Published in final edited form as:
Semin Radiat Oncol. 2022 October ; 32(4): 400–414. doi:10.1016/j.semradonc.2022.06.009.

Artificial Intelligence for Radiation Oncology Applications Using Public Datasets

Kareem A. Wahid[*], **Enrico Glerean**[†,‡], **Jaakko Sahlsten**[‡], **Joel Jaskari**[‡], **Kimmo Kaski**[‡], **Mohamed A. Naser**[*], **Renjie He**[*], **Abdallah S.R. Mohamed**[*], **Clifton D. Fuller**[*]

[*]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[†]Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, Espoo, Finland

[‡]Department of Computer Science, Aalto University School of Science, Espoo, Finland

Abstract

Artificial intelligence (AI) has exceptional potential to positively impact the field of radiation oncology. However, large curated datasets - often involving imaging data and corresponding annotations - are required to develop radiation oncology AI models. Importantly, the recent establishment of Findable, Accessible, Interoperable, Reusable (FAIR) principles for scientific data management have enabled an increasing number of radiation oncology related datasets to be disseminated through data repositories, thereby acting as a rich source of data for AI model building. This manuscript reviews the current and future state of radiation oncology data dissemination, with a particular emphasis on published imaging datasets, AI data challenges, and associated infrastructure. Moreover, we provide historical context of FAIR data dissemination protocols, difficulties in the current distribution of radiation oncology data, and recommendations regarding data dissemination for eventual utilization in AI models. Through FAIR principles and standardized approaches to data dissemination, radiation oncology AI research has nothing to lose and everything to gain.

Introduction

The rise of artificial intelligence (AI) and machine learning in recent years has seen an explosion in technical approaches to statistical modeling, driven mostly by innovation on large datasets in the computational science community. These groundbreaking efforts have benefitted immensely from the availability of large-scale highly annotated datasets, such as the ImageNet visual recognition task dataset,[1] or the Modified National Institute of Standards and Technology (MNIST) handwriting dataset.[2] These datasets have allowed researchers to train and benchmark AI algorithms using transparent and comparable methods through a known corpus with defined properties at large-scale, thereby substantially

Address reprint requests to: Clifton D. Fuller, Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX, 77030. CDFuller@mdanderson.org.

accelerating AI growth. In fact, some have purported that the modern AI age can be traced to the convolutional neural network AlexNet winning the ImageNet challenge in 2012;[3] clearly, this task could not have been accomplished without the careful construction and curation of the ImageNet platform several years earlier.[4] During the same interval, the “Big Cancer Data” era was arguably initiated by the availability of large-scale public cancer databases. Specifically, The Cancer Genome Atlas (TCGA) has had a transformative effect on oncologic research.[5] Crucial for the oncologic imaging community, and consequently the radiation oncology community, a subsidiary effort, The Cancer Imaging Archive (TCIA),[6] was originally developed as a repository for matched image data from TCGA cases; however, TCIA has become a leader in its own right for distributing reliable and high-quality datasets independent of TCGA. Subsequently, radiation oncology has now entered a new era for AI applications due to an increasing deluge of data afforded by resources such as TCIA. In the following sections, we review and discuss the role of data sharing in the radiation oncology ecosystem, important regulatory and ethical considerations, and contemporary databases available for public investigation.

FAIR Data Publication as a Normative Component of Modern Scientific Dissemination and Knowledge Generation as a Coherent Process

Importantly, the availability of big data on its own does not ensure the development of robust, reproducible, and clinically impactful datasets. This is particularly true in medical domains such as radiation oncology, where data can be scarce, heterogeneous, and often non-standardized.[7] To ensure the usefulness of large datasets and foster data re-usability, it is fundamental to follow the Findable, Accessible, Interoperable, Reusable (FAIR) principles for scientific data management.[8] These principles emphasize machine-actionability, that is, the capacity of computational systems to find, access, interoperate, and reuse data with minimal human intervention. To foster the widespread implementation of FAIR scientific principles in day-to-day scientific workflows, several steps can and should be taken by practitioners of open-science practices in the radiation oncology community. These practices should include the utilization of study preregistration, manuscript preregistration, open-access journals, and code/data sharing (Fig. 1). We briefly review these components here; an in-depth discussion on these concepts can be found in greater detail in Fuller et al.[9]

Scientific research is often plagued by widespread biases, for example, “p-hacking” (repeat measurements leading to significant findings),[10] among others.[11] Study preregistration serves as a tool to potentially reduce these pitfalls.[12] The concept of preregistration involves specifying details of a research plan on a registry, such as through the Open Science Framework,[13] before performing the research study. Notably, the practice of preregistration has been shown to increase analytical rigor and increase the publication of null findings.[14]

After study completion, the utilization of preprints, a manuscript that precedes formal peer review, allows for rapid dissemination of study findings to the general public.[15] Drawing inspiration from the well-known physics preprint server arXiv, in 2019 a clinical medicine preprint repository, medRxiv, was established.[16] Since its inception, medRxiv has quickly become a leading figure in medically related preprint dissemination.[17] The benefits of publishing preprints for radiation oncology are numerous, chief among which include near-immediate distribution of scientific findings, ease of use, circumvention of journal politics and dominant narratives, and complete access by the general public. Naturally, the use of preprints raises the potential for faulty analysis. However, these concerns seem minor compared to the relative benefit of rapid dissemination of information which could then be verified through indicators of internal validity and wider scientific community evaluation. Closely related to concepts regarding preprints is the use of open-access options for scientific manuscripts. By allowing wider distribution of scientific findings, open-access options may lead to a greater return on investment for scientific innovation.[18] As of 2008, the National Institutes of Health (NIH) has mandated publicly funded research be made available through PubMed Central.[19] Moreover, as of 2021, the Plan S initiative will require that all EU-funded efforts be published through open access journals or platforms.[20] These geopolitical pushes highlight the importance of open access models for manuscript dissemination and will likely be a core staple of future radiation oncology research.

Finally, data and code accessibility, which are particularly germane to AI-driven analyses in radiation oncology, form a more recently established central tenant for disseminating scientific information. An in-depth discussion of data repositories is described in other sections of this review (see “NCI Policy, Vision, and Supported Repositories” and “TCIA as a Model for Effective FAIR Data in Oncology”). Moreover, the need for structured data and corresponding annotations has prompted the rise of journals dedicated to publishing data and/or corresponding records explaining dataset contents, termed “data descriptors”.[21] Examples of journals publishing data descriptors include Medical Physics and Nature Scientific Data.[22] Finally, code repositories, such as GitHub, have allowed for the rapid and dynamic development of code related to scientific studies in the medical domain, and fostered community engagement for future developments.[23]

While the current components of scientific dissemination remain relatively independent, in the future, one could envision a modular framework where automated processes link these components in an integrated fashion (Fig. 2). Corresponding checklists ensuring proper completion of steps could allow for the routine integration of these components in scientific workflows. Through these processes, the quality of scientific dissemination would increase, which could have great potential for advances in radiation oncology research.

Annotation/Ontology Considerations for Public Datasets

Broadly, there are two main types of data relevant to AI model development using public datasets in radiation oncology - image data formats and image annotation formats. The majority of hospital picture archiving and communication system infrastructure stores medical images in Digital Imaging and Communications in Medicine (DICOM) format, which is the recognized international standard for image data storage.[24] For ease of distribution, DICOM data is often transformed into Neuroimaging Informatics Technology Initiative format, which is increasingly seen as a standard for reproducible imaging research.[25] While image data is often standardized and distributed relatively easily due to established nomenclature and customs, unfortunately, image annotations are not often stored in a single common format.[26] For example, image annotations can be stored in picture archiving and communication system and other systems as DICOM presentation state objects which may vary from vendor to vendor. The recently developed DICOM structured reporting (DICOM-SR) standard has allowed to somewhat combat this lack of standardization in annotation but there still remains problems with intersite variability of annotation templates.[27] For non-graphic annotations, the annotation and image markup format was recently developed and incorporated into DICOM-SR.[28] Within radiation oncology, the DICOM radiotherapy (DICOM-RT) standard has allowed for a systematic characterization of radiotherapy-related data.[29] However, DICOM-RT was not developed with AI/machine-learning applications in mind, thus annotations derived from these files can at times be challenging to incorporate directly into models; community-driven undertakings have developed some solutions to these problems.[30,31] Thus, ongoing efforts will be needed to reach a consensus on the best practices to standardize image annotation data for radiation oncology applications.

Outside of file objects and corresponding metadata standardization, there are additional considerations for radiation oncology annotation data that are crucial for public distribution. Specifically, nomenclature conventions can often vary across institutions. For example, nomenclature for target/non-target structures is often highly variable (Table 1). Subsequently, standardized nomenclatures applied to targets, normal tissue structures, and treatment planning concepts and metrics would allow for the more facile integration of radiation oncology data for standardized distribution and use in AI models. Therefore, from 2015–2018, the American Association of Physicists in Medicine (AAPM) collated a group of experts in imaging, radiation oncology, and machine learning (Task Group 263), to provide radiation oncology nomenclature guidelines for use in clinical trials, data-pooling initiatives, population-based studies, and routine clinical care.[32,33] Among the many accomplishments of AAPM Task Group 263, the group was able to standardize: (1) structure names across image processing and treatment planning system platforms, (2) nomenclature for dosimetric data, (3) templates for clinical trial groups, and (4) formalisms for nomenclature schema which could accommodate the addition of other structures defined in the future. Through these guidelines, annotation processes for radiation oncology datasets and corresponding public data deposition should be more easily standardized for use in reproducible and robust research applications.

International Differences in Privacy Guidelines

In research integrity and ethics, it is accepted across public, health, and research institutions around the world that the privacy of the study participants must always be respected and never compromised to avoid concrete risks such as identity theft, blackmailing, or any other possible adverse consequences that a subject might face if their identity is disclosed with the data. Institutional review boards and ethical committees ensure that scientific research follows the highest standards to protect the safety of human subjects. However, ethics is not law, and while the ethical review process is comparable across countries, the legal principles and interpretation of data protection vary considerably and even deter FAIR data reuse. Currently in the USA, the Health Insurance Portability and Accountability Act (HIPAA) advises the removal of 18 pieces of protected health information (PHI) when sharing de-identified data[34] (Table 2). By removing this PHI, human subject data in the

USA can be considered “anonymous” or at least “anonymous enough” so that their data is considered to not fall under HIPAA and can then be easily reused for scientific research. However, concerns have been raised regarding the potential for intricate biomedical data that could be used for re-identification purposes, for example, brain fingerprinting[35] and sophisticated algorithmic methods for re-identifying supposedly anonymous individuals. With these concerns in mind, the European Commission has worked on the General Data Protection Regulation (GDPR) which was unveiled in May 2018. GDPR’s main goal is to protect data subjects by ensuring that organizations respect and protect the personal data associated with an individual. While GDPR’s main goal is aligned with research ethics to enable full data protection of the data subject, its strict guidelines have however been perceived at times by the scientific community as an impediment to scientific progress.[36,37] The current conundrum of having FAIR data that should also adhere to GDPR is potentially solvable with improved techniques for de-identification as well as other approaches like differential privacy and federated analysis.[38]

Considerations for Anonymization, De-identification and Privacy-Enhancement (e.g., “De-facing”) for Public Datasets

Chief among concerns for utilizing patient-derived medical data in public datasets is proper anonymization and de-identification of PHI. The terms de-identification and anonymization, when applied to medical data can encompass a wide array of definitions, and at times can be vague, inconsistent, or even contradictory.[39] Often, data are said to be anonymized when PHI has been completely removed such that the data can no longer be associated with an individual in any manner, while de-identification refers to the general removal of PHI.[40] More technical definitions include de-identification referring to rule-based techniques to remove PHI, with anonymization referring to statistical/probabilistic techniques to remove PHI.[39] For the purposes of this review, we will refer to the terms de-identification and anonymization interchangeably.

For obvious pieces of medical information, such as patient name, medical record number, date of birth, and other demographics, the anonymization process is often straightforward in removing this data or meta-data from associated files. DICOM formatted files typically contain metadata that links images to information regarding patient demographics and image acquisition parameters, among other information.[41] In cases where meta-data is available in DICOM or similar files, these data can be stripped using commonly available tools, such as the Radiological Society of North America Clinical Trial Processor,[42] though care should still be taken to ensure all data has been properly modified. Moreover, PHI may be “burned” into existing images, such as when considering radiographs directly scanned into electronic medical records. These embedded pieces of information can often be removed using optical character recognition techniques or similar methods; however, they can provide additional challenges in ensuring the full removal of text-based PHI.[43]

Perhaps the subject of most controversy within the domain of publicly releasing patient medical image data concerns imaging data that contains readily identifiable facial features. HIPAA references “full-face photographs and any comparable images” as a part of PHI.

This raises specific concerns for high-resolution images where the intricacies of facial features can be reconstructed (i.e., CT scans, MRI, etc.) to generate similar "comparable" visualizations of a patient's face with relative ease. Several studies have shown the potential danger in releasing unaltered medical images containing facial features as they can often be easily recognized by humans or machines.[44–46] For example, using facial recognition software paired to MRI-derived facial reconstructions, 83% of research participants were able to be identified from their MRI scan.[46] Within the realm of radiation oncology, perhaps the avenue where the greatest amount of concern for facial identification is in the generation of public medical imaging datasets of head and neck cancer. While brain images are often processed such that obvious facial features are removed (i.e., skull stripping), these crude pre-processing techniques will likely remove important information for building predictive models with head and neck cancer imaging data, prohibiting their use in data resharing strategies. "De-facing" tools, where voxels that correspond to the areas of the patient's facial features are either removed or altered are one solution. However, they encounter the problem of information loss in important areas needed for predictive modeling or treatment planning. While several studies have investigated the effects of de-facing for applications in neuroimaging,[47–50] to our knowledge only one study has investigated the effects of de-facing tools for radiation oncology applications in head and neck cancer.[51] As an example use case, we demonstrate organs at risk are obscured with 4 state-of-the-art de-facing tools[49,52–54] (Fig. 3), which would have obvious downstream consequences for analysis. Importantly, Sahlsten et al. highlight that the specific selection of de-facing methods has a significant impact on the radiotherapy organs at risk for AI algorithmic development, in some cases rendering structures completely unusable.[51] While existing tools are seemingly unsatisfactory, future de-facing approaches based on deep learning may be a promising solution.[55]

NCI Policy, Vision, and Supported Repositories

Broadly, the sharing of biomedical data generated through research studies allows the scientific community to expedite the translation of findings into knowledge, products, and procedures to improve health.[56] As of 2003, the NIH has implemented a Data Sharing Policy which encourages data to "be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting con-fidential and proprietary data." Specifically, investigators were required to specify data sharing protocols within NIH grant applications. These protocols for data sharing could be accomplished through the use of data archives, data enclaves, or under the auspices of the principal investigator. Congruent with the increasing ability to generate, store, share, and combine data, in 2015 the NIH initiated a more comprehensive data sharing policy in tandem with efforts to modernize data sharing infrastructure in its Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research.[57] Effective in 2023, the NIH has issued a new Final NIH Policy for Data Management and Sharing which will require NIH funded researchers to prospectively submit a plan detailing how scientific data and metadata will be managed and shared taking into account potential restrictions or limitations[58]; this plan will replace the 2003 NIH Data Sharing Policy. These policies underscore the importance of data stewardship and management for nationally-funded research, and have influenced data sharing practices to funders of radiation oncology research, particularly the National Cancer Institute (NCI).

Within the Final NIH Policy for Data Management and Sharing, detailed documentation has been provided to help researchers select specific repositories for data deposition (“Selecting a Repository for Data Resulting from NIH-Supported Research”). Ideally, where applicable, data should be deposited in discipline or data-type specific repositories. The NIH has provided a list of approved specific repositories at https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html; a subset of radiation oncology related repositories is shown in Table 3. For an in-depth discussion on one specific NIH-approved data repository particularly salient for radiation oncology applications, we refer the reader to the section of our review titled “TCIA as a Model for Effective FAIR Data in Oncology”. When no appropriate disciplinary or data-type specific repositories are available, the NIH recommends the use of generalist or institutional repositories. For example, Figshare is an appropriate and well-established generalist repository that can permanently store datasets and assigns DOIs to all published research items.[59]

Driven by improvements and innovations in cloud-computing paradigms for use in big data research, the NCI has created the Cancer Research Data Commons (CRDC) as a component of a national cancer data ecosystem.[60] The NCI CRDC includes cloud-based domain-specific data repositories and analysis-focused cloud resources to facilitate collaborative and standardized research practices using diverse data types. Within the CRDC, several data repositories have been established, such as the Genomic Data Commons[61] and the Imaging Data Commons.[62] As opposed to previously described data repositories, the cloud-based infrastructure of the CRDC allows researchers to utilize data in real-time without the need for local downloading of data files. Moreover, through these data repositories, disparate data sources (imaging, genomic, proteomic, clinical trial, etc.) can be combined and investigated with compute resources provided by cloud environments, thereby providing researchers with the ability to perform robust harmonized analysis. An example of the IDC user portal interface is shown in Figure 4. The integrated analysis capabilities provided by infrastructure such as CRDC will likely have a significant impact on radiation oncology research in coming years.

TCIA as a Model for Effective FAIR Data in Oncology

TCIA is a recently developed service that de-identifies and hosts various medical imaging datasets and supporting data for public distribution. Launched in 2011 with funding from the NCI,[6] this integrated database has offered the imaging community large volumes of curated data for exploratory image analysis, computational model development, and model validation.[63] Provided human data originates from various sources, ranging from small-scale calibration studies to large-scale clinical trials. These imaging data have been crucial to developing contemporary medical AI models and catapulted TCIA as a de-facto leader of medical image data dissemination (Fig. 5). It stands to reason that TCIA will continue to provide much-needed high-quality datasets for clinical decision support tool development in the coming years, particularly for radiation oncology.

The overarching structure of the TCIA is stratified into individual collections defined by a common disease (e.g., brain cancer, head-neck cancer, lung cancer), imaging modality (e.g., MRI, CT, PET, histopathology), and/or research focus. Collections are assigned persistent digital object identifiers (DOIs), thereby allowing researchers to reference and acquire datasets.[64] DOIs contain “Primary Data”, that is, radiological or pathological images, which can be coupled to supporting data (demographics, clinical outcomes, annotations, genomic information, etc.). Before publication, datasets are rigorously curated to ensure acceptable image quality and data integrity. Currently, TCIA utilizes the Posda open-source framework[65] to aid in the curation process and remove any identifying information in metadata. Collection contents are described through “wiki pages”, which also list relevant publications and instructions for data use. Increasingly, focused “data descriptors,” in-depth manuscripts detailing individual datasets, such as those published through Nature Scientific Data,[22] are also generated for TCIA collections to engender greater transparency in data generation, collection protocols, and intended use-cases. For end-users, TCIA provides web interfaces and software (National Biomedical Imaging Archive Data Retriever) to easily retrieve and catalog collections on local computing infrastructure.

TCIA data is most often available through standardized imaging formats such as DICOM.[24] Importantly, TCIA has formed a corpus for not only raw imaging data, but also corresponding supporting data, such as region of interest segmentations through DICOM radiotherapy structure set (RTSTRUCT) files and clinical outcome data. Therefore, TCIA houses a rich stream of information for supervised machine learning segmentation and classification models. For radiation oncology applications, DICOM radiotherapy plan (RTPLAN) and DICOM radiotherapy dose (RTDOSE) are often also included in collections which can be used for model development germane to radiation therapy planning. A list of currently available TCIA collections that include RTPLAN and/or RTDOSE data is shown in Table 4. Currently, most collections with corresponding radiotherapy planning data correspond to head and neck cancer. Importantly, this subset of data comes with additional important considerations for re-use (discussed more in the “Considerations for Anonymization, De-identification and Privacy-Enhancement (e.g., “De-facing”) for Public Datasets” section). As TCIA continues to engender straightforward integration of community contributions, the number of collections that include radiotherapy-related data is expected to continue to rapidly increase over time.

Examples of Published “Challenges” Using Public Datasets

Data “challenges,” that is, competitions where datasets are publicly provided to interested participants to solve a specific problem, have been a staple in developing modern-day cutting-edge AI algorithms. For example, the ImageNet Large Scale Visual Recognition Challenge, a competition where participants are tasked with classifying photographic images of common objects, was the impetus for the rise of deep learning approaches for computer vision applications.[66] A similar trend has emerged in the medical domain, particularly for medical imaging, where anonymized data is provided to challenge participants to solve important healthcare problems.[67,68] The field of radiation oncology, which is heavily centered on image-based workflows, is no exception to this increasing trend, with several radiotherapy-related data challenges emerging in recent years. Here, we summarize a few key data challenges that have been particularly impactful for radiation oncology applications.

In 2016, inspired by the up-and-coming trend of radiomics, i.e., the use of quantitative features derived from medical imaging,[69] the University of Texas MD Anderson Cancer Center and the Medical Image Computing and Computer Assisted Intervention Society organized two public radiomics challenges in the head and neck radiation oncology domain.[70] Through the Kaggle InClass commercial educationally-oriented platform,[71] the organizers tasked participants to develop predictive models to: (1) classify patients based on human papillomavirus (HPV) status, and (2) predict local tumor recurrence status. A large number of contrast-enhanced CT images of oropharyngeal cancer patients and corresponding clinical data[72] were provided to participants to build models to solve the 2 tasks through evaluation of independent test data (Fig. 6). The majority of participants used pre-defined radiomic features extracted from images in combination with machine learning models to solve the 2 challenges. Many participants also utilized the provided clinical data in constructing their models. Interestingly, the winner of the HPV classification challenge only utilized radiomic features, while the winner of the recurrence prediction challenge only utilized clinical features. While the challenge has not been renewed for additional iterations, these important results highlighted the difficulty of integrating imaging and clinical data not only for designing data challenges in radiation oncology but also for eventual downstream model implementation.

In 2019, the AAPM hosted a 2-part competition, titled the “RT-MAC” challenge, for auto-segmentation of radiotherapy-related structures using MRI scans.[73] Participants were tasked with developing algorithms to segment parotid glands, submandibular glands, and various lymph node levels (Fig. 7). The challenge used a relatively limited number of training cases ($n = 35$), but participants were still able to generate segmentation results of reasonable quality on the independent pre-AAPM challenge test set ($n = 10$) and online challenge test set ($n = 10$). These datasets have become publicly available in their entirety through TCIA,[73] allowing for the community to continue improving upon methods for radiotherapy planning segmentation. Moreover, given the rapidly increasing interest in MRI-guided radiotherapy,[74] it is foreseeable that analogous datasets could be released in the near future to aid in adaptive radiotherapy auto-segmentation applications.

More recently, the HEad and neCK TumOR (HECKTOR) challenge was established to benchmark the utility of computational methods using PET/CT imaging for head and neck cancer radiotherapy-related applications. Initiated in 2020 through the Medical Image Computing and Computer Assisted Intervention Society, the first edition of the challenge sought to develop automatic methods to segment primary gross tumor volumes in patients with oropharyngeal cancer (Fig. 8).[75] The challenge utilized data from multiple Canadian and European medical institutions to provide participants with highly curated imaging and segmentation data. Participants were able to develop AI models, predominantly based on deep learning, to generate high-quality primary tumor segmentations on unseen test data. In the 2021 edition of the challenge, additional imaging data from a greater number of institutions were added to the training and testing datasets. Moreover, new tasks based on the prediction of progression-free survival were integrated into the challenge. While the official post-challenge analysis for the 2021 edition has not been made public yet, participants were shown to improve upon segmentation from the previous year's challenge[76] and demonstrate encouraging results to predict prognosis.[77,78] A new edition of the challenge has been planned for 2022, which will include incorporating metastatic cervical lymph node segmentations and additional datasets from more institutions. The HECKTOR challenge is increasingly seen as a leader in current day radiation oncology related data challenges; we anticipate it will lead to important clinical innovations for translational AI approaches in coming years as its corresponding datasets continue to mature.

Health Equity Considerations related to Public Datasets

While AI holds immense promise in improving the radiation oncology workflow through public medical imaging datasets, a thorough understanding of the current limitations of existing AI approaches is crucial before their widespread implementation. One such limitation is algorithmic bias/fairness, which if severe enough could further inequity and disparities in patient care. Algorithmic bias is not unique to advanced machine learning approaches. For example, a landmark paper by Obermeyer et al. demonstrated a widely used commercial algorithmic risk score based on simple demographic factors severely underestimated the health needs of the sickest marginalized groups (black patients) by focusing on financial costs.[79] Subsequently, it is crucial to capture the biases in AI systems before they can be deployed in large-scale clinical settings. Unfortunately, most guiding principles for machine learning in healthcare applications do not directly address model fairness in detail.[80] However, there is growing widespread interest in racial, gender, and socioeconomic disparities of AI-based healthcare algorithms.

Radiation oncology is not immune to biases in patient care, both at the level of the individuals,[81] and systemically.[82] It is well documented that marginalized racial groups often receive inferior care compared to wealthy or white patients,[83] so the potential amplification of inequality caused by AI software in radiation oncology is a significant concern. While racial disparities have been modestly investigated for healthcare in AI generally, specific mechanisms for the existence of these disparities in imaging data/models remains relatively unexplored. However, a recent study by Banerjee et al.[84] demonstrated that standard deep learning models could predict self-identified race from medical images with high performance. Importantly, they showed that this ability was not due to imaging-related surrogate covariates for race. Moreover, the performance of models persisted over a wide spectrum of clinical applications and image modalities, suggesting a significant and prevalent problem that warrants further investigation. These results are important since they suggest AI can trivially predict race where clinical experts cannot, thereby limiting human oversight and leading to potential downstream disparities, particularly for minorities.

In an ideal setting, machine learning models should be trained and evaluated on data that accurately represent real-world data. In designing public data challenges, these concepts should be considered in the curation of training and test sets. Since data-driven methods inherently recognize patterns in training data, any bias already present in the data will be propagated to downstream models. Naturally, all datasets at some level will contain biases inherently tied to the sampling procedures. Importantly, biased sampling may lead to inaccurate predictions in unseen evaluation data, as illustrated in Figure 9. For example, data from public access clinical repositories are often disproportionately represented by Caucasian males. Ensuring representative sampling across time and data sources is an important method to reduce bias inherent to training data.[85]

Recommendations for operationalizing fairness for AI in medical data have been previously suggested,[80] and these approaches should be subsequently implemented in imaging data for radiation oncology applications/data challenges. When curating datasets for data challenges or public dissemination, regardless of the target application, data on race, ethnicity, and socioeconomic status should also be collected and made available in order to assess their relationship to the underlying models where appropriate. Methods to circumvent algorithmic bias would include increases in model interpretability/explainability, either through inherently interpretable models or post-hoc techniques.[86] Moreover, when developing and conceptualizing new AI models, it stands to reason that individuals with a vested interest in combating inequalities should be included in discussions to address potential sources and consequences of bias.[87]

Conclusion

In summary, we describe how radiation oncology has benefited from FAIR scientific data distribution principles and will continue to benefit in the coming years. Given increasing attention by governing institutions and collaborative efforts, dissemination of radiation oncology data through structured repositories and public data challenges have led to algorithmic development and advancement, particularly with respect to AI-driven clinical decision support tools. A variety of concerns still plague the public dissemination of radiation oncology data, namely proper protection of patient PHI, ensuring standardized data objects and nomenclature, addressing health equity concerns, and consolidation of individual components of scientific dissemination. However, the future of public data distribution remains bright, and is certain to lead to continued innovation and clinical impact within the radiation oncology community.

Acknowledgments

This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672-44). K.A.W. is supported by the Dr. John J. Kopchick Fellowship through The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences, the American Legion Auxiliary Fellowship in Cancer Research, and an NIH/National Institute for Dental and Craniofacial Research (NIDCR) F31 fellowship (1 F31DE031502-01). C.D.F. received funding from the NIH/NIDCR (1R01DE025248-01/R56DE025248); an NIH/NIDCR Academic-Industrial Partnership Award (R01DE028290); the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the NCI Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825); the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148); an NIH/NCI Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50CA097007); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787).

References

1. Morid MA, Borjali A, Del Fiol G: A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 128, 2021:104115 [PubMed: 33227578]
2. Deng L: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag 29:141–142, 2012
3. Krizhevsky A, Sutskever I, Hinton GE: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25. Curran Associates, Inc., 2012.
4. Gershgorn D. The data that transformed AI research—and possibly the world. Quartz. Available at: https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/. Accessed February 1, 2022.
5. Tomczak K, Czerwińska P, Wiznerowicz M: The cancer genome atlas (TCGA): An immeasurable source of knowledge. Contemp Oncol (Pozn) 19:A68–A77, 2015 [PubMed: 25691825]
6. Clark K, Vendt B, Smith K, et al. : The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26:1045–1057, 2013 [PubMed: 23884657]
7. Mayo CS, Kessler ML, Eisbruch A, et al. : The big data effort in radiation oncology: Data mining or data farming? Adv Radiat Oncol 1:260–271, 2016 [PubMed: 28740896]
8. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. : The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:1–9, 2016
9. Fuller CD, van Dijk LV, Thompson RF, et al. : Meeting the challenge of scientific dissemination in the Era of COVID-19: Toward a modular approach to knowledge-sharing for radiation oncology. Int J Radiat Oncol Biol Phys 108:496–505, 2020 [PubMed: 32890543]
10. Sterling TD, Rosenbaum WL, Weinkam JJ: Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. Am Statistician 49:108–112, 1995
11. Nagarajan P, Garla B, Taranath M, et al. : The file drawer effect: A call for meticulous methodology and tolerance for non-significant results. Indian J Anesth 61, 2017
12. P Simmons J, D Nelson L, Simonsohn U: Pre-registration: Why and how. J Consum Psychol 31:151–162, 2021
13. Foster ED, Deardorff A: Open science framework (OSF). J Med Library Assoc 105:203, 2017
14. Allen C, Mehler DM: Open science challenges, benefits and tips in early career and beyond. PLoS Biol 17, 2019:e3000246 [PubMed: 31042704]
15. Bourne PE, Polka JK, Vale RD, et al. : Ten simple rules to consider regarding preprint submission. PLoS Comput Biol 13, 2017:e1005473. Public Library of Science San Francisco, CA USA [PubMed: 28472041]
16. Rawlinson C, Bloom T: New preprint server for medical research. BMJ 365, 2019. British Medical Journal Publishing Group
17. Krumholz HM, Bloom T, Sever R, et al. : Submissions and downloads of preprints in the first year of medRxiv. JAMA 324:1903–1905, 2020 [PubMed: 33170232]
18. Laakso M, Welling P, Bukvova H, et al. : The development of open access journal publishing from 1993 to 2009. PLoS One 6:e20961, 2011 [PubMed: 21695139]
19. Zerhouni EA: NIH public access policy. Science 306, 2004. 1895–1895 American Association for the Advancement of Science [PubMed: 15591188]
20. Plan S Accelerating the Transition to Full and Immediate Open Access to Scientific Publications. Science Europe, 2019.. Available at: https://www.scienceeurope.org/wp-content/uploads/2018/09/Plan_S. Accessed October 13, 2018. Title from the screen
21. Ward CH: Introducing the Data descriptor article. Integrating Materials and Manufacturing Innovation, 4. Springer, 190–191, 2015
22. Candela L, Castelli D, Manghi P, et al. : Data journals: A survey. J Assoc Inf Sci Technol 66:1747–1762, 2015
23. Bouquin DR: GitHub. J Med Libr Assoc 103:166, 2015
24. Mildenberger P, Eichelberg M, Martin E: Introduction to the DICOM standard. Eur Radiol 12:920–927, 2002 [PubMed: 11960249]

25. Li X, Morgan PS, Ashburner J, et al. : The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 264:47–56, 2016 [PubMed: 26945974]

26. Willemink MJ, Koszek WA, Hardell C, et al. : Preparing medical imaging data for machine learning. Radiology 295:4–15, 2020 [PubMed: 32068507]

27. Clunie DA: DICOM structured reporting and cancer clinical trials results. Cancer Inform 4, 2007. CIN. S37032

28. Channin DS, Mongkolwat P, Kleper V, et al. : The caBIG™ annotation and image markup project. J Digit Imaging 23:217–225, 2010 [PubMed: 19294468]

29. Law MYY, Liu B: DICOM-RT and its utilization in radiation therapy. Radiographics 29:655–667, 2009 [PubMed: 19270073]

30. Mason D: Pydicom: An open source DICOM library. Med Phys 38, 2011. 3493–3493

31. Anderson BM, Wahid KA, Brock KK, et al. : Simple python module for conversions between DICOM Images and radiation therapy structures, masks, and prediction arrays. Practical Radiat Oncol 11:226–229, 2021

32. Mayo CS, et al. : American association of physicists in medicine task group 263: Standardizing nomenclatures in radiation oncology. Int J Radiat Oncol Biol Phys 100:1057–1066, 2018 [PubMed: 29485047]

33. Mayo C, Moran JM, Bosch W, et al.: Standardizing Nomenclatures in Radiation Oncology: The Report of AAPM Task Group 263. Available at: https://www.aapm.org/pubs/reports/RPT_263.pdf

34. Meystre SM, Friedlin FJ, South BR, et al. : Automatic de-identification of textual documents in the electronic health record: A review of recent research. BMC Med Res Method 10:1–16, 2010

35. Ravindra V & Grama A De-anonymization attacks on neuroimaging datasets. in 2394–2398 (2021).

36. Peloquin D, DiMaio M, Bierer B, et al. : Disruptive and avoidable: GDPR challenges to secondary research uses of data. Eur J Hum Genet 28:697–705, 2020 [PubMed: 32123329]

37. Molnár-Gábor F, Korbel JO: Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall? EMBO Mol Med 12: e11421, 2020 [PubMed: 32072760]

38. Wei K, Li J, Ding M, et al. : Federated learning with differential privacy: Algorithms and performance analysis. IEEE Trans Inf Forensics Secur 15:3454–3469, 2020

39. Chevrier R, Foufi V, Gaudet-Blavignac C, et al. : Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. J Med Internet Res 21:e13484, 2019 [PubMed: 31152528]

40. Guidelines for Data De-Identification or Anonymization. EDUCAUSE, Boulder, Colorado, USA. Available at: https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/resources/information-security-guide/toolkits/guidelines-for-data-deidentification-or-anonymization.

41. Diaz O, Kushibar K, Osuala R, et al. : Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Physica Med 83:25–37, 2021

42. Aryanto KYE, Oudkerk M, van Ooijen PM A: Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol 25:3685–3695, 2015 [PubMed: 26037716]

43. Willemink MJ, Koszek WA, Hardell C, et al. : Preparing medical imaging data for machine learning. Radiology 295:4–15, 2020 [PubMed: 32068507]

44. Prior FW, Brunsden B, Hildebolt C, et al. : Facial recognition from volume-rendered magnetic resonance imaging data. IEEE Trans Inf Technol Biomed 13:5–9, 2008

45. Mazura JC, Juluru K, Chen JJ, et al. : Facial recognition software success rates for the identification of 3D surface reconstructed facial images: Implications for patient privacy and security. J Digit Imaging 25:347–351, 2012 [PubMed: 22065158]

46. Schwarz CG, Kremers WK, Therneau TM, et al. : Identification of anonymous MRI research participants with face-recognition software. N Engl J Med 381:1684–1686, 2019 [PubMed: 31644852]

47. Schwarz CG, Kremers WK, Wiste HJ, et al. : Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. Neuroimage 231, 2021:117845 [PubMed: 33582276]

48. Schimke N, Kuehler M, Hale J: Preserving privacy in structural neuroimages. In: IFIP Annual Conference on Data and Applications Security and Privacy, Springer, 301–308, 2011

49. Bischoff Grethe A, Ozyurt IB, Busa E, et al. : A technique for the deidenti-fication of structural brain MR images. Hum Brain Mapp 28:892–903, 2007 [PubMed: 17295313]

50. Theyers AE, Zamyadi M, O'Reilly M, et al. : Multisite comparison of MRI defacing software across multiple cohorts. Front Psychiatry 12:189, 2021

51. Sahlsten J, Wahid KA, Glerean E, et al. : Segmentation Stability of Human Head and Neck Medical Images for Radiotherapy Applications Under de-identification Conditions: Benchmarking for Data Sharing and Artificial Intelligence Use-Cases. medRxiv, 2022. 2022.01.22.22269695

52. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. : Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166:400–424, 2018 [PubMed: 29079522]

53. Gulban O, Nielson D, Poldrack R, et al. Poldracklab/pydeface: v2.0.0 [Internet]

54. Milchenko M, Marcus D: Obscuring surface anatomy in volumetric imaging data. Neuroinformatics 11:65–75, 2013 [PubMed: 22968671]

55. Jeong YU, Yoo S, Kim Y-H, et al. : De-identification of facial features in magnetic resonance images: software development using deep learning technology. J Med Internet Res 22:e22739, 2020 [PubMed: 33208302]

56. Tenopir C, Allard S, Douglass K, et al. : Data sharing by scientists: Practices and perceptions. PLoS One 6:e21101, 2011 [PubMed: 21738610]

57. National Institutes of Health: Plan for Increasing Access to Scientific Publications and Digital Scientific Data From NIH Funded Scientific Research. Bethesda: National Institutes of Health, Bethesday, Maryland, USA. 2015.

58. National Institutes of Health: Final NIH Policy for Data Management and Sharing. Bethesda, MD: Office of The Director, National Institutes of Health, 2020. NOT-OD-21–013. Vol NOT-OD-21–013. NIH Grants & Funding

59. Thelwall M, Kousha K: Figshare: A universal repository for academic resource sharing? Online Inform Rev 40:333–346, 2016

60. Hinkson IV, Davidsen TM, Klemm JD, et al. : A comprehensive infrastructure for big data in cancer research: Accelerating cancer research and precision medicine. Front Cell Dev Biol 5, 2017

61. Heath AP, Ferretti V, Agrawal S, et al. : The NCI genomic data commons. Nat Genet 53:257–262, 2021 [PubMed: 33619384]

62. Fedorov A, Longabaugh WJ, Pot D, et al. : NCI imaging data commons. Cancer Res 81:4188–4193, 2021 [PubMed: 34185678]

63. Kalpathy-Cramer J, Freymann JB, Kirby JS, et al. : Quantitative imaging network: data sharing and competitive algorithmvalidation leveraging the cancer imaging archive. Transl Oncol 7:147–152, 2014 [PubMed: 24772218]

64. Prior F, Smith K, Sharma A, et al. : The public cancer radiology imaging collections of the cancer imaging archive. Scientific Data 4, 2017:170124 [PubMed: 28925987]

65. Bennett W, Smith K, Jarosz Q, et al. : Reengineering workflow for curation of DICOM datasets. J Digit Imaging 31:783–791, 2018 [PubMed: 29907888]

66. Russakovsky O, Deng J, Su H, et al. : Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252, 2015

67. Prevedello LM, Halabi SS, Shih G, et al. : Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiology 1, 2019:e180031

68. Elhalawani H, Mak R: Are artificial intelligence challenges becoming radiology's new "Bee's Knees"? Radiology 3, 2021:e210056 [PubMed: 34138989]

69. Aerts HJ, Velazquez ER, Leijenaar RT, et al. : Decoding tumour pheno-type by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:1–9, 2014

70. Elhalawani H, Lin TA, Volpe S, et al. : Machine learning applications in head and neck radiation oncology: Lessons from open-source radiomics challenges. Front Oncol 8, 2018

71. Polak J, Cook D: A study on student performance, engagement, and experience with kaggle inclass data challenges. J Stat Data Sci Educ 29:63–70, 2021

72. Elhalawani H, Mohamed AS, White AL, et al. : Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci Data 4, 2017:170077 [PubMed: 28675381]

73. Cardenas CE, Mohamed AS, Yang J, et al. : Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. Med Phys 47:2317–2322, 2020 [PubMed: 32418343]

74. Pollard JM, Wen Z, Sadagopan R, et al. : The future of image-guided radiotherapy will be MR guided. Br J Radiol 90, 2017:20160667

75. Oreiller V, Andrearczyk V, Jreige M, et al. : Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med Image Anal 77, 2022:102336 [PubMed: 35016077]

76. Naser MA, Wahid KA, Dijk LV et al.: Head and Neck Cancer Primary Tumor Auto Segmentation Using Model Ensembling of Deep Learning in PET/CT Images. In 3D Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 121–133. Springer, Cham, 2021.

77. Naser MA, Wahid KA, Mohamed AS, et al.: Progression Free Survival Prediction for Head and Neck Cancer Using Deep Learning Based on Clinical and PET/CT Imaging Data. In 3D Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 287–299. Springer, Cham, 2021.

78. Wahid KA, He R, Dede C, et al.: Combining tumor segmentation masks with PET/CT images and clinical data in a deep learning framework for improved prognostic prediction in head and neck squamous cell carcinoma. In 3D Head and Neck Tumor Segmentation in PET/CT Challenge, pp. 300–307. Springer, Cham, 2021.

79. Obermeyer Z, Powers B, Vogeli C, et al. : Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453, 2019 [PubMed: 31649194]

80. Wawira Gichoya J, McCoy LG, Celi LAG. et al. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform. 28:e100289, 2021.10.1136/bmjhci-2020-100289.

81. Diaz DA, Suneja G, Jagsi R, et al. : Mitigating implicit bias in radiation oncology. Adv Radiat Oncol 6, 2021:100738 [PubMed: 34381930]

82. Chapman CH, Gabeau D, Pinnix CC, et al. : Why racial justice matters in radiation oncology. Adv Radiat Oncol 5:783–790, 2020 [PubMed: 32838067]

83. Brondolo E, Gallo LC, Myers HF: Race, racism and health: disparities, mechanisms, and interventions. J Behav Med 32:1–8, 2009 [PubMed: 19089605]

84. Banerjee I, Bhimireddy AR, Burns JL, et al. : Reading race: AI recognises patient's racial identity. Medical Images. arXiv, 2021. preprint arXiv:2107.10356

85. Chua IS, Gaziel-Yablowitz M, Korach ZT, et al. : Artificial intelligence in oncology: Path to implementation. Cancer Med 10:4138–4149, 2021 [PubMed: 33960708]

86. Vokinger KN, Feuerriegel S, Kesselheim AS: Mitigating bias in machine learning for medicine. Commun Med 1:25, 2021 [PubMed: 34522916]

87. Algebra H: How can we keep algorithmic racism out of Canadian health care's AI toolkit? Globe Mail 2121. https://www.theglobeandmail.com/life/health-and-fitness/article-how-can-we-keep-algorithmic-racism-out-of-canadian-health-cares-ai/

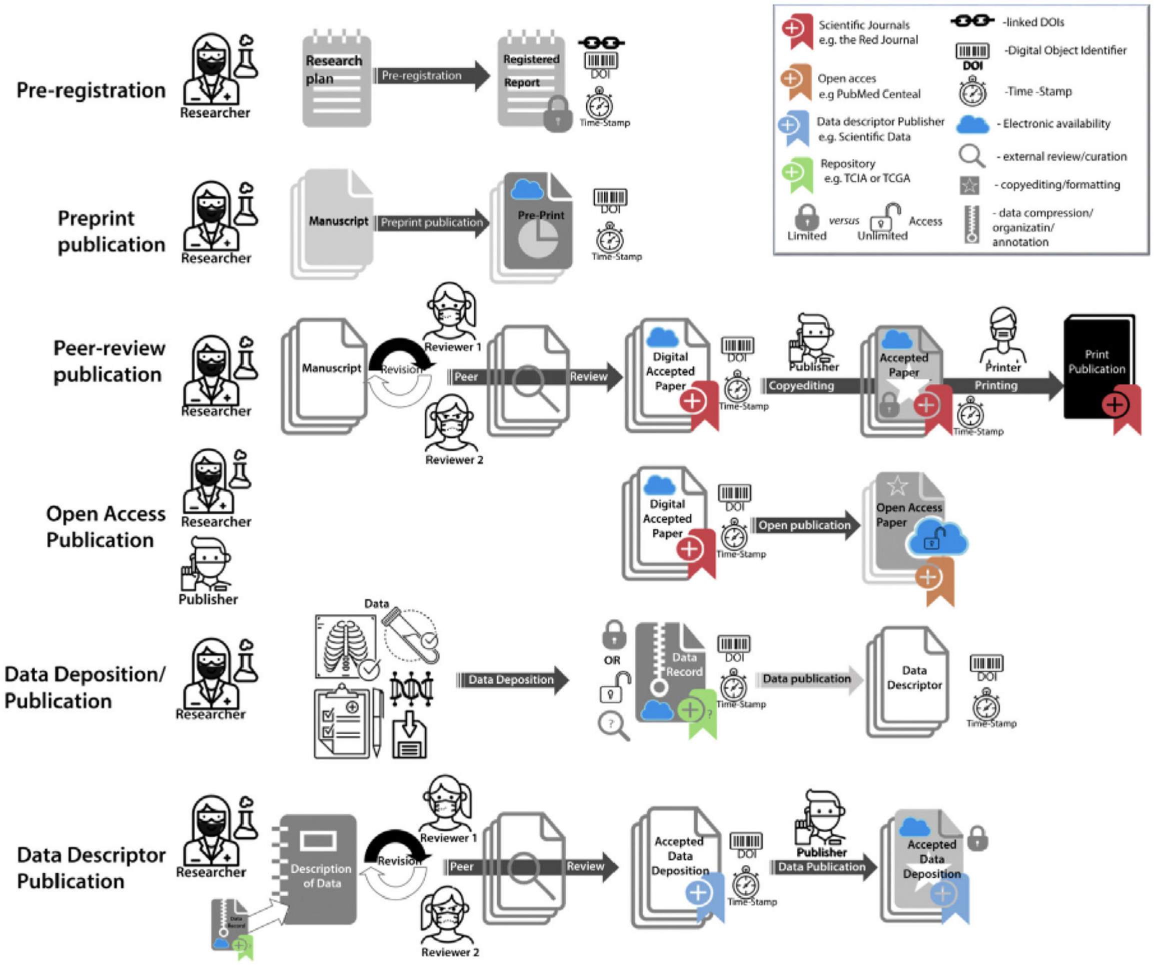


Figure 1.

Graphical representation of current independent steps in scientific dissemination process. Reprinted from Fuller et al.[9]

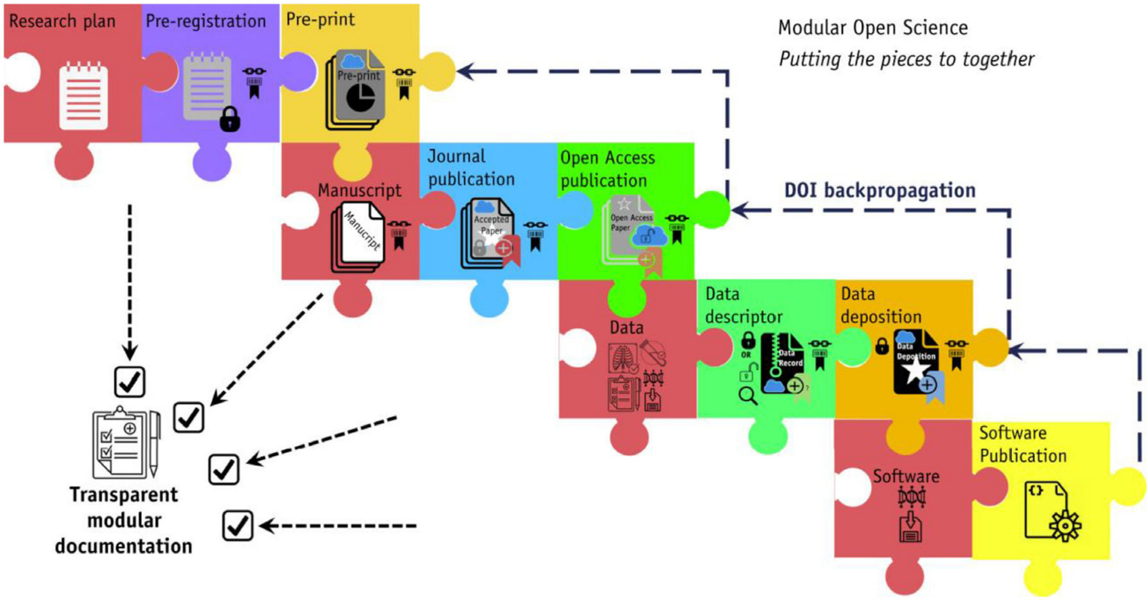


Figure 2.

Proposed transparent modular scientific dissemination process, using metadata or digital object identifier (DOI) to link individual processes. Reprinted from Fuller et al.[9]

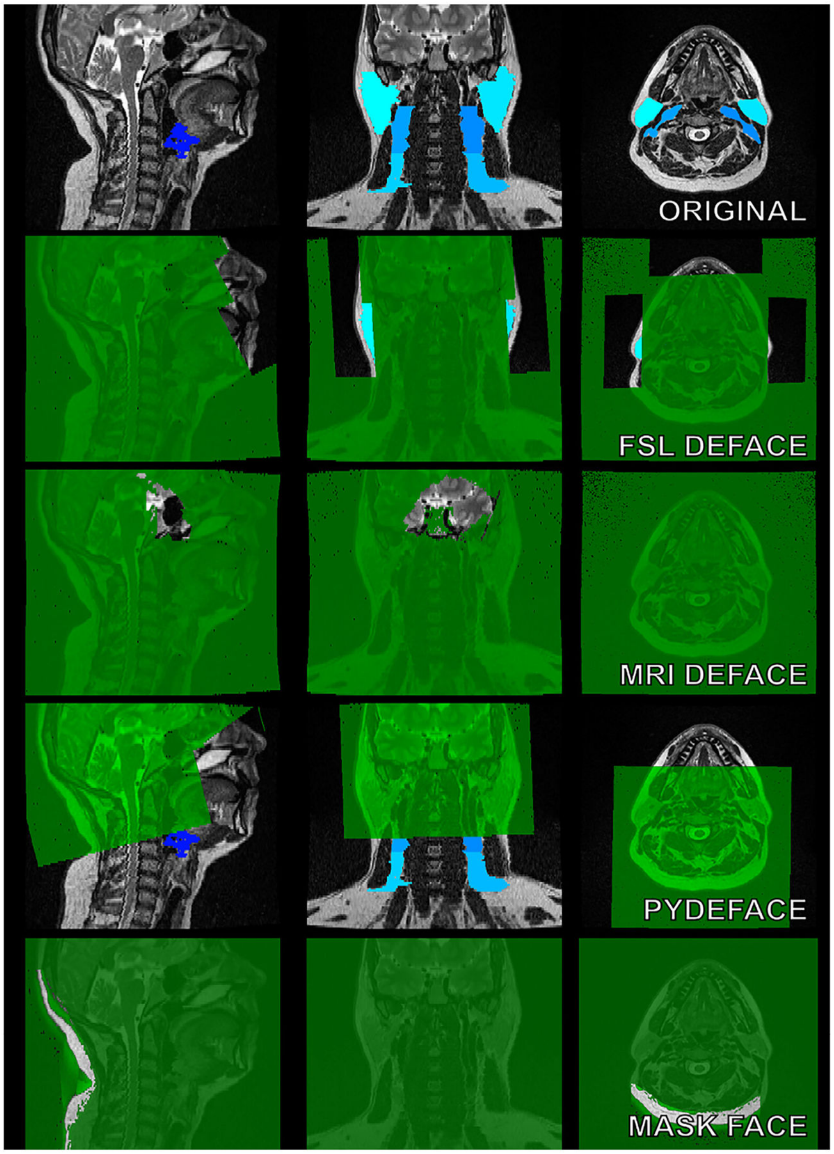
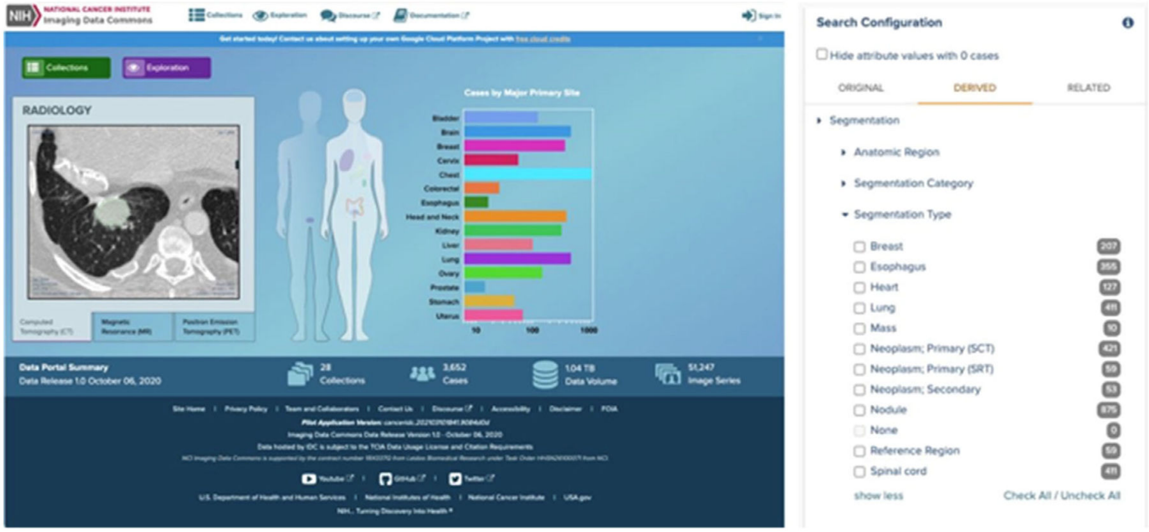


Figure 3.

Comparison of MRI de-facing tools. We tested the performance of 4 state-of-the-art tools for face de-identification. On the top row, the original T2-weighted MRI with tissue annotations (lymph node levels, glands) are shown in blue. Masks that each tool automatically creates to remove facial structures from the image volume are shown in green. These tools are popular within the neuroimaging community and were designed for defacing MRIs while preserving brain structures. In many cases, tissues of interest for radiation oncology applications are obscured or removed through the use of these tools.


NIH
NATIONAL CANCER INSTITUTE
Imaging Data Commons
Collections
Exploration
Discourse
Documentation
Sign in
Get started today! Contact us about setting up your own Google Cloud Platform Project with free cloud credits
Collections
Exploration
RADIOLOGY
Computed Tomography (CT)
Magnetic Resonance (MR)
Positron Emission Tomography (PET)
Cases by Major Primary Site
Bladder
Brain
Breast
Cervix
Chest
Colorectal
Esophagus
Head and Neck
Kidney
Liver
Lung
Ovary
Prostate
Stomach
Uterus
10
100
1000
Data Portal Summary
Data Release 1.0 October 06, 2020
28 Collections
3,652 Cases
1.04 TB Data Volume
51,247 Image Series
Site Home | Privacy Policy | Team and Collaborators | Contact Us | Discourse | Accessibility | Disclaimer | FOIA
Imaging Data Commons Data Release Version 1.0 - October 06, 2020
Data hosted by IDC is subject to the TCIA Data Usage License and Citation Requirements
Youtube | GitHub | Twitter
U.S. Department of Health and Human Services | National Institutes of Health | National Cancer Institute | USA.gov
NIH... Turning Discovery Into Health®

Search Configuration

□ Hide attribute values with 0 cases

ORIGINAL DERIVED RELATED

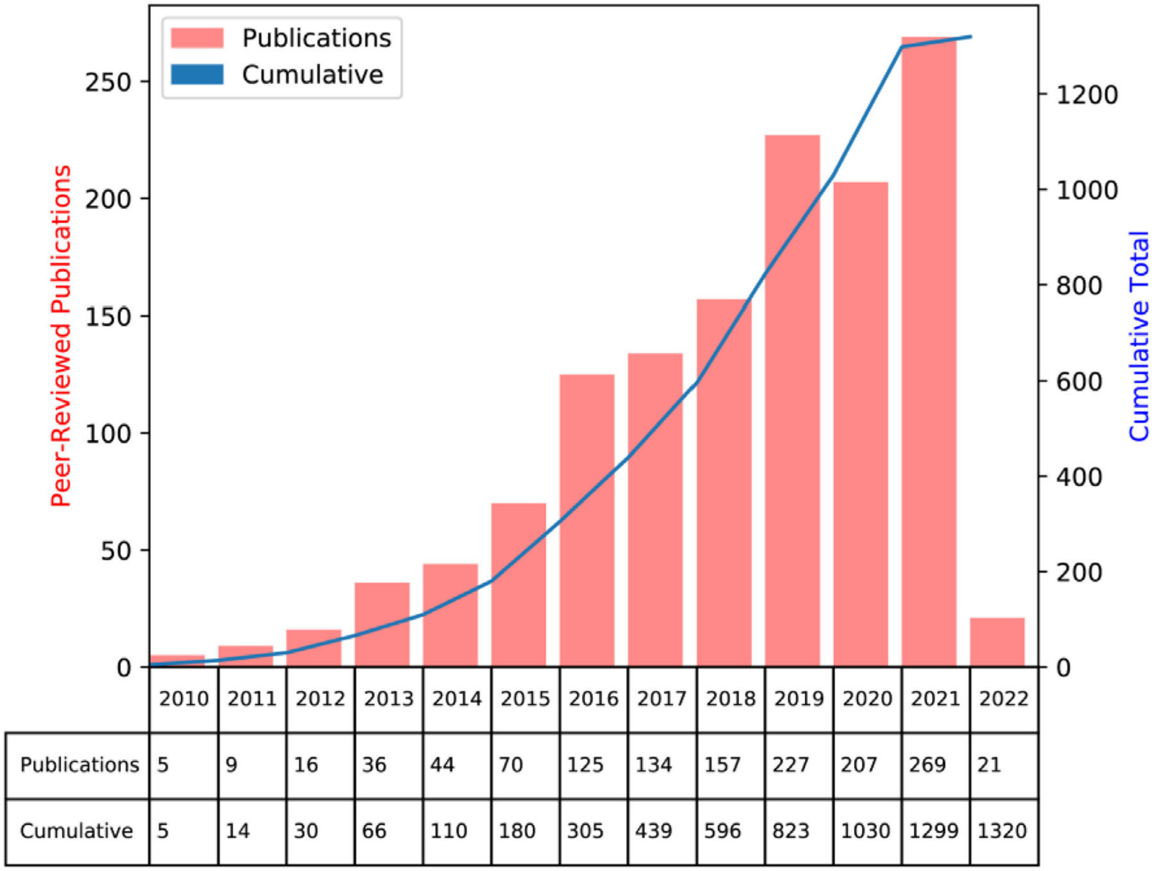
- Segmentation
  - Anatomic Region
  - Segmentation Category
  - Segmentation Type

□ Breast	207
□ Esophagus	355
□ Heart	127
□ Lung	411
□ Mass	10
□ Neoplasm; Primary (SCT)	421
□ Neoplasm; Primary (SRT)	59
□ Neoplasm; Secondary	53
□ Nodule	875
□ None	0
□ Reference Region	59
□ Spinal cord	411

show less Check All / Uncheck All

Figure 4.

Example of imaging data commons (IDC) portal user interface. Reprinted from Fedorov et al.[62]



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Publications	5	9	16	36	44	70	125	134	157	227	207	269	21
Cumulative	5	14	30	66	110	180	305	439	596	823	1030	1299	1320

Figure 5.

Publications over time related to the cancer imaging archive (TCIA) databases. Graph generated from TCIA website (https://www.cancerimagingarchive.net/publications/) on February 18, 2022.

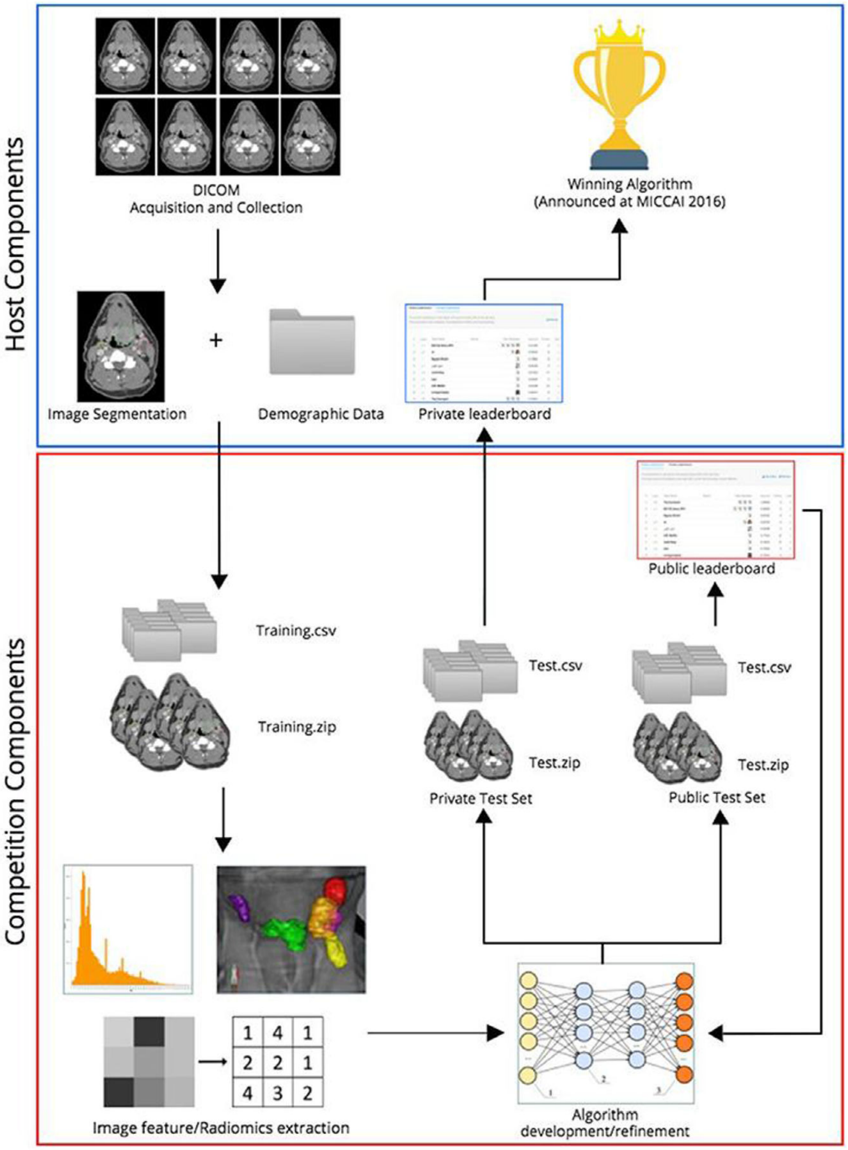


Figure 6.

Overview of 2016 oropharynx cancer (OPC) radiomics challenge. Reprinted from Elhalawani et al.[70]

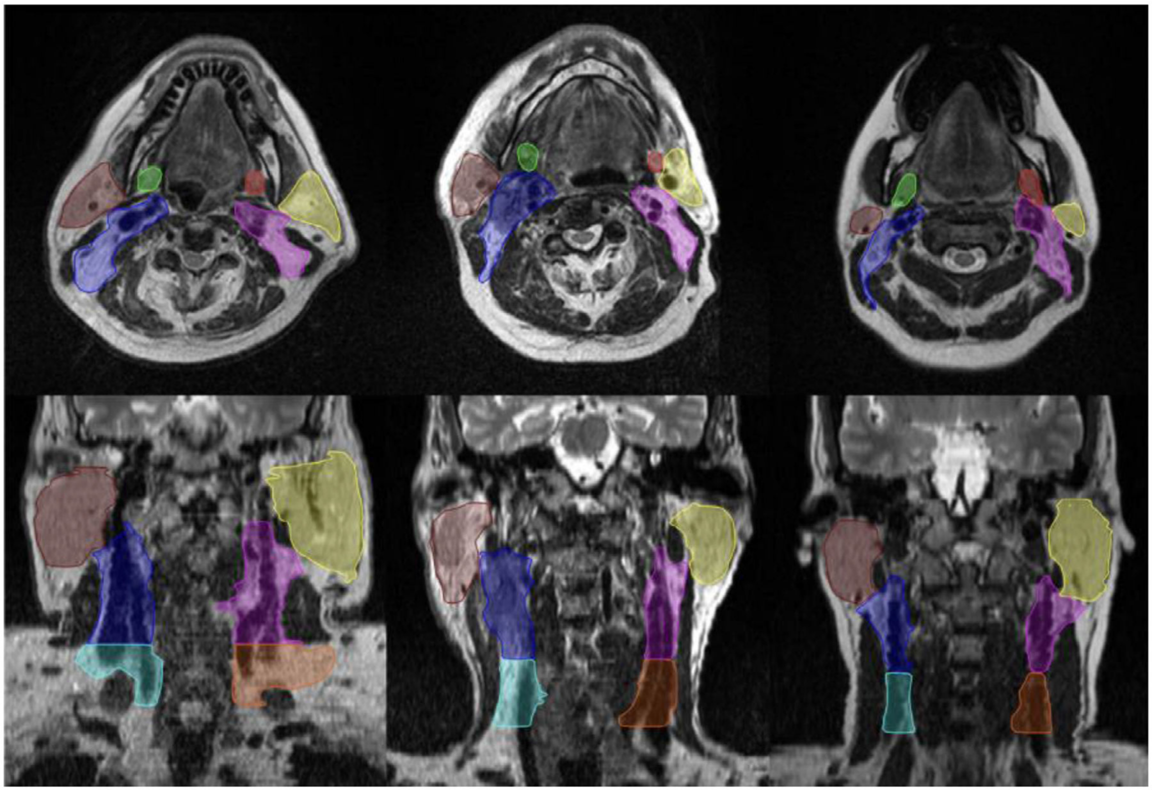
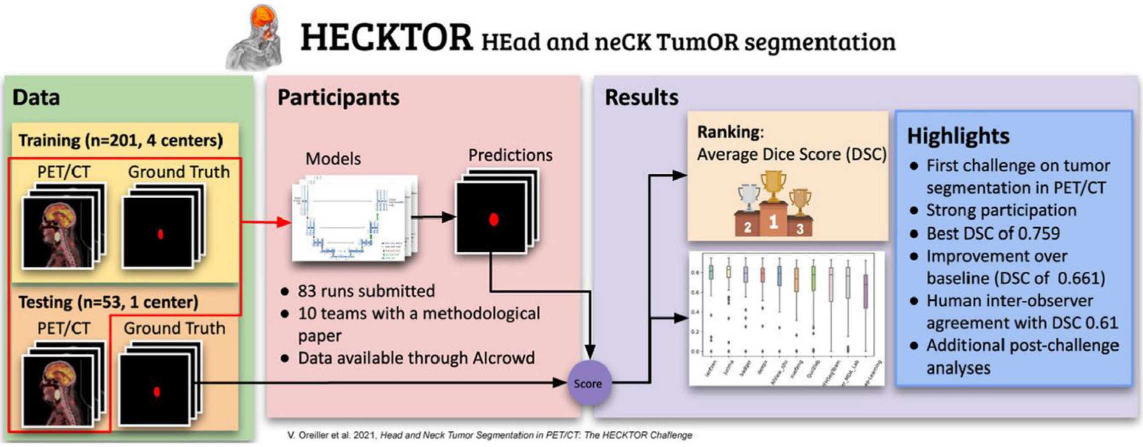


Figure 7.

Contoured structure [left submandibular gland (red), right submandibular gland (green), left parotid gland (yellow), right parotid gland (brown), left lymph node level II (blue), right lymph node level II (pink), left lymph node level III (orange), right lymph node level III (light-blue)] for 2019 “RT-MAC” challenge. Reprinted from Cardenas et al.[73]



- 83 runs submitted
- 10 teams with a methodological paper
- Data available through AIcrowd

- First challenge on tumor segmentation in PET/CT
- Strong participation
- Best DSC of 0.759
- Improvement over baseline (DSC of 0.661)
- Human inter-observer agreement with DSC 0.61
- Additional post-challenge analyses

V. Oreiller et al. 2021, *Head and Neck Tumor Segmentation in PET/CT: The HECKTOR Challenge*

Figure 8.

Overview of 2020 HEad and neCK TumOR (HECKTOR) challenge. Reprinted from Oreiller et al.[75]

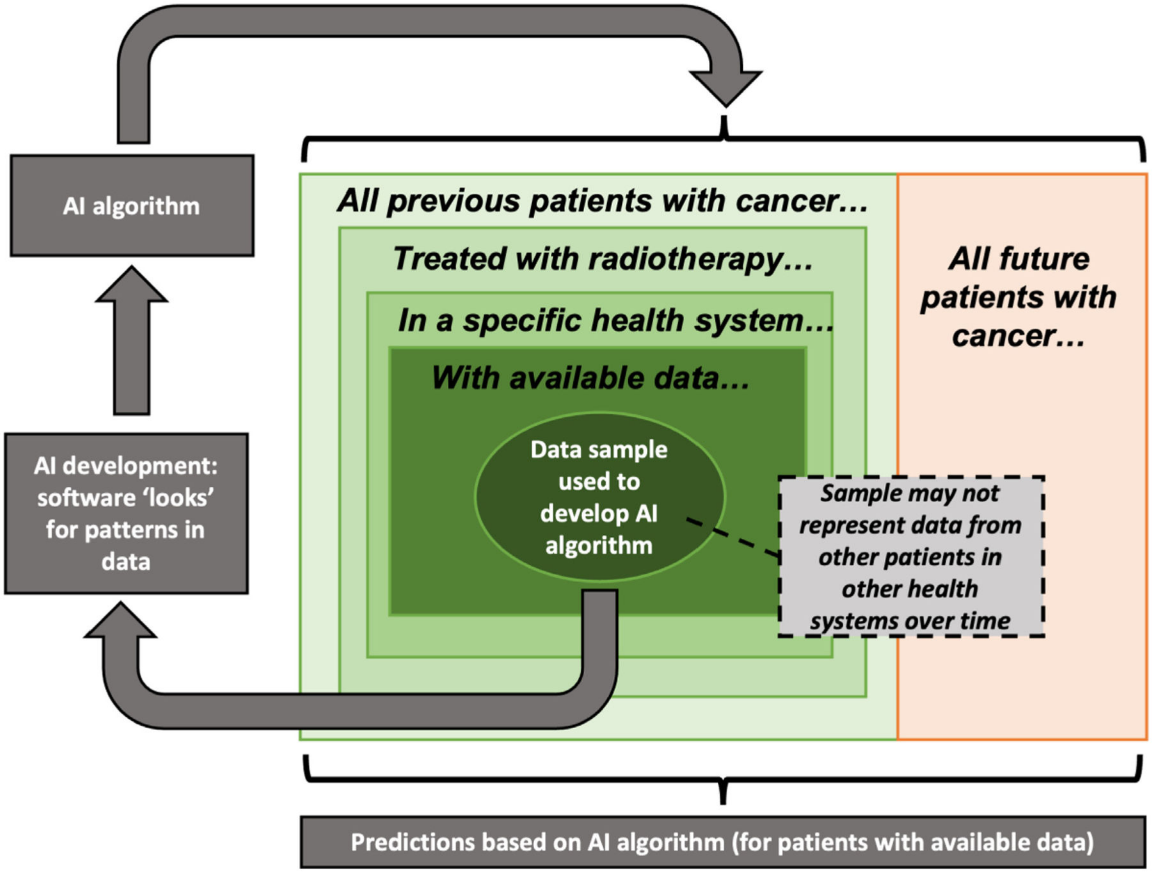

AI algorithm
AI development: software 'looks' for patterns in data
All previous patients with cancer…
Treated with radiotherapy…
In a specific health system…
With available data…
Data sample used to develop AI algorithm
All future patients with cancer…
Sample may not represent data from other patients in other health systems over time
Predictions based on AI algorithm (for patients with available data)

Figure 9.

Statistical biases associated with AI predictions. Adapted from Chua et al.[85]

Table 1

Example of Variations in Standardized Nomenclatures Reported for Non-Target Structures by 16 Institutions

Structure	Number of Institutions	Examples
Left optic nerve	12	Lt Optic Nerve, OPTICN_L, OPTNRV_L, optic_nrv_l, L_optic_nerve, OPTIC_NRV_L, OpticNerve_L, LOPTIC, OpticNerve_L, Lef Optic Nerve, ON_L
Left lung	12	Lt Lung, Lung_L, LUNG_L, lung_l, L_lung, LLUNG, L Lung
Both lungs	12	Lungs, LUNGs, LUNG_TOTAL, lung_total, combined_lung, LUNG, LUNGS, Lung,BilatLung, Lung_Both
8th Cranial nerve	7	CN_VIII, cn_viii, CN8, CN_8
Right external iliac artery	2	A_ILIAC_E_R, a_iliac_e_r

Data adapted from American Association of Physicists in Medicine Task Group 263 Report.[33]

Table 2

HIPAA Data Elements That Encompass the “Safe Harbor” Method

Data Element	Description
Name	-
Address	All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial 3 digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: **1** The geographic unit formed by combining all ZIP codes with the same 3 initial digits contains more than 20,000 people; and **2** The initial 3 digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.
Dates related to individual	birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
Telephone numbers	-
Fax numbers	-
E-mail address	-
Social security number	-
Medical record number	-
Health plan beneficiary number	-
Account number	-
Certificate or license number	-
Vehicle/other device serial number	That is, license plate numbers
Device identifiers/serial numbers	-
Universal resource locators	That is, web URLs
Internet Protocol address	-
Biometric identifiers	Including finger- and voice- prints
Photographic image	Full-face photographs and any comparable images
Misc.	Any other characteristic that could uniquely identify the individual

Table 3

Current National Institute of Health (NIH)-Supported Domain-Specific Data Repositories Related to Radiation Oncology Research

Repository	Description
The cancer imaging archive (TCIA)	TCIA is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. The data are organized as “Collections”, typically patients related by a common disease (e.g., lung cancer), image modality (MRI, CT, etc.) or research focus. DICOM is the primary file format used by TCIA for image storage. Supporting data related to the images such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available.
Metabolomics work-bench (MetWB)	The Metabolomics Program’s Data Repository and Coordinating Center (DRCC), housed at the San Diego Supercomputer Center (SDSC), University of California, San Diego, has developed the Metabolomics Workbench. MetWB will serve as a national and international repository for metabolomics data and metadata and will provide analysis tools and access to metabolite standards, protocols, tutorials, training, and more.
Cancer nanotechnology laboratory (caNanoLab)	caNanoLab is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in bio-medicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical, in vitro, and in vivo assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.
Genomic data commons (GDC)	The mission of the GDC is to provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. The GDC contains clinical, biospecimen, and molecular data from several cancer research programs.
Proteomic data commons (PDC)	The Proteomic Data Commons hosts mass spectra and process data from cancer proteomic experiments. Many datasets have corresponding genomic and/or imaging data available in other nodes of the Cancer Research Data Commons.
The pediatric genomic data inventory (PGDI)	PGDI is an open-access resource for identifying and locating genomic datasets that can be used to further the understanding of childhood cancers and develop better treatment protocols for sick children. This resource lists ongoing and completed molecular characterization projects of pediatric cancer cohorts from the United States and other countries, along with some basic details and reference metadata. To contribute data to the PGDI, researchers need to create a secure submitter account through the PGDI Contributor Application at https://ocg.cancer.gov/programs/target/pgdi/contributor-application. The PGDI catalog will be continually updated as new information is deposited by the research community.
FaceBase	FaceBase is a NIDCR-funded data hub that hosts variety of data generated through dental, oral, and craniofacial research using model organisms and humans. The data offer spotlights high-throughput genetic, molecular, biological, imaging and computational techniques, as well as the database of 3D Facial Norms, developmental atlases and the Ontology of Craniofacial Development and Malformation (OCDM), Human Genome Analysis Interface (HGAI) and other resources.
ClinicalTrials.gov	ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world.

Repositories and descriptions were derived from NIH Data Sharing Repositories online page (https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html) on February 20, 2022.

Table 4

Currently Published Data Collections on The Cancer Imaging Archive (TCIA) Related to Radiation Oncology Applications

Collection	Location	Subjects	Image Types	Supporting Data
Vestibular-Schwannoma-SEG	Ear	242	MR, RTSTRUCT, RTPLAN, RTDOSE	Image Analyses
HNSCC	Head-Neck	627	CT, PT, MR, RTSTRUCT, RTPLAN, RTDOSE	Clinical, Image Analyses
NSCLC-Cetuximab (RTOG-0617)	Lung	490	CT, RTSTRUCT, RTDOSE, RTPLAN	Clinical
TCGA-HNSC	Head-Neck	227	CT, MR, PT, RTSTRUCT, RTPLAN, RTDOSE, Pathology	Clinical, Genomics
HNSCC-3DCT-RT	Head-Neck	31	CT, RTSTRUCT, RTDOSE	Image Analyses
Head-Neck-PET-CT	Head-Neck	298	PT, CT, RTSTRUCT, RTPLAN, RTDOSE	Clinical, Image Analyses, Software/Source Code
NRG-1308 (RTOG 1308)	Lung	12	CT, RTSTRUCT, RTPLAN, RTDOSE	Image Analyses
Head-Neck Cetuximab (RTOG 0522)	Head-Neck	111	CT, PT, RTSTRUCT, RTPLAN, RTDOSE	Image Analyses

Data were generated from TCIA collections page (https://www.cancerimagingarchive.net/collections/) using RTDOSE or RTPLAN as filtering criteria on February 18, 2022.