitations of the review, and future perspectives.

Distributed learning from federated databases makes data centralization unnecessary. Distributed algorithms iteratively analyze separate databases, essentially sharing research questions and answers between databases instead of sharing the data. In other words, one can learn from separate and isolated datasets without patient data ever leaving the individual clinical institutes.

Distributed learning promises great potential to facilitate big data for medical application, in particular for international consortiums. Our purpose is to review the major implementations of distributed learning in health care.

JCO Clin Cancer Inform 4:184-200. © 2020 by American Society of Clinical Oncology

Licensed under the Creative Commons Attribution 4.0 License

INTRODUCTION

Law and ethics seek to produce a governance framework for the processing of patient data that produces a solution to the issues that arise between the competing desires of individuals in society for privacy and advances in health care. Traditional safeguards to achieve this governance have come from, for example, the anonymization of data or informed consent. These are not adequate safeguards for the new big data and artificial intelligence methodologies in research; it is increasingly difficult to create anonymous data (rather than pseudonymized/coded data) or to maintain it against re-identification (through linking of datasets causing accidental or deliberate re-identification). The technology of big data and artificial intelligence, however, itself increasingly offers safeguards to solve the governance problem. In this article we explore how privacy-preserving distributed machine learning from federated databases might assist governance in health care. The article first outlines the basic parameters of the law and ethics issues and then discusses machine learning and deep learning. Thereafter, the results of the review are presented and discussed. The methodology for this research is that distributed machine learning is an evolving field in computing, with 665 articles published between 2001 and 2018; the study is based on a literature search, focuses on the medical applications of distributed machine learning, and provides an up-to-date summary of the field.

THE LEGAL CONTEXT FOR PATIENT DATA RESEARCH

The challenges in law and ethics in relation to big data and artificial intelligence are well documented and discussed[1-16]. The issue is one of balance: privacy of health data and access to data for research. This issue is likely to become more pronounced with the foreseeable developments in health care, notably in relation to rising cost, aging population, precision medicine, universal health coverage, and the increase of noncommunicable diseases. However, recent developments in law, for example, in the European Union's General Data Protection Regulation (GDPR), appear to maintain the traditional approach that seems to favor individualism above solidarity. Individualism is strengthened in the new legislation. There is

ASSOCIATED CONTENT

Appendix

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on January 16, 2020 and published at ascopubs.org/journal/cci on March 5, 2020: DOI https://doi.org/10.1200/CCI.19.00047

CONTEXT

Key Objective

Review the contribution of distributed learning to preserve data privacy in health care.

Knowledge Generated

Data in health care are greatly protected; therefore, accessing medical data is restricted by law and ethics. This restriction has led to a change in research practice to adapt to new regulations. Distributed learning makes it possible to learn from medical data without these data ever leaving the medical institutions.

Relevance

Distributed learning allows learning from medical data while guaranteeing preservation of patient privacy.

a narrowing of the definition of informed consent in Article 4.11 of the GDPR, with the unclear inclusion of the necessity for broad consent in scientific research included in Recital 33.

In relation to the continuing ambiguity of the unclear legal landscape for research using and reusing large datasets and linking between datasets, the GDPR is not clear in the area of re-identification of individuals. For the GDPR, part of the problem is clear—when data have the potential when added to other data to identify an individual, then those data are personal data and subject to regulation. The question is, is this absolute (any possibility, regardless of remoteness), or is there a reasonableness test? Recital 26 includes such a reasonable test: "To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."[16a]

From this overview of legal difficulties, it is clear that there are obstacles to processing data in big data, machine learning, and artificial intelligence methodologies and environments. It must be stressed that the object is not to circumvent the rights of patients or to suggest that privacy should be ignored. The difficulty is that where the law is unclear, there is a tendency toward restrictive readings of the law to avoid liability, and, in the case of the methodologies and applications of data science discussed here, the effect of unclear law and restrictive interpretations of the law will be to block potentially important medical and scientific developments and research. Each of the uncertainties will require regulators to take a position on the best interpretation of the meaning of the law according to the available safeguards. The question for the data science community is, how far can that community itself address concerns about privacy, about re-identification, and about safeguarding autonomy of individuals and their legitimate expectations to dignity in their treatment through the proper treatment of their personal data? How far distributed learning might contribute a suitable safeguard is the question addressed in the remainder of this paper.

MACHINE LEARNING

Machine learning comes from the possibility to apply algorithms on raw data to acquire knowledge.[1] These algorithms are implemented to support decision making in different domains, including health care, manufacturing, education, financial modeling, and marketing.[2,3] In medical disciplines, machine learning has contributed to improving the efficiency of clinical trials and decision-making processes. Some examples of machine learning applications in medicine are the localization of thoracic diseases,[4] early diagnosis of Alzheimer disease,[5] personalized treatment,[6] outcome prediction,[7,8] and automated radiology reports.[9]

There are three main categories of machine learning algorithms. First, in supervised learning, the algorithm generates a function for mapping input variables to output variables. In unsupervised learning, the applied algorithms do not have any outcome variable to estimate, and the algorithms generate a function mapping for the structure of the data. The third type is referred to as reinforcement learning, whereby in the absence of a training dataset the algorithm trains itself by learning from experiences to make increasingly improved decisions. A reinforcement agent decides what action to perform to accomplish a given task.[10,11] Table 1 provides a brief description of selected popular machine learning algorithms across the three categories.

DEEP LEARNING

Deep learning is a subset of machine learning, which, in turn, is a subset of artificial intelligence,[12] as represented in Figure 1. The learning process of a deep neural network architecture cascades through multiple nodes in multiple layers, where nodes and layers use the output of the previous nodes and layers as input.[13] The output of a node is calculated by applying an activation function to the weighted average of this node's input. As described by Andrew Ng[14], "The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data that we can feed in to these algorithms," meaning that the more data are fed into the model the better the performance. Yet, this continuous improvement of the performance in concordance with the

TABLE 1. Examples of Machine Learning Algorithms

Example	Algorithm	Description	Distributed Algorithm Available?
Supervised learning	SVM	An algorithm performing classification tasks by composing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.[2,3]	Yes[2]
	Logistic regression	An algorithm used for discrete values estimation tasks based on a given set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function that maps probabilities with values lying between 0 and 1. Hence, it is also known as logit regression.[81,82]	Yes[83-85]
	Decision tree	A nonparametric method used for both classification and regression problems. As the name suggests, it uses a tree-like decision model by splitting a dataset into two or more subsets on the basis of conditional control statements. It can also be used to visually represent decision-making processes.[10]	Yes[86,87]
	Random forest	Random forest is a method performing classification and regression tasks. A random forest is a composition of two initials: forest because it represents a collection of decision trees, and random because the forest randomly selects observations and features from which it puts up various decision trees. The results are then averaged. Each decision tree in the forest has access to a random set of the training data, chooses a class, and the most selected class is then the predicted class.[10]	No
	KNN	KNN can be used for regression problems; however, it is widely used for classification problems. In KNN, the assumption is that similar data elements are close to each other. Given K (positive integer) and a test observation, KNN first groups the K closest elements to the test observation. Then, in the case of regression, it returns the mean of the K labels, or in the case of classification, it returns the mode of the K labels.[10]	Yes[88,89]
Unsupervised learning	K-means	An algorithm mainly used for clustering in data mining. K-means nomination comes from its functionality, which is partitioning of N observations to K clusters, where each and every observation is part of the cluster with the nearest mean.[90]	Yes[90-93]
	Apriori algorithm	A classic algorithm in data mining. Used for mining frequent item groups and relevant association rules in a transactional database.[94]	Yes[95]

(Continued on following page)

TABLE 1. Examples of Machine Learning Algorithms (Continued)

Example	Algorithm	Description	Distributed Algorithm Available?
Reinforcement learning	MDP	Introduced in 1950s,[96] MDP is a discrete stochastic control process providing a framework for modeling decision making when final outcomes are ambiguous. Given S_O (current state) and S (new state), the decision process is made by steps. The process has a state at each step and the decision maker can choose any available action in S_O. The process then moves randomly into S, the new state. The chosen action influences the probability of moving from S_O to S. In other words, the next state depends only on the current state, not the previous states, and the action taken by the decision maker, satisfying the Markovian property, from which comes the algorithm's name.[97]	No
	Q-learning	Useful for optimization of action selection of any finite MDP. Q-learning algorithm provides agents in a process the ability to know what action to take under what situation.[98]	Yes[99]

Abbreviations: KNN, K-nearest neighbors; MDP, Markov decision process; SVM, support vector machine.

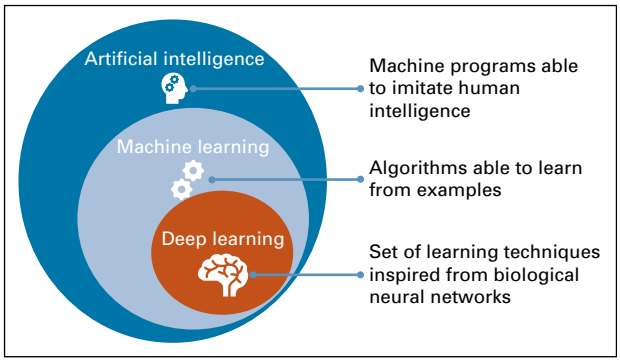

FIG 1. Relationship between artificial intelligence, machine learning, and deep learning.

amount of the data are not correct for traditional machine learning algorithms reaching a steady performance level that does not improve with the increase of the amount of the training data.[15]

METHODS AND MATERIAL SELECTION

A PubMed search was performed to collect relevant studies concerning the utilization of distributed machine learning in medicine. We used the search strings: "distributed learning," "distributed machine learning," and "privacy preserving data mining." The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement was adopted to select and compare distributed learning literature.[16] The PRISMA flow diagram and checklist are slightly modified and presented in Appendix Figure A1 and Appendix Table A1, respectively. The last search for distributed machine learning articles was performed on February 28, 2019.

SEARCH RESULTS

A total of 127 articles were identified in PubMed using the search query: ("distributed learning" OR "distributed machine learning" OR "privacy preserving data mining"). Six papers were screened; a brief summary of each article is presented in Table 2.

DISTRIBUTED LEARNING

Distributed learning ensures data safety by only sharing mathematical parameters (or metadata) and not the actual data or in any instance data that might enable tracking back the patient information (such as patient ID, name, or date of birth). In other words, distributed algorithms iteratively analyze separate databases and return the same solution as if data were centralized, essentially sharing research questions and answers between databases instead of data.[17] Also, before processing with the learning process, researchers must make sure all data have been successfully anonymized and secured by means of hashing algorithms and semantic web techniques, respectively, as can be seen in Figure 2, in addition to post-processing methods to address the multicenter variabilities.[19]

Distributed Machine Learning

A large quantity of training data is required for machine learning to be applied, especially in outcome modeling, where multiple factors influence learning. Provided there are sufficient and appropriate data, machine learning typically results in accurate and generalizable models.[20,21] However, the sensitivity of the personal data greatly hinders the conventional centralized approach to machine learning, whereby all data are gathered in a single data store. Distributed machine learning resolves legal and ethical privacy concerns by learning without the personal data ever leaving the firewall of the medical centers.[22]

The euroCAT[23] and ukCAT[24] projects are a proof of distributed learning being successfully implemented into clinical settings to overcome data access restrictions. The purpose of the euroCAT project was to predict patient outcomes (eg, post-radiotherapy dyspnea for patients with lung cancer) by learning from data stored within clinics without sharing any of the medical data.

Distributed Deep Learning

Training a deep learning model typically requires thousands to millions of data points and is therefore computationally expensive as well as time consuming. These challenges can be mitigated with different approaches. First, because it is possible to train deep learning models in a parallelized fashion,[25] using dedicated hardware (graphics processing units, tensor processing units)[26] reduces the computational time. Second, as the memory of this dedicated hardware is often limited, it is possible to divide the training data into subsets called batches. In this situation, the training process iterates over the batches, only considering the data of one batch at each iteration.[27] On top of easing the computing burden, using small batches during training improves the model's ability to generalize.[28]

These approaches address computation challenges but do not necessarily preserve data privacy. As for machine learning, deep learning can be distributed to protect patient data.[29,30] Moreover, distributed deep learning also improves computing performance, as in the case of wireless sensor networks, where centralized learning is inefficient in terms of both communication and energy.[31,32]

An example of distributed deep learning in the medical domain is that of Chang et al,[33] who deployed a deep learning model across four medical institutions for image classification purposes using three distinct datasets: retinal fundus, mammography, and ImageNet. The results were compared with the same deep learning model trained on centrally hosted data. The comparison showed that the distributed model accuracy is similar to the centrally hosted model.[33] In a different study, McClure et al[34] developed a distributed deep neural network model to reproduce FreeSurfer brain segmentation. FreeSurfer is an open source tool for preprocessing and analyzing (segmentation,

 © 2020 by American Society of Clinical Oncology

TABLE 2. Summary of Methods and Results of Distributed Machine Learning Studies Grouping More Than One Health Care Center

Reference	Data and Target	Methods and Distributed Learning Approach	Tools	Accomplishments and Results
Jochems[61]	Clinical data from 287 patients with lung cancer, treated with curative intent with CRT or RT alone were collected and stored in five different medical institutes: MAASTRO: (the Netherlands), Jessa (Belgium), Liège (Belgium), Aachen (Germany), and Eindhoven (the Netherlands).	A Bayesian network model is adapted for distributed learning using data from five hospitals.	Varian learning portal	AUC, 0.61
	Target: predict dyspnea.	Patient data were extracted and stored locally in the hospitals. Only the weights were then sent to the master server.		
Deist[23]	Clinical data from 268 patients with NSCLC from five different medical institutes: Aachen (Germany), Eindhoven (The Netherlands), Hasselt (Belgium), Liège (Belgium), and Maastricht (the Netherlands).	Alternating Direction Method of Multipliers was used to learn SVM models.	Varian learning portal	AUC, 0.62 for training set and 0.66 for validation set
	Target: predict dyspnea grade ≥ 2.	The data were processed simultaneously in local databases. Then, the updated model parameters were sent to the master machine to compare and update them and check if the learning process has converged. The process is repeated until convergence criteria were met.		
Dluhoš[66]	258 patients with first-episode schizophrenia and 222 healthy controls originating from four datasets were collected: two datasets from University Hospital Brno (Czech Republic), University Medical Center Utrecht (The Netherlands), and the last dataset originates from the Prague Psychiatric Center and Psychiatric Hospital Bohnice.	All images were preprocessed: normalized, segmented, and standardized.	VBM8 toolbox	Joint and meta models had similar classification performance, which was better than performance of local models.
	Target: classification of patients with first-episode schizophrenia	Create four local SVM models. Then create multisample models (joint model and meta model) based on the individual models created previously. This process was repeated four times, by setting each time three training datasets, with remaining one as the validation set.	MATLAB statistics and machine learning toolbox	

(Continued on following page)

TABLE 2. Summary of Methods and Results of Distributed Machine Learning Studies Grouping More Than One Health Care Center (Continued)

Reference	Data and Target	Methods and Distributed Learning Approach	Tools	Accomplishments and Results
Jochems[63]	Clinical data from 698 patients with lung cancer, treated with curative intent with CRT or RT alone were collected and stored in two medical institutes: MAASTRO (Netherlands) and Michigan University (United States).	Distributed learning for a Bayesian network using data from three hospitals	Varian learning portal	AUC, 0.662
	Target: prediction of NSCLC 2-year survival after radiation therapy	The model used the T category and N category, age, total tumor dose, and WHO performance for predictions.		The discriminative performance of centralized and distributed models on the validation set was similar.
Brisimi[65]	Electronic health records from Boston Medical Center of patients with at least one heart-related diagnosis between 2005 and 2010.	Soft-margin l1-regularized sparse SVM classifier.	Not provided	AUC, 0.56
	The data are distributed between 10 hospitals.			Developed an iterative cPDS algorithm for solving the large-scale SVM problem in a decentralized fashion. The system then predicted patient's hospitalization for cardiac events in upcoming calendar year.
	Target: prediction of heart cardiac events.			cPDS converged faster than centralized methods.
Tagliaferri[64]	227 variables extracted from thyroid cancer data from six Italian cancer centers. Each has four properties: name, form, type of field, and levels.	Inferential regression analysis.	COBRA framework	Thyroid COBRA: based on COBRA-Storage System. A new software BOA "Beyond Ontology" supporting two different models: Cloud-based large database model and distributed learning model
	Target: prediction of survival and toxicity.	Learning Analyzer Proxy (module of BOA only in distributed mode) sends algorithms directly to local research proxies, taking back from them only the results of each iteration step, with no need to work with shared data in the Cloud anymore.		

Abbreviations: AUC, area under the curve; BOA, Beyond Ontology Awareness; COBRA, Consortium for Brachytherapy Data Analysis; cPDS, cluster Primal Dual Splitting; CRT, chemoradiation; NSCLC, non–small-cell lung cancer; RT, radiotherapy; SVM, support vector machine.

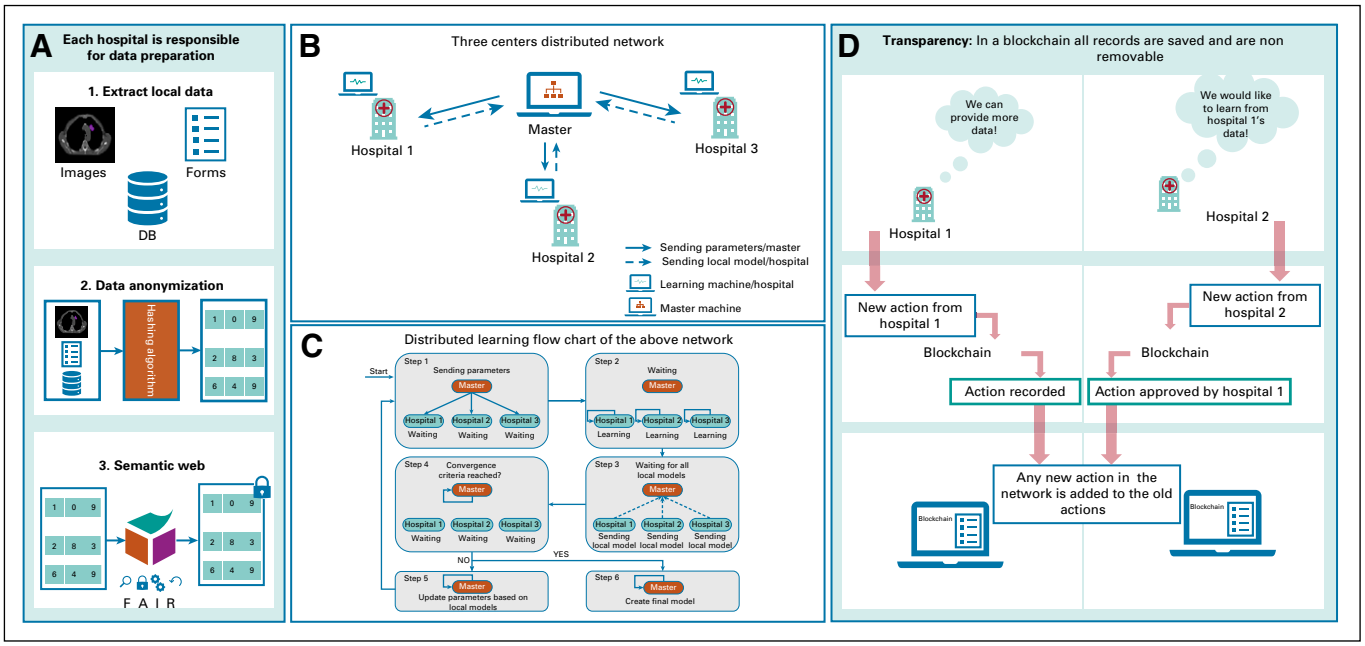

FIG 2. Schematic representation of the processes in a transparent distributed learning network. (A) Data preparation steps. (B) Distributed learning network, which is composed of three hospitals, each of which is equipped with a learning machine that can communicate with a master machine responsible for sending model parameters and checking convergence criteria. (C) Flowchart of the distributed learning network described in B. (D) Example of an action that can be tracked by blockchain (designed and implemented according to needs agreed among network members) and keep all network participants aware of any new activity taken in the network. DB, database; FAIR, findable, accessible, interoperable, reusable.

thickness estimation, and so on) of human brain magnetic resonance images.[35] The results demonstrated performance improvement on the test datasets. Similar to the previous study, a brain tumor segmentation was successfully performed using distributed deep learning across 10 institutions (BraTS distribution).[36]

In the matter of distributed deep learning, the training weights are combined to train a final model, and the raw data are never exposed.[35,37] In the case of sharing the locale gradients,[25] it might be possible to retrieve estimations of the original data from these gradients. Training the local models on batches may prevent retrieving all the data from the gradients, as these gradients correspond to single batches rather than all the local data.[38] However, setting an optimal batch size needs to be considered[25] to assure data safety and the model's ability to generalize.[28,39,40]

PRIVACY AND INTEGRATION OF DISTRIBUTED LEARNING NETWORKS

Privacy in a distributed learning network addresses three main areas: data privacy, the implemented model's privacy, and the model's output privacy. Data privacy is achieved by means of data anonymization and data never leaving the medical institutions. The distributed learning model can be secured by applying differential privacy techniques,[41] preventing leakage of weights during the training, and cryptographic techniques.[42] These cryptographic techniques provide a set of multiparty protocols that ensure security of the computations and communication. Once the model is ready, not only can the network participants use it to learn from their data, but this learning should be able to be performed locally and under highly private and secure conditions to protect the model's output.[23]

The users of a machine/deep learning model are not necessarily the model's developers. Hence, documentation and the integration of automated data eligibility tests have two important assets:

- The documentation ensures providing a clear view of what the model is designed for, a technical description of the model, and its use.
- The eligibility tests are important to ensure that correct input data are extracted and provided before executing the model. In euroCAT,[23] a distributed learning expert installed quality control via data extraction pipelines at every participant point in the network. The pipeline automatically allowed data records fulfilling the model training eligibility criteria to be used in the training. The experts also test the extraction pipeline thoroughly in addition to the machine learning testing. However, there were post-processing compensation methods to correct for the variations caused by using different local protocols.[19]

DISCUSSION

If one examines oncology, for instance, cancer is clearly one of the greatest challenges facing health care. More than 16 million new cancer cases were reported in 2017 alone.[43] This number climbed to 18.1 million cases in 2018.[44] This

increasing number of cancer incidences[45] means that there are undoubtedly sufficient data worldwide to put machine/deep learning to meaningful work. However, as highlighted earlier, this requires access to the data and, as also highlighted earlier, distributed learning enables this in a manner that resolves legal and ethical concerns. Nonetheless, integration of distributed learning into health care is much slower compared with other fields, which raises the question of why this should be. Here, we summarize a set of methodologies to facilitate the adoption of distributed learning and provide future directions.

CURRENT STATE OF MEDICAL DATA STORAGE AND PREPROCESSING

Information Communication Technology

Every hospital has its own storage devices and architecture.[38,39] In this case, the information communication technology preparation for distributed learning requires significant energy, time, and manpower, which can be costly. This same process (data acquisition and preprocessing) needs to be repeated for each participating hospital,[46-48] and subsequently development and adoption of medical data standardization protocols need to be developed for this implementation process.

Make the Data Readable: Findable, Accessible, Interoperable, Reusable Data Principles

One way to enable a virtuous circle network effect is to embrace another community engaged in synergistic activities (joining a distributed learning network is worthwhile if it links to another large network). The Findable, Accessible, Interoperable, Reusable (FAIR) Guiding Principles for data management and stewardship have gained substantial interest, but delivering scientific protocols and workflows that are aligned with these principles is significant.[49] A description of FAIR principles is represented in Figure 3. Technological solutions are urgently needed that will enable researchers to explore, consume, and produce FAIR data in a reliable and efficient manner,

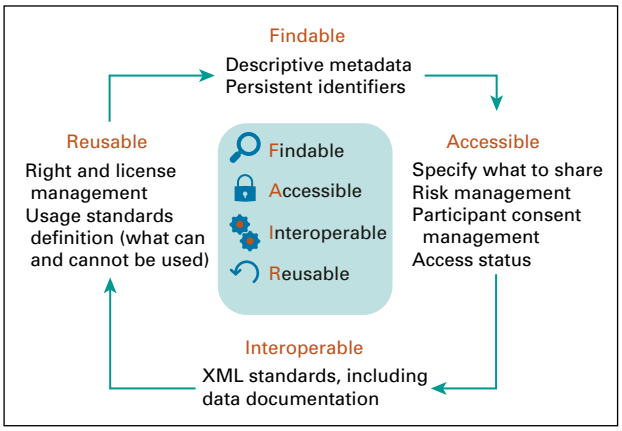

FIG 3. Description of findable, accessible, interoperable, reusable (FAIR) principles.

to publish and reuse computational workflows, and to define and share scientific protocols as workflow templates.[50] Such solutions will address emerging concerns about the nonreproducibility of scientific research, particularly in data science (eg, poorly published data, incomplete workflow descriptions, limited ability to perform meta-analyses, and an overall lack of reproducibility).[51,52] Because workflows are fundamental to research activities, FAIR has broad applicability, which is vital in the context of distributed learning with medical data.

WHY NOT PUBLICLY SHARE MEDICAL DATA?

Some studies were conducted trying to facilitate and secure data-sharing procedures to encourage related researchers and organizations to publicly share their data and embrace transparency,[53] by proposing data-sharing procedures and protocols aiming to harmonize regulatory frameworks and research governance.[54,55] Despite the efforts made toward data-sharing globalization, the sociocultural issues surrounding data sharing remain pertinent.[56] Large clinical trials also face limitations in the data collection capabilities because of limited data storage capacities and manpower. To retrospectively perform additional analysis, all the participating centers need to be contacted again, which is time consuming and delays research.[57]

Furthermore, medical institutions prefer not to share patient data to ensure privacy protection.[58] This is, of course, in no small part about ensuring the trust and confidence of patients who display a wide range of sensitivities toward the use of their personal data.

ORGANIZATIONAL CHANGE MANAGEMENT

The adoption of distributed learning will require a change in organizational management (such as making use of newest data standardization techniques and adapting the roles of employees to more technically oriented tasks, such as data retrieval). Provided knowledge and understanding of proper change management concepts, health care providers can implement the latter successfully.[59] Change management principles, such as defining a global vision, networking, and continuous communicating, could facilitate the integration of new technologies and bring up the clinical capabilities. However, this process of change management can be complicated, because it requires the involvement of multiple health care centers from different countries and continents. This diversity can trigger a fear of loss (one of the major factors of financial decision making), which stems from differences of opinion and regulation,[60] and the absence of data standardization, making the processes of data acquisition and preprocessing harder. In addition, the lack of knowledge about the new technology leads to resistance to accept the change and innovation.[60,61] Therefore, it is important to help health care organizations understand the need for distributed learning by explaining the context of the change in terms of traditional ways of learning to distributed learning and

 © 2020 by American Society of Clinical Oncology

a long-term vision of the improvements that it can bring, including time and money savings for both hospitals and patients. This could in turn improve patient lives, in addition to conducting more studies on research databases to consolidate proof of safety and quality of distributed models.

As can be seen in Table 2, distributed learning has been applied to train different models that can predict different outcomes for a variety of pathologies, including lung cancer,[23,62,63,63a] thyroid cancer,[64] heart cardiac events,[65] and schizophrenia,[66] in addition to the continuous development of tools and algorithms facilitating the adoption of distributed learning, such as the variant learning portal, the alternating direction method of multipliers algorithm,[2] as well as the application of FAIR data principles. The cited studies provide a proof that distributed learning can ensure patient data privacy and guarantee that accurate models are built that are the equivalent of centralized models.

LIMITATIONS OF THE EXISTING DISTRIBUTED LEARNING IMPLEMENTATIONS

A shared limitation of the studies presented in Table 2 is that the number of institutes involved in the distributed network is rather small. The size of the network varies from four to 10 institutions. With few medical institutes involved, the models were trained using the data of only a few hundred patients. By promoting the use of distributed learning, it should instead be possible to train the models using data from thousands or even millions of patients.

FUTURE PERSPECTIVES

An automated monitoring system accessible by the partners or medical centers participating in the distributed learning network can promote transparency, traceability, and trust.[67] Recent advances of information technology, such as blockchain, can be integrated into a distributed learning network.[68] Blockchain allows trusted partners to visualize the history of the transactions and actions taken in the distributed network. This integration of blockchain should help in easing the resistance to the new distributed technology among health care workers as it provides both provenance and enforceable governance.

In 2008, Satoshi Nakamoto[69] introduced the concept of a peer-to-peer electronic cash system known as Bitcoin. Blockchain was made famous as the public transaction ledger of this cryptocurrency.[69,70] It ensures security by using cryptography in a decentralized, immutable distributed ledger technology.[71] It is easy to manage as it can be made public, whereby any individual can participate, or it can be made private, where all participants are known to each other.[72] It is an efficient monitoring system, as records cannot be deleted from the chain. By these means, blockchain exceeds its application as a cryptocurrency to a permanent trustful tracing system. Figure 4 illustrates a visual representation of blockchain.

Boulos et al[71] demonstrated how blockchain could be used to contribute in health care: securing patient information and provider identities, managing health supply chains,

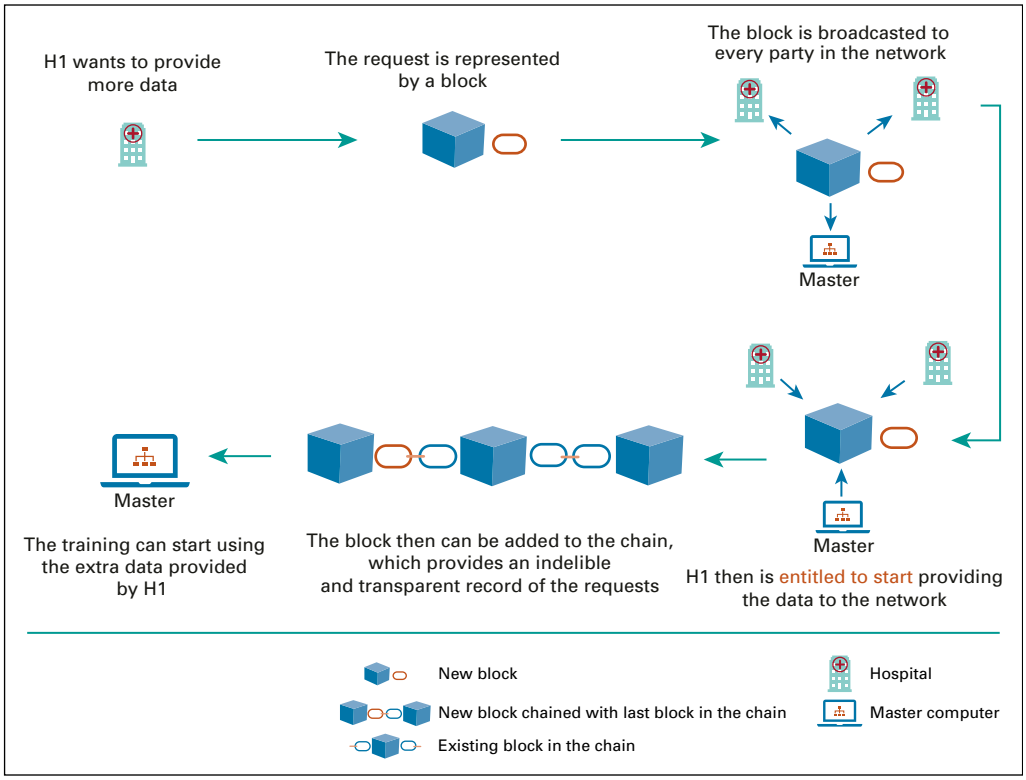

FIG 4. Visual representation of blockchain. Adapted from Rennock et al.[18]

monetizing clinical research and data (giving patients the choice to share), processing claims, detecting fraud, and managing prescriptions (replace incorrect and outdated data). In addition to the above-mentioned uses of blockchain, it has been also used to maintain security and scalability of clinical data sharing,[73] secure medical record sharing,[74] prevent drug counterfeiting,[75] and secure a patient's location.[76]

It is essential that the use of distributed machine/deep learning and blockchain be harmonized with the available security-preserving technologies (ie, continues development and cybersecurity), which begins at the user levels (use strong passwords, connect using only trusted networks, and so on) and ends with more complex information technology infrastructures (such as data anonymization and user ID encryption).[77] Cybersecurity is a key aspect in preserving privacy and ensuring safety and trust among patients and health care systems.[78] The continuous development or postmarketing surveillance can be seen as the set of checks and integrations that should occur when a distributed learning network is launched. This practice should make it possible to identify any weak security measures in the network or non-up-to-date features that may require re-implementation.[79,80]

The distributed learning and blockchain technologies presented here show that there are emerging data science solutions that begin to meet the concerns and shortcomings of the law. The problems of re-identification are greatly reduced and managed through the technologies. Clearly, there are conceptual issues of understanding the impact of these technologies on privacy, and the relationship between privacy and confidentiality, but there are significant technical developments for the regulators to consider that could answer a number of their concerns.

SUMMARY

Currently, a combination of regulations and ethics makes it difficult to share data even for scientific research purposes. The issues relate to the legal basis for processing and anonymization. Specifically, there has been reluctance to move away from informed consent as the legal basis for processing toward processing in the public interest, and there are concerns about the re-identification of individuals where data are de-identified and then shared in aggregated environments. A solution could be to allow researchers to train their machine learning programs without the data ever having to leave the clinics, which in this paper we have established as distributed learning. This safe practice makes it possible to learn from medical data and can be applied across various medical disciplines. A limitation to its application, however, is that medical centers need to be convinced to participate in such practice, and regulators also need to know suitable safeguards have been established. Moreover, as can be seen in Table 2, even with the use of distributed learning, the size of the data pool learned from remains rather small. In the future, the integration of blockchain technology to distributed learning networks could be considered, as it ensures transparency and traceability while following FAIR data principles and can facilitate the implementation of distributed learning.

AFFILIATIONS

[1]The D-Lab, Department of Precision Medicine, GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands
[2]Oncoradiomics, Liège, Belgium
[3]Department of Radiation Oncology, University Hospital Zurich and University of Zurich, Zurich, Switzerland
[4]Department of Health, Ethics, and Society, CAPHRI (Care and Public Health Research Institute), Maastricht University, Maastricht, The Netherlands

CORRESPONDING AUTHOR

Fadila Zerka, PhD, Universiteit Maastricht, Postbus 616, Maastricht 6200 MD, the Netherlands; e-mail: f.zerka@maastrichtuniversity.nl.

SUPPORT

Supported by European Research Council advanced grant ERC-ADG-2015 Grant No. 694812, Hypoximmuno; the Dutch technology Foundation Stichting Technische Wetenschappen Grant No. P14-19 Radiomics STRaTegy, which is the applied science division of Dutch Research Council (De Nederlandse Organisatie voor Wetenschappelijk) the Technology Program of the Ministry of Economic Affairs; Small and Medium-Sized Enterprises Phase 2 RAIL Grant No. 673780; EUROSTARS, DART Grant No. E10116 and DECIDE Grant No. E11541; the European Program PREDICT ITN Grant No. 766276; Third Joint Transnational Call 2016 JTC2016 "CLEARLY" Grant No. UM 2017-8295; Interreg V-A Euregio Meuse-Rhine "Euradiomics" Grant No. EMR4; and the Scientific Exchange from Swiss National Science Foundation Grant No. IZSEZ0_180524.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Financial support: Sean Walsh
Administrative support: Sean Walsh
Provision of study material or patients: Sean Walsh
Collection and assembly of data: Fadila Zerka, Samir Barakat, Ralph T.H. Leijenaar
Data analysis and interpretation: Fadila Zerka, Samir Barakat, Ralph T.H. Leijenaar, David Townend, Philippe Lambin
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

 © 2020 by American Society of Clinical Oncology

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Fadila Zerka
Employment: Oncoradiomics
Research Funding: PREDICT

Samir Barakat
Employment: PtTheragnostic
Leadership: PtTheragnostic

Sean Walsh
Employment: Oncoradiomics
Leadership: Oncoradiomics
Stock and Other Ownership Interests: Oncoradiomics
Research Funding: Varian Medical Systems (Inst)

Ralph T. H. Leijenaar
Employment: Oncoradiomics
Leadership: Oncoradiomics
Stock and Other Ownership Interests: Oncoradiomics
Patents, Royalties, Other Intellectual Property: Image analysis method supporting illness development prediction for a neoplasm in a human or animal body (PCT/NL2014/050728)

Arthur Jochems
Stock and Other Ownership Interests: Oncoradiomics, Medical Cloud Company

Benjamin Miraglio
Employment: OncoRadiomics

Philippe Lambin
Employment: Convert Pharmaceuticals
Leadership: DNAmito
Stock and Other Ownership Interests: BHV, Oncoradiomics, Convert Pharmaceuticals, The Medical Cloud Company
Honoraria: Varian Medical
Consulting or Advisory Role: BHV, Oncoradiomics
Research Funding: ptTheragnostic
Patents, Royalties, Other Intellectual Property: Co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issued patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three nonpatentable inventions (software) licensed to ptTheragnostic/DNAmito, Oncoradiomics, and Health Innovation Ventures.
Travel, Accommodations, Expenses: ptTheragnostic, Elekta, Varian Medical

David Townend
Consulting or Advisory Role: Newron Pharmaceuticals (I)

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Simone Moorman for editing the manuscript.

REFERENCES

1. Mitchell TM: Machine Learning International ed., [Reprint.]. New York, NY, McGraw-Hill, 1997
2. Boyd S, Parikh N, Chu E, et al: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3:1-122, 2010
3. Cardoso I, Almeida E, Allende-Cid H, et al: Analysis of machine learning algorithms for diagnosis of diffuse lung diseases. Methods Inf Med 57:272-279, 2018
4. Wang X, Peng Y, Lu L, et al: ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Presented at 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, July 21-26, 2017
5. Ding Y, Sohn JH, Kawczynski MG, et al: A deep learning model to predict a diagnosis of Alzheimer disease by using ^{18}F-FDG PET of the brain. Radiology 290:456-464, 2019
6. Emmert-Streib F, Dehmer M: A machine learning perspective on personalized medicine: An automized, comprehensive knowledge base with ontology for pattern recognition. Mach Learn Knowl Extr 1:149-156, 2018
7. Deist TM, Dankers FJWM, Valdes G, et al: Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Med Phys 45:3449-3459, 2018
8. Lambin P, van Stiphout RG, Starmans MH, et al: Predicting outcomes in radiation oncology multifactorial decision support systems. Nat Rev Clin Oncol 10:27-40, 2013
9. Wang S, Summers RM: Machine learning and radiology. Med Image Anal 16:933-951, 2012
10. James G, Witten D, Hastie T, et al: An introduction to statistical learning: With applications in R. New York, NY, Springer, 2017
11. Sutton RS, Barto AG: Reinforcement Learning: An Introduction. https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf
12. Deng L: Deep learning: Methods and applications. Foundations and Trends in Signal Processing 7:197-387, 2014
13. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521:436-444, 2015
14. Garling C: Andrew Ng: Why 'deep learning' is a mandate for humans, not just machines. Wired 2015. https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/
15. Pesapane F, Codari M, Sardanelli F: Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2:35, 2018
16. Liberati A, Altman DG, Tetzlaff J, et al: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Med 6:e1000100, 2009
16a. Intersoft Consulting: General Data Protection Regulation: Recitals. https://gdpr-info.eu/recitals/no-26/
17. MAASTRO Clinic: euroCAT: Distributed learning. https://youtu.be/nQpqMIuHyOk
18. Rennock MJW, Cohn A, Butcher JR: Blockchain technology and regulatory investigations. https://www.steptoe.com/images/content/1/7/v2/171967/LIT-FebMar18-Feature-Blockchain.pdf
19. Orlhac F, Frouin F, Nioche C, et al: Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology 291:53-59, 2019
20. Goodfellow I, Bengio Y, Courville A: Deep Learning. https://www.deeplearningbook.org/
21. Lambin P, Roelofs E, Reymen B, et al: Rapid Learning health care in oncology - an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol 109:159-164, 2013
22. Lustberg T, van Soest J, Jochems A, et al: Big Data in radiation therapy: Challenges and opportunities. Br J Radiol 90:20160689, 2017

23. Deist TM, Jochems A, van Soest J, et al: Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clin Transl Radiat Oncol 4:24-31, 2017
24. Price G, van Herk M, Faivre-Finn C: Data mining in oncology: The ukCAT project and the practicalities of working with routine patient data. Clin Oncol (R Coll Radiol) 29:814-817, 2017
25. Dean J, Corrado G, Monga R, et al: Large Scale Distributed Deep Networks. Advances in Neural Information Processing Systems 25, 2012, 1223-1231. https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012
26. Cireşan D, Meier U, Schmidhuber J: Multi-column deep neural networks for image classification. http://arxiv.org/abs/1202.2745
27. Radiuk PM: Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. Information Technology and Management Science 20:20-24, 2017
28. Keskar NS, Mudigere D, Nocedal J, et al: On large-batch training for deep learning: generalization gap and sharp minima. http://arxiv.org/abs/1609.04836
29. Papernot N, Abadi M, Erlingsson Ú, et al: Semi-supervised knowledge transfer for deep learning from private training data. http://arxiv.org/abs/1610.05755
30. Shokri R, Shmatikov V: Privacy-preserving deep learning, in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15. Denver, Colorado, ACM Press, 2015, pp 1310-1321.
31. Predd JB, Kulkarni SB, Poor HV: Distributed learning in wireless sensor networks. IEEE Signal Process Mag 23:56-69, 2006
32. Ji X, Hou C, Hou Y, et al: A distributed learning method for ℓ 1 -regularized kernel machine over wireless sensor networks. Sensors (Basel) 16:1021, 2016
33. Chang K, Balachandar N, Lam C, et al: Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 25:945-954, 2018
34. McClure P, Zheng CY, Kaczmarzyk J, et al: Distributed Weight Consolidation: A Brain Segmentation Case Study. https://arxiv.org/abs/1805.10863
35. FreeSurferWiki: FreeSurfer. http://freesurfer.net/fswiki/FreeSurferWiki
36. Sheller MJ, Reina GA, Edwards B, et al: Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. http://arxiv.org/abs/1810.04304
37. Li W, Milletarì F, Xu D, et al: Privacy-preserving federated brain tumour segmentation. http://arxiv.org/abs/1910.00962
38. Abadi M, Chu A, Goodfellow I, et al: Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security – CCS'16. 308-318, 2016
39. Mishkin D, Sergievskiy N, Matas J: Systematic evaluation of convolution neural network advances on the Imagenet. Comput Vis Image Underst 161:11-19, 2017
40. Lin T, Stich SU, Patel KK, et al: Don't use large mini-batches, use local SGD. http://arxiv.org/abs/1808.07217
41. Biryukov A, De Cannière C, Winkler WE, et al: Discretionary access control policies (DAC), in van Tilborg HCA, Jajodia S (eds): Encyclopedia of Cryptography and Security. Boston, MA, Springer, 2011, pp 356-358
42. Pinkas B: Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor 4:12-19, 2002
43. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2017. CA Cancer J Clin 67:7-30, 2017
44. Bray F, Ferlay J, Soerjomataram I, et al: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394-424, 2018
45. Siegel R, DeSantis C, Virgo K, et al: Cancer treatment and survivorship statistics, 2012. CA Cancer J Clin 62:220-241, 2012
46. Shortliffe EH, Barnett GO: Medical data: Their acquisition, storage, and use, in Shortliffe EH, Perreault LE (eds): Medical Informatics. New York, NY, Springer, 2001, pp 41-75
47. Shabani M, Vears D, Borry P: Raw genomic data: Storage, access, and sharing. Trends Genet 34:8-10, 2018
48. Langer SG: Challenges for data storage in medical imaging research. J Digit Imaging 24:203-207, 2011
49. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018, 2016
50. Wilkinson MD, Sansone S-A, Schultes E, et al: A design framework and exemplar metrics for FAIRness. Sci Data 5:180118, 2018
51. Dumontier M, Gray AJG, Marshall MS, et al: The health care and life sciences community profile for dataset descriptions. PeerJ 4:e2331, 2016
52. Jagodnik KM, Koplev S, Jenkins SL, et al: Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. J Biomed Inform 71:49-57, 2017
53. Polanin JR, Terzian M: A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. J Clin Epidemiol 106:60-69, 2018
54. Azzariti DR, Riggs ER, Niehaus A, et al: Points to consider for sharing variant-level information from clinical genetic testing with ClinVar. Cold Spring Harb Mol Case Stud 4:a002345, 2018
55. Boué S, Byrne M, Hayes AW, et al: Embracing transparency through data sharing. Int J Toxicol 10.1177/1091581818803880
56. Poline J-B, Breeze JL, Ghosh S, et al: Data sharing in neuroimaging research. Front Neuroinform 6:9 2012
57. Cutts FT, Enwere G, Zaman SMA, et al: Operational challenges in large clinical trials: Examples and lessons learned from the Gambia pneumococcal vaccine trial. PLoS Clin Trials 1:e16 2006
58. Xia W, Wan Z, Yin Z, et al: It's all in the timing: Calibrating temporal penalties for biomedical data sharing. J Am Med Inform Assoc 25:25-31, 2018
59. Fleishon H, Muroff LR, Patel SS: Change management for radiologists. J Am Coll Radiol 14:1229-1233, 2017
60. Delaney R, D'Agostino R: The challenges of integrating new technology into an organization. https://digitalcommons.lasalle.edu/cgi/viewcontent.cgi?article=1024&context=mathcompcapstones
61. Agboola A, Salawu R: Managing deviant behavior and resistance to change. Int J Bus Manage 6:235, 2010
62. Jochems A, Deist TM, van Soest J, et al: Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. Radiother Oncol 121:459-467, 2016
63. Jochems A, Deist TM, El Naqa I, et al: Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. Int J Radiat Oncol Biol Phys 99:344-352, 2017

63a. Deist TM, Dankers FJWM, Ojha P, et al: Distributed learning on 20 000+ lung cancer patients - The Personal Health Train. Radiother Oncol 144:189-200, 2020

64. Tagliaferri L, Gobitti C, Colloca GF, et al: A new standardized data collection system for interdisciplinary thyroid cancer management: Thyroid COBRA. Eur J Intern Med 53:73-78, 2018
65. Brisimi TS, Chen R, Mela T, et al: Federated learning of predictive models from federated Electronic Health Records. Int J Med Inform 112:59-67, 2018
66. Dluhoš P, Schwarz D, Cahn W, et al: Multi-center machine learning in imaging psychiatry: A meta-model approach. Neuroimage 155:10-24, 2017

67. Dhillon V, Metcalf D, Hooper M: Blockchain in health care, in Dhillon V, Metcalf D, Hooper M (eds): Blockchain Enabled Applications: Understand the Blockchain Ecosystem and How to Make it Work for You. Berkeley, CA, Apress, 2017, pp 125-138
68. Lugan S, Desbordes P, Tormo LXR, et al: Secure architectures implementing trusted coalitions for blockchained distributed learning (TCLearn). http://arxiv.org/abs/1906.07690
69. Nakamoto S: Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf
70. Gordon WJ, Catalini C: Blockchain technology for healthcare: Facilitating the transition to patient-driven interoperability. Comput Struct Biotechnol J 16:224-230, 2018
71. Kamel Boulos MN, Wilson JT, Clauson KA: Geospatial blockchain: Promises, challenges, and scenarios in health and healthcare. Int J Health Geogr 17:25 2018
72. Pirtle C, Ehrenfeld J: Blockchain for healthcare: The next generation of medical records? J Med Syst 42:172, 2018
73. Zhang P, White J, Schmidt DC, et al: FHIRChain: Applying blockchain to securely and scalably share clinical data. Comput Struct Biotechnol J 16:267-278, 2018
74. Dubovitskaya A, Xu Z, Ryu S, et al: Secure and trustable electronic medical records sharing using blockchain. AMIA Annu Symp Proc 2017:650-659, 2018
75. Vruddhula S: Application of on-dose identification and blockchain to prevent drug counterfeiting. Pathog Glob Health 112:161, 2018
76. Ji Y, Zhang J, Ma J, et al: BMPLS: Blockchain-based multi-level privacy-preserving location sharing scheme for telecare medical information systems. J Med Syst 42:147, 2018
77. Coventry L, Branley D: Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas 113:48-52, 2018
78. Jalali MS, Kaiser JP: Cybersecurity in hospitals: A systematic, organizational perspective. J Med Internet Res 20:e10059, 2018
79. Vlahović-Palčevski V, Mentzer D: Postmarketing surveillance, in Seyberth HW, Rane A, Schwab M (eds): Pediatric Clinical Pharmacology. Berlin, Springer, 2011, pp 339-351
80. Parkash R, Thibault B, Philippon F, et al: Canadian Registry of Implantable Electronic Device outcomes: Surveillance of high-voltage leads. Can J Cardiol 34:808-811, 2018
81. Ing EB, Ing R: The use of a nomogram to visually interpret a logistic regression prediction model for giant cell arteritis. Neuroophthalmology 42:284-286, 2018
82. Tirzīte M, Bukovskis M, Strazda G, et al: Detection of lung cancer with electronic nose and logistic regression analysis. J Breath Res 13: 016006, 2018
83. Ji Z, Jiang X, Wang S, et al: Differentially private distributed logistic regression using private and public data. BMC Med Genomics 7:S14, 2014 (suppl 1)
84. Jiang W, Li P, Wang S, et al: WebGLORE: A web service for Grid LOgistic REgression. Bioinformatics 29:3238-3240, 2013
85. Wang S, Jiang X, Wu Y, et al: EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed privacy-preserving online model learning. J Biomed Inform 46:480-496, 2013
86. Desai A, Chaudhary S: Distributed decision tree. Proceedings of the Ninth Annual ACM India Conference, Gandhinagar, India, ACM Press, 2016, pp 43-50
87. Caragea D, Silvescu A, Honavar V: Decision tree induction from distributed heterogeneous autonomous data sources, in Abraham A, Franke K, Köppen M (eds): Intelligent Systems Design and Applications. Berlin, Springer, 2003, pp 341-350
88. Plaku E, Kavraki LE: Distributed computation of the knn graph for large high-dimensional point sets. J Parallel Distrib Comput 67:346-359, 2007
89. Xiong L, Chitti S, Liu L: Mining multiple private databases using a kNN classifier, in Proceedings of the 2007 ACM symposium on Applied computing – SAC '07. Seoul, Korea, ACM Press, 2007, p 435
90. Huang Z: Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2:283-304, 1998
91. Jagannathan G, Wright RN: Privacy-preserving distributed k-means clustering over arbitrarily partitioned data, in Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining – KDD '05. Chicago, Illinois, USA, ACM Press, 2005, p 593
92. Jin R, Goswami A, Agrawal G: Fast and exact out-of-core and distributed k-means clustering. Knowl Inf Syst 10:17-40, 2006
93. Jagannathan G, Pillaipakkamnatt K, Wright RN: A new privacy-preserving distributed k -clustering algorithm, in Proceedings of the 2006 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2006, pp 494-498
94. Ye Y, Chiang C-C: A parallel apriori algorithm for frequent itemsets mining, in Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06). Seattle, WA, IEEE, 2006, pp 87-94
95. Cheung DW, Ng VT, Fu AW, et al: Efficient mining of association rules in distributed databases. IEEE Trans Knowl Data Eng 8:911-922, 1996
96. Bellman R: A Markovian decision process. Indiana Univ Math J 6:679-684, 1957
97. Puterman ML: Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York, NY, John Wiley & Sons, 2014
98. Watkins CJCH, Dayan P: Q-learning. Mach Learn 8:279-292, 1992
99. Lauer M, Riedmiller M: An algorithm for distributed reinforcement learning in cooperative multi-agent systems, in Proceedings of the Seventeenth International Conference on Machine Learning. Burlington, MA, Morgan Kaufmann, 2000, pp 535-542. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.772

APPENDIX

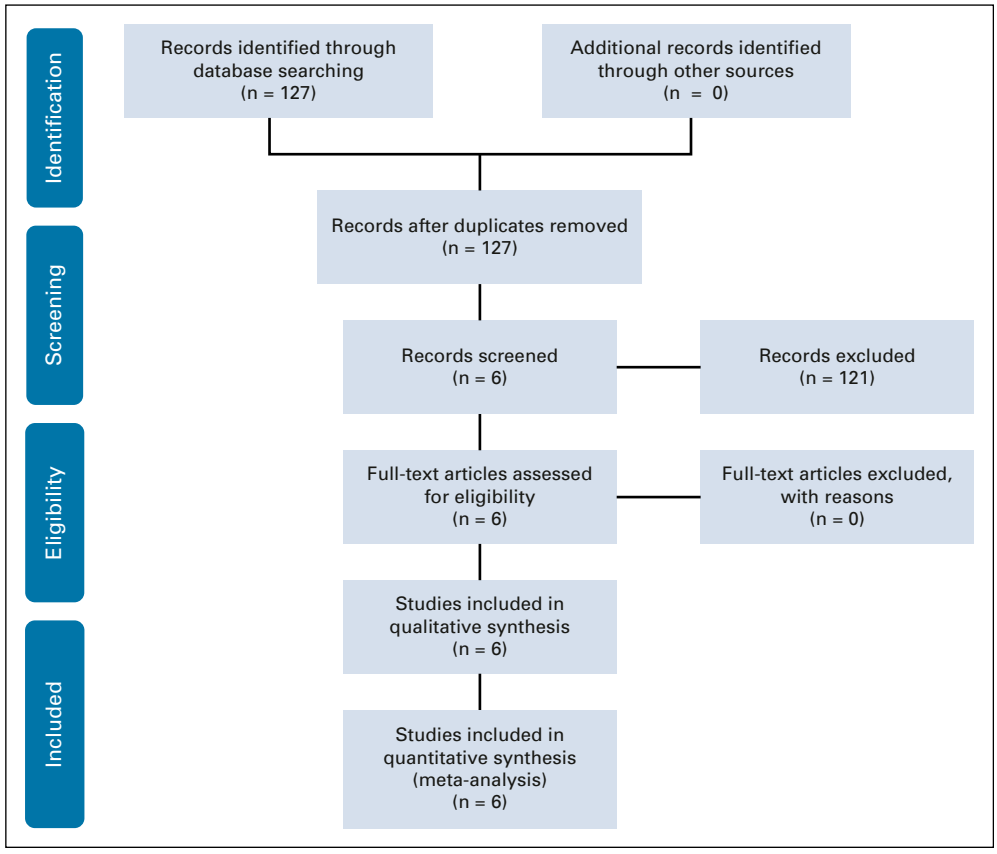

FIG A1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 flow diagram.

 © 2020 by American Society of Clinical Oncology

TABLE A2. PRISMA 2009 Checklist

Section/Topic	No.	Checklist Item	Reported on Page No.
Title			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
Abstract			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known.	1-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to PICOS.	2
Methods			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (eg, Web address), and, if available, provide registration information including registration number.	5
Eligibility criteria	6	Specify study characteristics (eg, PICOS, length of follow-up) and report characteristics (eg, years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (eg, databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	5
Study selection	9	State the process for selecting studies (ie, screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5 (and Fig A1)
Data collection process	10	Describe method of data extraction from reports (eg, piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (eg, PICOS, funding sources) and any assumptions and simplifications made.	N/A
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level) and how this information is to be used in any data synthesis.	N/A
Summary measures	13	State the principal summary measures (eg, risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (eg, I^2) for each meta-analysis.	5
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (eg, publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (eg, sensitivity or subgroup analyses, meta-regression), if done, indicating which were prespecified.	N/A
Results			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	5 (and Fig A1)
Study characteristics	18	For each study, present characteristics for which data were extracted (eg, study size, PICOS, follow-up period) and provide the citations.	5-8

(Continued on following page)

TABLE A2. PRISMA 2009 Checklist (Continued)

Section/Topic	No.	Checklist Item	Reported on Page No.
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome-level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group, and (b) effect estimates and confidence intervals, ideally with a forest plot.	5-8
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	5-8
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (eg, sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
Discussion			
Summary of evidence	24	Summarize the main findings, including the strength of evidence for each main outcome; consider their relevance to key groups (eg, health care providers, users, and policy makers).	8
Limitations	25	Discuss limitations at study and outcome level (eg, risk of bias), and at review level (eg, incomplete retrieval of identified research, reporting bias).	10
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	11
Funding			
Funding	27	Describe sources of funding for the systematic review and other support (eg, supply of data); role of funders for the systematic review.	11

Abbreviations: N/A, not applicable; PICOS, participants, interventions, comparisons, outcomes, and study design; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

 © 2020 by American Society of Clinical Oncology