Article published online: 2023-07-06

Clinical Informatics Approaches to Facilitate Cancer Data Sharing

Sanjay Aneja[1,2,3], **Arman Avesta**[1,2], **Hua Xu**[3], **Lucila Ohno Machado**[3]

[1] Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT, USA
[2] Center for Outcomes Research and Evaluation at Yale, New Haven, CT, USA
[3] Department of Bioinformatics and Data Science, Yale School of Medicine, New Haven, CT, USA

Summary

Objectives: Despite growing enthusiasm surrounding the utility of clinical informatics to improve cancer outcomes, data availability remains a persistent bottleneck to progress. Difficulty combining data with protected health information often limits our ability to aggregate larger more representative datasets for analysis. With the rise of machine learning techniques that require increasing amounts of clinical data, these barriers have magnified. Here, we review recent efforts within clinical informatics to address issues related to safely sharing cancer data.

Methods: We carried out a narrative review of clinical informatics studies related to sharing protected health data within cancer studies published from 2018-2022, with a focus on domains such as decentralized analytics, homomorphic encryption, and common data models.

Results: Clinical informatics studies that investigated cancer data sharing were identified. A particular focus of the search yielded studies on decentralized analytics, homomorphic encryption, and common data models. Decentralized analytics has been prototyped across genomic, imaging, and clinical data with the most advances in diagnostic image analysis. Homomorphic encryption was most often employed on genomic data and less on imaging and clinical data. Common data models primarily involve clinical data from the electronic health record. Although all methods have robust research, there are limited studies showing wide scale implementation.

Conclusions: Decentralized analytics, homomorphic encryption, and common data models represent promising solutions to improve cancer data sharing. Promising results thus far have been limited to smaller settings. Future studies should be focused on evaluating the scalability and efficacy of these methods across clinical settings of varying resources and expertise.

Keywords

Data sharing, federated learning, encryption, common data models, cancer

Yearb Med Inform 2023:104-10
http://dx.doi.org/10.1055/s-0043-1768721

1 Introduction

Clinical informatics has transformed cancer diagnosis, treatment, and surveillance over the past decade. Informatics efforts have allowed oncologists to better understand rare tumors by creating larger cohorts of similar patients. Additionally, with improvements in machine learning techniques, our ability to decode the complexity of cancer has dramatically improved across different data streams, spanning genomics, diagnostic imaging, and the electronic health record [1].

Despite increasing enthusiasm regarding the role of informatics and machine learning in improving our understanding of cancer, the availability of clinical data remains a persistent bottleneck to progress. This problem has become more apparent because modern machine learning techniques often rely on larger amounts of high-dimensional clinical data. Restrictions on sharing protected health information have made it increasingly difficult to aggregate larger amounts of cancer data for study. Moreover, when data sharing agreements are put in place, heterogeneity across institutions makes the process of obtaining usable harmonized data for analysis a labor-intensive process. The process of data harmonization itself serves as an additional barrier, as institutions must devote resources to either adjusting models to account for differences in data or harmonizing data before collaboration.

In response to these challenges, the cancer informatics community has attempted to build solutions to ease the burden of sharing patient-usable data safely between institutions. Here, we present a review of current difficulties in sharing clinical cancer data and key literature on informatics solutions that have shown success. Specifically, we focus on decentralized machine learning approaches, homomorphic encryption technology, and common data models.

2 Methods

For this narrative review, we performed a search of MEDLINE with a focus on prominent clinical informatics journals, including *JCO Clinical Cancer Informatics*, the *Journal of the American Informatics Association*, *Applied Clinical Informatics*, *BMJ Health & Care Informatics, Informatics for Health and Social Care, International Journal of Medical Informatics, Methods of Information in Medicine,* and the IMIA Yearbook of Medical Informatics. We reviewed articles published from 2018 to 2022 that were relevant to our discussion, with an emphasis on articles from the past two years. We limited our review to studies in three domains: 1) decentralized analytics/machine learning, 2) homomorphic encryption methods, and 3) common data models. Although these domains do not cover all methods for data sharing, we chose them because of their application in cancer informatics and their usefulness across different types of data seen in oncology research. Finally, we limited our review to studies that examined cancer across genomic, imaging, and clinical data streams.

3 Barriers to Sharing Patient Data

In contrast to sharing non-clinical data streams, the sharing of patient-linked healthcare data poses considerably higher risk. Both the United States Health Insurance Portability and Accountability Act (HIPAA) and the European General Data Protection Regulation (GDPR) impose strict rules regarding the exchange of personally identifiable healthcare data for non-care purposes, including research. Due to such restrictions, multi-institutional efforts to aggregate large patient cohorts are often burdensome, time-consuming, and require significant resources [2, 3]. Successful examples of large collaborative datasets include the AACR-sponsored Project GENIE, which provides genomic data with limited outcomes data across 18 institutions representing 160,000 cancer patients in 6 countries [4]. Given such difficulties, large data collaborations are often limited to a small number of like-minded institutions with adequate resources and enthusiasm. Although such small efforts have value, they risk potential bias because they lack data representing the entire spectrum of cancer patients. This is of particular importance given that institutions and health systems without resources to participate in such studies may represent different patient populations. Kaushal *et al.* [5] found that a minority of clinical deep learning research studies (24%) analyzed multiple institutional data. Moreover, they found that many of the studies used data from only three states within the US (California, Massachusetts, and New York), with a paucity of studies from the other 47 states.

Patients have expressed concerns regarding the privacy of their healthcare data. Khullar *et al*. [6] conducted a survey of patient perspectives on artificial intelligence within healthcare found that over 70% of surveyed patients expressed concerns about privacy breaches associated with healthcare data. This proportion was higher among non-white (74% vs. 68%) and older respondents (72% vs. 69%).

Another barrier to sharing patient data is a lack of interoperability between institutional datasets. Despite increasing digitization of healthcare data, there is a lack of standardization across many clinical data streams, most notably the clinical data within the electronic health record [7]. Although the Digital Imaging and Communications in Medicine (DICOM) provide a data standard in diagnostic imaging, differences in imaging parameters, equipment, and protocols among different institutions make models trained on multi-institutional datasets less likely to generalize across different patient populations and regions [7]. Additionally, although data standards such as FHIR/HL7 have the potential to increase interoperability, adoption has been lacking [8].

4 Anonymization

Anonymization remains the most common way of sharing patient level data that has shown some success. Removing identifiable features from a dataset mitigates the risks associated with sharing data across institutions. A main theoretical advantage of data anonymization is that it allows aggregated data to be shared within the public domain for further research studies. Successful examples of de-identified public cancer data repositories include the Cancer Genome Atlas (TCGA) [9] and Cancer Imaging Archive (TCIA) [10] which have both required considerable investment by the US National Institute of Health (NIH). The TCGA tiers its data into different layers of sensitivity, with only the most general layer being openly available to users. However, anonymization still presents challenges that make data sharing difficult. One of the challenges is that true anonymization is difficult to achieve, and there is always a risk that patient data can be re-identified despite best efforts to remove information. Second, effectively anonymizing various data streams requires significant resources, expertise, and time which can be difficult for resource limited institutions [11]. Third, anonymization may remove important information needed for analysis [12]. Lastly, there are differing views of what constitutes truly anonymized data, often placing different institutional policies at odds with one another when trying to share data [13, 14].

5 Decentralized Analytics

One emerging solution to help mitigate risks and resources required to anonymize data for shared informatics projects is decentralized analysis. Decentralized analysis allows institutions to keep healthcare data locally but conduct informatics analysis collaboratively. These solutions have become popular in the setting of machine learning techniques which have large data requirements. Typical decentralized machine learning techniques involve sharing some parameters of a model during the training process without sharing actual protected health information.

Chang *et al.* [15] demonstrated the efficacy of cyclical weight transfers to train deep learning models to classify breast mammograms. Cyclical weight transfer involves multiple institutions transferring parameters of a model during the training process. Because the model is training on data from both institutions intermittently, there is less likelihood of overfitting to one institutions dataset. The authors found that cyclical weight transfer created a model with similar accuracy to a model created with all data aggregated centrally. The method was scalable to 20 hypothetical institutions and maintained relatively strong performance when one institution had lower quality data or imbalanced datasets. One potential disadvantage of this method is that performance is strongest when all institutions are training at the same time. Also, the cycle is dependent on the computational speed of each institution meaning training may be limited by the resources of the slowest institution.

Deist *et al.* [16] demonstrated one of the largest successful algorithms trained in a decentralized manner. The authors trained a logistic regression model to predict 2-year survival based on TNM staging on over 20,000 non-small cell lung cancer patients across five countries. The model which used TMN stage as the inputs was trained in a distributed manner using the Alternative Direction Method of Multipliers technique. It is unclear whether this method is applicable to more complex data streams with larger numbers of parameters, but nevertheless demonstrates the potential of decentralized machine learning techniques to aggregate large international cohorts of patients.

Federated machine learning techniques were initially developed in 2015 and represent a different method for decentralized machine learning [17]. Federated learning algorithms rely on distributing copies of a machine learning algorithm to institutions which house their own protected data [18]. Training iterations are completed locally and return results to a central repository for aggregation. The central repository then provides a new global model to re-distribute to devices for further training. The major benefits of federated learning compared to alternative decentralized machine learning techniques is the flexibility to operate when some devices are off-line. The disadvantage of federated learning is that performance potentially decreases if the incorrect aggregation strategy is chosen when configuring the global model.

Federated learning has been shown to be effective for cancer image analysis. Recently the German Cancer Consortium released a Joint Imaging Platform for federated clinical image analysis [19]. The developers successfully created a federated learning platform which was implemented across 10 institutions within Germany. The platform is currently being implemented to house data for six multi-center clinical trials investigating a variety of different cancer types. The platform highlights the potential of federated learning to improve the efficiency of large clinical trials in oncology which are frequently multi-institutional in nature.

Sarma *et al.* [20] demonstrated the utility of federated learning for biomedical image segmentation. The authors trained a deep learning algorithm which successfully segmented prostates on MRI across three institutions. The federated learning algorithm was more generalizable and accurate when compared to models trained locally (Dice 0.812 vs. 0.889, p<.001).

Federated has also been applied to histopathological image analysis. Agbley *et al.* [21] demonstrated the ability to identify invasive carcinoma on breast pathology specimens across three separate institutions. The authors noted that the federated learning underperformed when there was significant class imbalance between institutions. This suggests that federated learning systems may not perform well when combining very disparate patient populations.

Lu *et al.* [22] similarly demonstrated federated learning to be an effective method for classification of histopathological data. The authors were able to train image classification models to predict survival from whole slide images across four different institutions. Although the authors found federated models did show strong performance, the models did underperform when compared to centrally trained models. It is unclear from the study if the performance gap between central and federated models would be mitigated if they had a larger number of institutions.

Although federated approaches have most often been focused on supervised machine learning tasks, it has been shown to also be applicable to unsupervised learning algorithms. Bercea *et al.* [23] developed a framework to train an unsupervised autoencoder to identify high grade gliomas on brain MRIs across four institutions. The authors found their federated approach improved glioma identification by 80% compared to locally trained models. Specifically, the authors noted that the dramatic increase in performance was because each individual institution did not possess enough cases to adequately train their model. Only through a federated approach across multiple institutions were the authors able to have a sample size large enough to successfully complete their task.

Federated learning using clinical data has been shown to be effective in small settings. Rajendran *et al.* [24] demonstrated both neural network and logistic regression models predicting risk of developing lung cancer from electronic health data could successfully be trained in a cloud based federated environment. Notably the authors found that the logistic regression model did not show significant improvements in performance when trained using larger federated data source. In contrast the neural network showed significant improvements in performance when trained on a larger dataset via federated learning.

Hansen *et al.* [25] similarly used clinical data across three countries to build a cox regression model to identify factors associated with larynx cancer outcomes. The federated model showed strong discriminatory ability with AUCs ranging from 0.67 to 0.77 but did not significantly outperform non-federated localized cox models. The authors do note that the federated model shows slightly better separation of risk groups compared to localized models. These findings further highlight that certain data intensive machine learning methods may benefit the most from federated environments.

Federated learning does have known challenges. Scalability of federated learning infrastructure on cancer problems remains understudied. A systematic review of studies using federated databases completed by Zerka *et al.* [26] found that published studies on federated learning involved less than 10 institutions and often only analyzed a few hundred patients.

One of the largest demonstrations of federated learning at scale was published by Pati *et al.* [27] With a group of 71 institutions across six continents the authors trained a successful auto-segmentation model for glioblastomas on brain MRIs. As expected, the authors found increased data improved overall model performance and made federated models more robust to potential data quality issues at individual institutions.

Although federated learning does allow individual devices to retain protected data, they may still be sensitive to privacy threats. These threats include attempted extraction of training data information from intermediate/final models and corrupting models to produce inaccurate results [28-30]. Evidence suggests that nefarious actors maybe able to reconstruct individual data located on devices if given access to model parameters while training a federated model [31].

6 Homomorphic Encryption

Encryption is alternative technique to facility patient data sharing which is thought to be less sensitive to privacy threats than decentralized analytics. Based on the fundamentals of number theory, encryption techniques transform original data into an encoded format [12]. Among the most popular encryption techniques within healthcare is homomorphic encryption. Homomorphic encryption is a specific type of encryption which enables primitive mathematical operations (for example addition, multiplication) directly on encoded data. The advantage of homomorphic encryption techniques is that they offer more certain privacy of data. The disadvantage is the significant computational resources to successfully encrypt medical data. There appears to be an efficiency security trade-off where the most efficient scalable encryption methods are likely less secure [18].

Improving the efficiency of homomorphic encryption methods remains an area of continued research. In 2018, the iDASH Privacy and Security Workshop organized a special competition track to create secure parallel genome wide association studies using homomorphic encryption [32]. The winning homomorphic encryption solutions from Duality Technologies [33] and UCSD [32] successfully completed full GWAS for 1,000 individuals in approximately 4 and 2 minutes respectively. Both solutions chose a similar common encryption framework. CKKS/HEAAN appears to be amenable to numerical optimization for problems that involve machine learning and statistical learning. Recently the group from Duality Technologies improved upon their iDASH winning solution resulting in improved computational efficiency and reduce computer memory usage [34]. They also demonstrated the scalability of their method on a larger dataset of 25,000 individuals.

Homomorphic encryption on images is particularly challenging given difficulty encoding visual images. Khilji *et al.* [35] successfully demonstrated the ability to train a deep learning classification model using homomorphic encryption. The model which attempted to diagnose the present of Acute Lymphoblastic Leukemia from pathologic images had an accuracy of 77.9%. Their classification model although somewhat accurate did underperform compared to non-encrypted models (77.9% vs. 80.0%).

Homomorphic encryption on clinical data has been less studied. Son *et al.* [36] demonstrated the ability homomorphic encryption to securely train model for breast cancer recurrence using clinical data from 13,000 patients. Interestingly the authors found performance of the model trained using their homomorphic encryption method performed equivalently to models trained on un-encrypted data.

Paddock *et al.* [37] demonstrated the feasibility of homomorphic encryption to identify exceptional tumor responders within a real-word dataset. The authors were able to identify all exceptional responders in 21 months over the course of the simulated study. Although the encryption and decryption process was computationally expensive (often requiring hours of computation), the authors argue the rate limiting step compares favorably to cohort identification and aggregation using traditional methods.

Combining federated learning and homomorphic encryption is arguably the safest solution to protect health information. Froelicher *et al.* [29] described a novel federated learning platform which leverages multiparty homomorphic encryption to enable privacy preservation on distributed datasets. The authors showed their platform reproduced two published centrally trained models: 1) predicting survival after the receipt of immunotherapy and 2) predicting HIV viral load from genetic data. These findings suggest that combining data sharing approaches may best ensure patient data security.

7 Common Data Models

Both decentralized analytics and encryption techniques help address privacy concerns which limit data sharing, but do not address interoperability issues that plague cancer informatics projects. Common data models represent a potential solution to help improve data sharing while also addressing interoperability issues among institutions with different informatics infrastructure. Given there exists several different common data models which are already in production across healthcare, these common data model solutions may be the most likely to impact clinical practice in the short-term.

Anonymization is a potential advantage of common data models. Common data models often are designed with privacy in mind and can easily eliminated elements with protected health information when sharing across institutions. Additionally, common data models can identify sensitive data elements and collectively build data sharing policies which allow for tiered access based on privacy risk. The use of common data models is not mutually exclusive to decentralized analytics or encryption techniques. Combining techniques may potentially provide superior methods for data sharing. Combining common data models with homomorphic encryption has been proposed by Rosario *et al.* [38] The authors demonstrated the feasibility to use homomorphic encryption and apply it to the i2b2 common data model. This allows the potential to share common data model elements in a secure fashion or more safely store them within a cloud environment.

One emerging disadvantage to common data models is difficulty harmonizing between popular common data models [39]. Some common data models employ traditional relational database design in which each table corresponds to a clinical domain (e.g., PCORnet CDM). In contrast, other common data models use alternative structures. The commonly employed i2b2 common data model employs a star-schema format which leverages one large 'fact' table to connect various concepts. Similarly, the Observational Medical Outcomes Partnership (OMOP) common data model uses a hybrid approach which blends domain tables and 'fact' tables [39]. Recognizing the difficulty of harmonizing common data models, the ONC launched a common data model harmonization initiative in collaboration with the FDA, NCI, NIH and NLM. The three year project which was completed in 2020 developed a common data architecture to facilitate interoperability between four common data models commonly used in clinical medicine (Sentinel, PCORNet, i2b2, and OMOP) [40]. Additionally, common data models are limited to certain data elements and organizational paradigms which may not align well with all clinical specialties. For example, various common data models record radiation therapy for cancer treatment differently with different degrees of detail [41]. Specifically within oncology, significant efforts have been made to attempt to align current common data models to represent cancer data in a way which is most relevant to cancer research [42, 43]. Although considerable literature has been devoted to common data models in healthcare, we will review advances in common data models specifically within the domain of oncology.

The OMOP common data model is among the most popular common data models used in medicine. Developed by Observational Health Data Science and Informatics (OHDSI) program, OMOP enables analysis of disparate observational databases through common terminologies, vocabularies, and coding schemes [44, 43]. There exist a number of extensions created by the OHDSI Oncology Subgroup and a number of independent groups have also attempted to create their own ways to extend OMOP for cancer informatics research.

Warner *et al.* [44] in collaboration with OHDSI have made OMOP more useful for cancer informatics by extending OMOP to better capture the structure of chemotherapy regimens. Leveraging a HemOnc.org a curated website of chemotherapy regimens, the authors were able to successfully map content from HemOnc.org to a relational data model that is compatible with the OMOP common data model. In addition to increasing the chemotherapy information available to OMOP users, the authors also created an extension to the OMOP CDM to handle episodes of care allowing for the capture cancer treatment information with temporal information [45].

In addition to chemotherapy information OHDSI recently released an OMOP Oncology Module [43] which extends the OMOP CDM and standardized vocabularies to better represent cancer diagnoses, treatments, and episodes. The module incorporates information from seven existing standards including the WHO International Classification of Diseases for Oncology, HemOnc.org, North American Association for Central Cancer Registries, College of American Pathologists Electronic Cancer Checklists, Nebraska Medical Clinical Ontology Application, National Cancer Institute Thesaurus, and the Anatomical Therapeutic Chemical Drug classification system. The module was successfully pilot tested at six institutions. The developers noted that the integration of the electronic health record to institutional tumor registry information was necessary to successfully fill the OMOP module.

Yu *et al.* [46] demonstrated the potential utility of common data models to collect adverse events in patients receiving immunotherapy. Using the OMOP common data model, the authors were able to identify ipilimumab induced hypopituitarism four months earlier than the FDA Adverse Event Reporting System.

The OMOP common data model has also been used to conduct epidemiological studies regarding cancer incidence. Lee *et al.* [47] used an OMOP created dataset across three hospitals in South Korea to test the association between thiazide usage and non-melanomatous skin cancer prevalence. The investigators were able to leverage OMOP common data elements to create a cohort of over 600,000 patients over the course of years. Such large-scale studies would be difficult to complete without the use of a common data model to efficiently aggregate data elements of interest.

Despite the popularity of the OMOP common data model across a variety of clinical disciplines [48, 49], one disadvantage for cancer informatics research is the lack of genetic information included in the standard OMOP tables. To address this Shin *et al.* [50] developed a genomic common data model which allows genomic information to be integrated within standard OMOP data tables. The authors showed that successful implementation of their proposed common data model allowed for successful comparison of genetic data between the Cancer Genome Atlas and Ajou University Hospital in South Korea.

Like the OMOP, PCORnet is an alternative common data model which was proposed by the Patient Center Outcomes Research Institute to facilitate large-scale patient centered research. Carnahan *et al.* [51] examined the utility of PCORnet to evaluate the utilization of molecular-guided cancer treatment and testing across nine clinical research networks and two health plan research networks. The authors found traditional billing codes to be effective at identifying molecular testing, but unable to adequately capture cancer specific details regarding the analyte being tested. In a sub-analysis, the authors found that PCORnet failed to capture all patients who received molecular guided therapy and recommended linkage of PCORnet with tumor registries to improve data capture.

One consistent challenge across proposed common data models remains the lack of specificity to clinical oncology. Most common data models lack important cancer-specific variables (staging, molecular testing, adverse events) that are important to cancer informatics researchers. To address growing concerns, the Minimal Common Oncology Data Elements (mCODE) initiative was started in 2018 [52]. The mCODE initiative which is led by the American Society of Clinical Oncology (ASCO) attempts to providing infrastructure regarding data elements which can be used across electronic health records. Some critiques of the mCODE initiative are the lack of variables capturing smoking and drinking status and the ability for it to be implemented outside of the United States [53]. Although relatively recent, mCODE has been created partly with the hope of guiding future common data models to include more cancer specific data elements [54]. There is increasing enthusiasm regarding the use of mCODE to help better capture and standardize oncology data. Specifically, CodeX the HL7 FHIR accelerator focused on interoperability is implementing and testing the use of mCODE within specific use cases [52]. Use cases include cancer registry reporting, EHR derived endpoints for cancer clinical trials, cancer clinical trial matching, prior authorization, and capturing radiation therapy treatment data for cancer patients.

An alternative to mCODE named OSIRIS was recently proposed by the French Institute National du Cancer (INCa) [55]. OSIRIS is a minimum data set framework composed of 67 clinical and 65 omic items which was validated on 300 patients across six clinical trials across different cancer types. OSIRIS is compatible with the HL7 Fast Healthcare Interoperability Resources (FHIR) format. Features of the OSIRIS common data include temporal structure to capture longitudinal cancer events, a blend of omics and clinical concepts, ability to integrate future data streams (e.g., proteomic), and the presence international terminologies.

8 Future Directions and Conclusions

Enabling data sharing, whether through use of common data models, federated learning, or other approaches is key to reaching the vision of learning from every cancer patient. There have been considerable advances to help facilitate data sharing within cancer informatics. Decentralized machine learning, encryption, and common data models represent three promising solutions which can help ease the burden of aggregating large cancer datasets. Each solution poses advantages and limitations which are being investigated in parallel. It is likely that no one informatics approach will be appropriate for all types of data and clinical settings. More likely tailored approaches will be personalized based on a specific informatics task, resources, and research expertise. Additionally, combined solutions may offer the safest and most applicable method to facilitate collaboration.

Recent regulatory changes have implications for promoting the sharing of healthcare data and advancing research. The Office of National Coordinator of Health Information Technology (ONC) finalized the 21st Century Cures Act and its companion Cures Act Final Rule on April 5th, 2021, which aim to facilitate data sharing with patients and may provide momentum for institutions to create infrastructure for safe and effective sharing of PHI with researchers [56]. In addition, the newly enacted NIH Data Management and Sharing Policy requires prospective planning for managing and sharing scientific data during the application process for NIH-funded projects, promoting greater transparency and data sharing [57]. Such initiatives are expected to encourage institutions to invest in data sharing technology and infrastructure at scale. Furthermore, the NIH Cloud Platform Interoperability effort represents a significant investment by the NIH in promoting infrastructure for democratizing data for researchers and is likely to play a larger role in promoting data sharing in the future.

Although significant progress has been made in prototyping these solutions, there are few examples of wide adoption at scale within oncology. Current progress was frequently limited to organizations with significant informatics expertise and resources. The next iteration of research in this area will likely focus on developing scalable solutions which are user-friendly for clinicians with varying levels of expertise. Given the current prototyped solutions across different data types, oncology is a field well positioned to develop such scalable solutions and become among the leading clinical fields in this arena.

No conflict of interest has been declared by the author(s).

References

1. Warner JL, Patt D. Cancer Informatics in 2019: Deep Learning Takes Center Stage. Yearb Med Inform 2020 Aug;29(01):243–6. doi: 10.1055/s-0040-1701993.
2. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR, et al. Clinical impact of COVID-19 on patients with cancer (CCC19): a cohort study. The Lancet 2020 Jun;395(10241):1907–18. doi: 10.1016/S0140-6736(20)31187-9.
3. Connor M, Paulino AC, Ermoian RP, Hartsell WF, Indelicato DJ, Perkins S, et al. Variation in Proton Craniospinal Irradiation Practice Patterns in the United States: A Pediatric Proton Consortium Registry (PPCR) Study. Int J Radiat Oncol 2022 Mar;112(4):901–12. doi: 10.1016/j.ijrobp.2021.11.016.
4. The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov 2017 Aug 1;7(8):818–31. doi: 10.1158/2159-8290.CD-17-0151.
5. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA 2020 Sep 22;324(12):1212. doi: 10.1001/jama.2020.12067.
6. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of Patients About Artificial Intelligence in Health Care. JAMA Netw Open 2022 May 4;5(5):e2210309. doi: 10.1001/jamanetworkopen.2022.10309.
7. Adnan M, Kalra S, Cresswell JC, Taylor GW, Tizhoosh HR. Federated learning and differential privacy for medical image analysis. Sci Rep 2022 Feb 4;12(1):1953. doi: 10.1038/s41598-022-05539-7.
8. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform 2022 Jul 19;10(7):e35724. doi: 10.2196/35724.
9. The Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013 Oct;45(10):1113–20. doi: 10.1038/ng.2764.
10. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 2013 Dec;26(6):1045–57. doi: 10.1007/s10278-013-9622-7.
11. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng 2022 Jul 4;6(12):1330–45. doi: 10.1038/s41551-022-00898-y.
12. Bonomi L, Huang Y, Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet 2020 Jul;52(7):646–54. doi: 10.1038/s41588-020-0651-0.
13. Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. Curr Opin Neurol 2019 Dec;32(6):850–6. doi: 10.1097/WCO.0000000000000761.
14. Thompson RF, Valdes G, Fuller CD, Carpenter CM, Morin O, Aneja S, et al. Artificial Intelligence in Radiation Oncology Imaging. Int J Radiat Oncol 2018 Nov;102(4):1159–61. doi: 10.1016/j.ijrobp.2018.05.070.
15. Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed deep learning networks among institutions for medical imaging. J Am Med Inform Assoc 2018 Aug 1;25(8):945–54. doi: 10.1093/jamia/ocy017.
16. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. Radiother Oncol 2020 Mar;144:189–200. doi: 10.1016/j.radonc.2019.11.019.
17. Konečný J, McMahan B, Ramage D. Federated Optimization:Distributed Optimization Beyond the Datacenter. arXiv; 2015 [cited 2023 Mar 29]. [Available from: http://arxiv.org/abs/1511.03575]
18. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2020 Jun 8;2(6):305–11. doi: 10.1038/s42256-020-0186-1
19. Scherer J, Nolden M, Kleesiek J, Metzger J, Kades K, Schneider V, et al. Joint Imaging Platform for Federated Clinical Data Analytics. JCO Clin Cancer Inform 2020 Nov;(4):1027–38. doi: 10.1200/CCI.20.00045.
20. Sarma KV, Harmon S, Sanford T, Roth HR, Xu Z, Tetreault J, et al. Federated learning improves site performance in multicenter deep learning without data sharing. J Am Med Inform Assoc 2021 Jun 12;28(6):1259–64. doi: 10.1093/jamia/ocaa341.
21. Agbley BLY, Li J, Hossin MA, Nneji GU, Jackson J, Monday HN, et al. Federated Learning-Based Detection of Invasive Carcinoma of No Special Type with Histopathological Images. Diagnostics 2022 Jul 9;12(7):1669. doi: 10.3390/diagnostics12071669.
22. Lu MY, Chen RJ, Kong D, Lipkova J, Singh R, Williamson DFK, et al. Federated learning for computational pathology on gigapixel whole slide images. Med Image Anal 2022 Feb;76:102298. doi: 10.1016/j.media.2021.102298.
23. Bercea CI, Wiestler B, Rueckert D, Albarqouni S. Federated disentangled representation learning for unsupervised brain anomaly detection. Nat Mach Intell 2022 Aug 25;4(8):685–95. doi: 10.1038/s42256-022-00515-2.
24. Rajendran S, Obeid JS, Binol H, D Agostino R, Foley K, Zhang W, et al. Cloud-Based Federated Learning Implementation Across Medical Centers. JCO Clin Cancer Inform 2021 Dec;(5):1–11. doi: 10.1200/CCI.20.00060.
25. Hansen CR, Price G, Field M, Sarup N, Zukauskaite R, Johansen J, et al. Larynx cancer survival model developed through open-source federated learning. Radiother Oncol 2022 Nov;176:179–86. doi: 10.1016/j.radonc.2022.09.023.
26. Zerka F, Barakat S, Walsh S, Bogowicz M, Leije-

naar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform 2020 Nov;(4):184–200. doi: 10.1200/CCI.19.00047.
27. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. Nat Commun. 2022; Nat Commun 2022 Dec 5;13(1):7346. doi: 10.1038/s41467-022-33407-5.
28. Nasr M, Shokri R, Houmansadr A. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In: 2019 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE; 2019 [cited 2023 Jan 29]. p. 739–53. [Available from: https://ieeexplore.ieee.org/document/8835245/].
29. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun 2021 Oct 11;12(1):5910. doi: 10.1038/s41467-021-25972-y.
30. Joel MZ, Umrao S, Chang E, Choi R, Yang DX, Duncan JS, et al. Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. JCO Clin Cancer Inform 2022 May;(6):e2100170. doi: 10.1200/CCI.21.00170.
31. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin Bhagoji A, et al. Advances and Open Problems in Federated Learning. Found Trends® Mach Learn 2021;14(1–2):1–210.
32. Kuo TT, Jiang X, Tang H, Wang X, Bath T, Bu D, et al. iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching. BMC Med Genomics 2020 Jul;13(S7):98, s12920-020-0715–0. doi: 10.1186/s12920-020-0715-0.
33. Blatt M, Gusev A, Polyakov Y, Rohloff K, Vaikuntanathan V. Optimized homomorphic encryption solution for secure genome-wide association studies. BMC Med Genomics 2020 Jul;13(S7):83. doi: 10.1186/s12920-020-0719-9.
34. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. Proc Natl Acad Sci 2020 May 26;117(21):11608–13. doi: 10.1073/pnas.1918257117.
35. Khilji IQ, Saha K, Amin J, Iqbal M. Application of Homomorphic Encryption on Neural Network in Prediction of Acute Lymphoid Leukemia. Int J Adv Comput Sci Appl 2020 [cited 2023 Jan 26];11(6). [Available from: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=6&Code=IJACSA&SerialNo=46].
36. Son Y, Han K, Lee YS, Yu J, Im YH, Shin SY. Privacy-preserving breast cancer recurrence prediction based on homomorphic encryption and secure two party computation. Vijayakumar P, editor. PLoS One 2021 Dec 20;16(12):e0260681. doi: 10.1371/journal.pone.0260681.
37. Paddock S, Abedtash H, Zummo J, Thomas S. Proof-of-concept study: Homomorphically encrypted data can support real-time learning in personalized cancer medicine. BMC Med Inform Decis Mak 2019 Dec;19(1):255. doi: 10.1186/s12911-019-0983-9.
38. Raisaro JL, Klann JG, Wagholikar KB, Estiri H, Hubaux JP, Murphy SN. Feasibility of Homomorphic Encryption for Sharing I2B2 Aggregate-Level Data in the Cloud. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:176-85.
39. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019 Feb 19;14(2):e0212463. doi: 10.1371/journal.pone.0212463.
40. Common Data Model Harmonization Project FInal Report. 2020 Aug. [Available from : https://aspe.hhs.gov/sites/default/files/private/pdf/259016/CDMH-Final-Report-14August2020.pdf].
41. Hayman JA, Dekker A, Feng M, Keole SR, McNutt TR, Machtay M, et al. Minimum Data Elements for Radiation Oncology: An American Society for Radiation Oncology Consensus Paper. Pract Radiat Oncol 2019 Nov;9(6):395–401. doi: 10.1016/j.prro.2019.07.017.
42. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building Data Infrastructure to Evaluate and Improve Quality: PCORnet. J Oncol Pract 2015 May;11(3):204–6. doi: 10.1200/JOP.2014.003194.
43. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Inform 2021 Dec;(5):12–20. doi: 10.1200/CCI.20.00079.
44. Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform 2019 Aug;96:103239. doi: 10.1016/j.jbi.2019.103239.
45. Jeon H, You SC, Park J, Park RW. Conversion of Diagnosis and Chemotherapy Data in Electronic Health Records to Episode-based Oncology Extension of OMOP-CDM. [Available from : https://www.ohdsi.org/2019-us-symposium-showcase-12/].
46. Yu Y, Ruddy KJ, Wen A, Zong N, Chen J, Shah ND, et al. Integrating Electronic Health Record Data into the ADEpedia-on-OHDSI Platform for Improved Signal Detection: A Case Study of Immune-related Adverse Events. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:710-719.
47. Lee SM, Kim K, Yoon J, Park SK, Moon S, Lee SE, et al. Association between Use of Hydrochlorothiazide and Nonmelanoma Skin Cancer: Common Data Model Cohort Study in Asian Population. J Clin Med 2020 Sep 9;9(9):2910. doi: 10.3390/jcm9092910.
48. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: Clinical decision support and machine learning as a service – A training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services. PLoS One 2019 Oct 3;14(10):e0223010. doi: 10.1371/journal.pone.0223010.
49. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open 2021 Jul 31;4(3):ooab001. doi: 10.1093/jamiaopen/ooab001.
50. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. J Med Internet Res 2019 Mar 26;21(3):e13249. doi: 10.2196/13249.
51. Carnahan RM, Waitman LR, Charlton ME, Schroeder MC, Bossler AD, Campbell WS, et al. Exploration of PCORnet Data Resources for Assessing Use of Molecular-Guided Cancer Treatment. JCO Clin Cancer Inform 2020 Nov;(4):724–35. doi: 10.1200/CCI.19.00142.
52. Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. JCO Clin Cancer Inform 2020 Nov;(4):993–1001. doi: 10.1200/CCI.20.00059.
53. Chen J, Chiang Y. Applying the Minimal Common Oncology Data Elements (mCODE) to the Asia-Pacific Region. JCO Clin Cancer Inform 2021 Dec;(5):252–3. doi: 10.1200/CCI.20.00181.
54. Potter D, Brothers R, Kolacevski A, Koskimaki JE, McNutt A, Miller RS, et al. Development of CancerLinQ, a Health Information Learning Platform From Multiple Electronic Health Record Systems to Support Improved Quality of Care. JCO Clin Cancer Inform 2020 Nov;(4):929–37. doi: 10.1200/CCI.20.00064.
55. Guérin J, Laizet Y, Le Texier V, Chanas L, Rance B, Koeppel F, et al. OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology. JCO Clin Cancer Inform 2021 Dec;(5):256–65. doi: 10.1200/CCI.20.00094.
56. Everson J, Patel V, Adler-Milstein J. Information blocking remains prevalent at the start of 21st Century Cures Act: results from a survey of health information exchange organizations. J Am Med Inform Assoc 2021 Mar 18;28(4):727–32. doi: 10.1093/jamia/ocaa323.
57. Kozlov, M. NIH issues a seismic mandate: share data publicly. Nature. 2022 Feb 24;602(7898):558–9. doi: 10.1038/d41586-022-00402-1.

Correspondence to:
Dr. Sanjay Aneja
Department of Therapeutic Radiology
Yale School of Medicine
330 Cedar Street, CB326
New Haven, CT 06510
USA
E-mail: sanjay.aneja@yale.edu
Tel + 1 203 200 2000