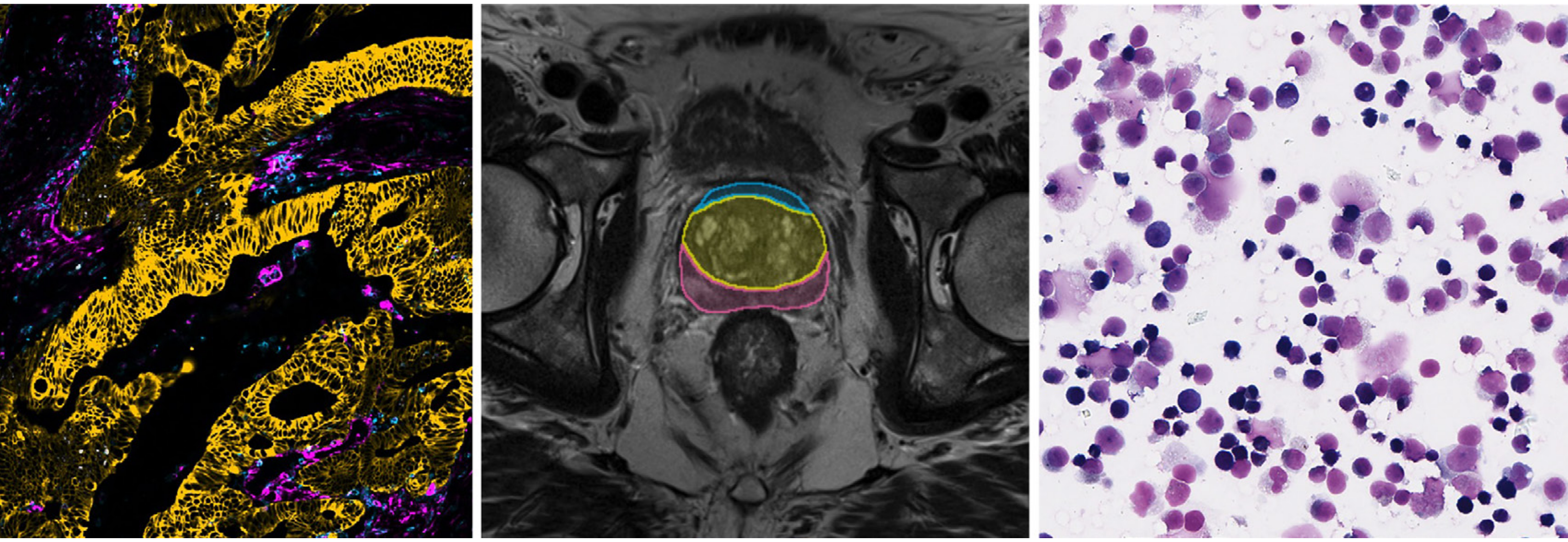
National Cancer Institute Imaging Data Commons: Toward Transparency, Reproducibility, and Scalability in Imaging Artificial Intelligence

Andrey Fedorov, PhD • William J. R. Longabaugh, MS • David Pot, PhD • David A. Clunie, MBBS • Steven D. Pieper, PhD • David L. Gibbs, PhD
Christopher Bridge, DPhil • Markus D. Herrmann, MD, PhD • André Homeyer, PhD • Rob Lewis, BS • Hugo J. W. L. Aerts, PhD
Deepa Krishnaswamy, PhD • Vamsi Krishna Thiriveedhi, MS • Cosmin Ciausu, MS • Daniela P. Schacherer, MS
Dennis Bontempi, MS • Todd Pihl, PhD • Ulrike Wagner, BS • Keyvan Farahani, PhD • Erika Kim, PhD • Ron Kikinis, MD

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

The remarkable advances of artificial intelligence (AI) technology are revolutionizing established approaches to the acquisition, interpretation, and analysis of biomedical imaging data. Development, validation, and continuous refinement of AI tools requires easy access to large high-quality annotated datasets, which are both representative and diverse. The National Cancer Institute (NCI) Imaging Data Commons (IDC) hosts large and diverse publicly available cancer image data collections. By harmonizing all data based on industry standards and colocalizing it with analysis and exploration resources, the IDC aims to facilitate the development, validation, and clinical translation of AI tools and address the well-documented challenges of establishing reproducible and transparent AI processing pipelines. Balanced use of established commercial products with open-source solutions, interconnected by standard interfaces, provides value and performance, while preserving sufficient agility to address the evolving needs of the research community. Emphasis on the development of tools, use cases to demonstrate the utility of uniform data representation, and cloud-based analysis aim to ease adoption and help define best practices. Integration with other data in the broader NCI Cancer Research Data Commons infrastructure opens opportunities for multiomics studies incorporating imaging data to further empower the research community to accelerate breakthroughs in cancer detection, diagnosis, and treatment.



Introduction

The remarkable advances of artificial intelligence (AI) technology are revolutionizing established approaches to the acquisition, interpretation, and analysis of biomedical imaging data. Many promising AI-based tools have been introduced both in the clinic and in the laboratory (1). Development, continuous refinement, and validation of such tools require easy access to large high-quality annotated datasets that are both representative and diverse (2). Establishing such datasets in the field of medical imaging comes with numerous complexities (3). Acquisition of medical images requires highly specialized and complex equipment and personnel. Specialized expertise and significant effort are required to correctly de-identify and curate such data (4). Storage and retrieval of large imaging datasets can present additional challenges, as does orchestrating computation on this scale.

RadioGraphics 2023; 43(12):e230180
https://doi.org/10.1148/rg.230180

Content Codes: AI, IN
Abbreviations: AI = artificial intelligence, CRDC = Cancer Research Data Commons, DICOM = Digital Imaging and Communications in Medicine, FAIR = Findable, Accessible, Interoperable, Reusable, IDC = Imaging Data Commons, NCI = National Cancer Institute, TCIA = The Cancer Imaging Archive

TEACHING POINTS

- Many promising AI-based tools have been introduced both in the clinic and in the laboratory. Development, continuous refinement, and validation of such tools require easy access to large high-quality annotated datasets that are both representative and diverse. Establishing such datasets in the field of medical imaging comes with numerous complexities. Acquisition of medical images requires highly specialized and complex equipment and personnel. Specialized expertise and significant effort are required to correctly de-identify and curate such data.
- As a data commons, IDC enables its users to explore and analyze data and share the generated analysis results.
- The core underpinning of the FAIR Data Principles is in "enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals." To achieve this vision, it is mandatory that metadata follows consistent conventions, which necessitates the use of a standard.
- DICOM representation is rich with metadata (both structured and unstructured) that can enable searching and processing of the images. DICOM enables interoperability, which means that IDC can use off-the-shelf tools (both commercial and open source) implementing the standard, reducing development and maintenance costs, and supporting reuse of the analysis workflow components.
- IDC is an established "home" for findable, accessible, interoperable, and reusable cancer imaging data within the national cancer data ecosystem. IDC is continuously evolving with the goal to better meet the needs of a broad community. Concerted focus on the conversion of images and image-derived data into DICOM representation empowers data exploration and enables interoperability. Balanced use of established commercial products with open-source solutions, interconnected by standard interfaces, allows us to provide value and performance, while preserving sufficient agility to address the evolving needs of the research community.

The National Cancer Institute (NCI), part of the U.S. National Institutes of Health, has invested significant resources into the collection of large amounts of health-related data, including imaging (5–8). Efforts were undertaken to support de-identification, curation, and access to the imaging data with the introduction of The Cancer Imaging Archive (TCIA) about 10 years ago (9). With the nascent effort to establish the national cancer data ecosystem, as one of the priorities for the Cancer Moonshot (10), the emphasis is shifting beyond supporting data archival and access and toward enabling the collaborative use and analysis of these datasets within data commons (11). A core component of the ecosystem is the NCI Cancer Research Data Commons (CRDC)—a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data, along with the analytical and visualization tools for data analysis across domains.

NCI Imaging Data Commons (IDC) (12) *(https://imaging.datacommons.cancer.gov/)* is a component of the CRDC infrastructure that hosts publicly available cancer imaging data colocated with analysis and exploration resources. Since the initial release in 2020, the platform has been evolving, expanding both the data offering and the capabilities supporting data use. Today, IDC is an established imaging data science platform, providing de-identified Digital Imaging and Communications in Medicine (DICOM) imaging data and metadata, analysis results collections, capabilities for viewing, cohort selection, and downstream AI and machine learning (ML) development and analysis using customized or cloud-native tools (13). IDC collections are publicly available, versioned and harmonized into DICOM representation, to meet Findable, Accessible, Interoperable, Reusable (FAIR) principles (14).

IDC is uniquely positioned to help improve transparency, reproducibility, and scalability of the emerging AI analysis tools in biomedical imaging. Lack of transparency (the availability of details accompanying the analysis to enable scientific understanding of how the analysis was performed [15]) and limited reproducibility (the ability to replicate the analysis given the same input data [16]) of AI imaging analysis workflows is broadly recognized as a major obstacle to clinical translation (15). Scalability characterizes the system as being capable of efficiently supporting increasing load sizes. Given both the growing computational complexity of the modern AI algorithms and the sizes of imaging datasets, scalability becomes a critical attribute to enable evaluation and application of imaging AI advances in practice.

In this article, we summarize the key implementation principles of IDC, highlight its current features and capabilities, and discuss major developments and updates since its initial release. As a data commons, IDC enables its users to explore and analyze data and share the generated analysis results. To illustrate this, we highlight some of the recent applications and projects utilizing IDC, with the emphasis on how IDC is positioned to enable its users to improve transparency, reproducibility, and scalability of their analyses.

Overview of NCI IDC

As of the writing this article, the IDC contains over 67 TB (including prior versions) of imaging data spanning a range of image acquisition techniques (eg, radiology imaging modalities, digital pathology, and multiplexed fluorescence imaging) and devices, cancer and tissue types, and organ systems. There are several aspects of the IDC that differentiate it from other repositories.

DICOM for Data Harmonization

The core underpinning of the FAIR Data Principles is in "enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals" (14). To achieve this vision, it is mandatory that metadata follows consistent conventions, which necessitates the use of a standard. All of the images and image-derived data (ie, annotations, segmentations of the regions of interest, image-derived features, analysis results) hosted by IDC are natively encoded as, or harmonized into, DICOM representations (13). In situations where data are supplied in a research format or vendor-specific representation, conversion of the data into DICOM is done by the IDC team. While DICOM was originally developed to support clinical workflows focused on radiology, it has demonstrated utility to support other imaging types and applications (17,18) by enabling interoperability and providing a consistent data model and metadata conventions. DICOM representation is rich with metadata (both structured and unstructured) that can enable searching and processing of the images. DICOM enables interoperability, which means that IDC can use off-the-shelf tools (both commercial and open source) implementing the standard, reducing development and maintenance costs, and supporting reuse of the analysis workflow components. There is no alternative standard for harmonizing representation of pixel data and metadata that can address the breadth of use cases in medical imaging.

Cloud-based Hosting of Data

IDC data are hosted using public cloud services to streamline exploration, search, and analysis of the data (as we discuss in the example use cases later in this article), while leveraging flexibility of the cloud to enable security and scalability. IDC relies on the services provided by both Google Cloud Platform (GCP; Google) and Amazon Web Services (AWS; Amazon). While hosting IDC data in the cloud makes it easier to use cloud-based tools, users can also use their own computational resources for analyzing IDC data.

Public Availability of All Data

All of the data hosted by IDC are available publicly and can be downloaded for either cloud-based or on-premises analysis. Most of the collections in IDC are governed by nonrestrictive licenses allowing use of the images in research and in development of medical products. A small number of collections in IDC limit use to noncommercial activities.

Data Versioning

As the content of individual collections evolves, IDC provides persistent access to the prior versions of each file. Data can be removed from IDC under rare and exceptional circumstances, such as the retraction of the dataset, for example, if protected health information (PHI) is discovered in the data.

Balance of Open-Source and Commercial Components

IDC is implemented using a combination of open-source and commercial tools. Commercial offerings from the leading cloud providers are used to enable scalability, competitive pricing for the access to cloud-based resources, and manageable operational costs. Open-source components enable support of the evolving needs of the cancer imaging research community.

Data Acceptance Criteria

To deposit a dataset to IDC, the submitters must establish its quality and scientific value (eg, by demonstrating the collection is supported by a funded initiative, or it is accompanied by a peer-reviewed article). Data must be de-identified before depositing to IDC. Unless the data are de-identified by an entity that is approved by NCI Security, the submitter must complete risk mitigation documentation describing de-identification approaches and procedures to follow in case PHI is discovered in the data. Contributors must be comfortable with the public (as opposed to restricted) release of the dataset. Finally, the dataset must be released under a permissive license. Most of the data in IDC are covered by the Creative Commons By Attribution license, which permits commercial use of the data. Resources of the IDC team to harmonize incoming data are always going to be limited. Assuming these criteria are satisfied, prioritization of a specific dataset for ingestion is determined by the IDC stakeholders. We expect this procedure to be refined to better address the evolving needs of the community in implementing the recently introduced National Institutes of Health Data Management and Sharing Policy.

Content, Capabilities, and Intended Users

Content

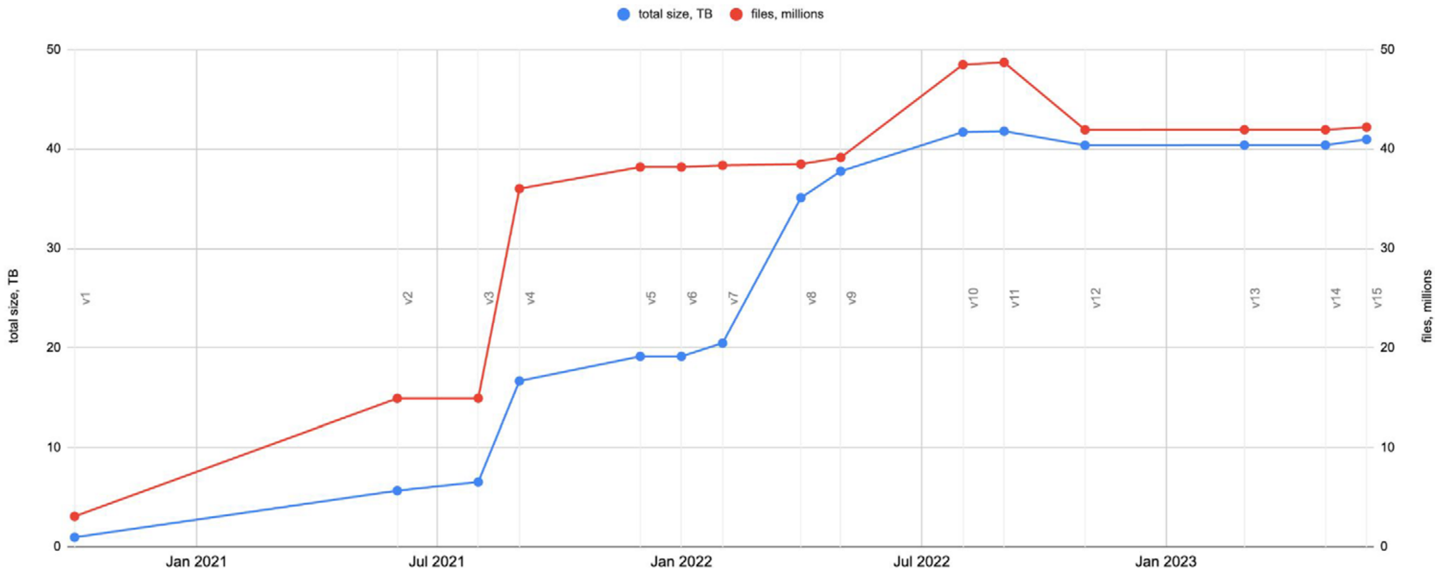
Initially IDC focused on ingesting public DICOM radiology collections already de-identified and curated by TCIA (9). TCIA continues to be an active partner of IDC: data contributors are encouraged to submit images to TCIA with a well-defined pathway to making the data available in IDC. Beyond these initial radiology collections from TCIA, IDC proceeded to ingest digital pathology components of the data collected by The Cancer Genome Atlas (TCGA), The Clinical Proteomic Tumor Analysis Consortium (CPTAC) (6), and the National Lung Screening Trial (NLST) (8), some of which were also hosted in proprietary formats by TCIA. Harmonization of the initial release of the imaging data collected by the Human Tumor Atlas Network (HTAN) (7), including multichannel fluorescence images, was the next important milestone. Importantly, the tools and procedures used by the IDC team to perform conversion into DICOM representation are documented and publicly available, paving the path for broader adoption of DICOM.

In most applications, meaningful interpretation and analysis of images can be challenging without the data describing the clinical characteristics of the patient. Such information is often stored in nonstandardized attachments that accompany the images. To better meet the FAIR principles for the accompanying clinical data, we implemented curation procedures that parse such attachments and ingest clinical metadata into IDC, along with the dictionaries describing their content (when available). All of the clinical data in IDC is searchable using the Standard Query Language (SQL) interface and is linked with the imaging data via the patient or case identifiers.

As of data release version 15 (June 2023; the annotated timeline of releases is shown in Fig 1), IDC contains most of the public radiology collections from TCIA and a range of collections that are unique to IDC. Such IDC-specific content includes DICOM-converted digital pathology collections, originally distributed in a vendor-specific format by the TCGA, CPTAC, NLST, and HTAN initiatives, and the National Library of Medicine (NLM) Visible Human Project (20) dataset, which until recently was only available in a proprietary vendor format from NLM. Current content of IDC is summarized in Figure 2, with a sample of highlight images shown in Figure 3.

Capabilities

IDC is an actively maintained data commons that is continuously growing to include new cancer imaging data collections but also is a resource to support interaction with and use of the



v1: pilot release
v2: most of public TCIA radiology collections
v3: initial DICOM pathology collections, pathology image visualization
v4: production release, data versioning, NLST collection is public
v5: Google Public Dataset Program integration
v6: head collections not accessible due to TCIA policy change
v7: CPTAC DICOM digital pathology collections
v8: TCGA DICOM digital pathology collections
v9: support for differential license (pathology/radiology separate licensing)
v10: fluorescence and H&E imaging collections from HTAN v1 release
v11: ingestion of image-related clinical data
v12: removal of head image collections that were made inaccessible in v6
v13: first collection of AI-generated annotations results
v14: Amazon Open Data Sponsorship integration
v15: NLM Visible Human Project dataset

Figure 1. Annotated timeline of IDC data releases and major development milestones. An up-to-date version of the timeline is available in the IDC data release notes documentation page (19). CPTAC = The Clinical Proteomic Tumor Analysis Consortium, H&E = hematoxylin and eosin, HTAN = Human Tumor Atlas Network, NLM = National Library of Medicine, NLST = National Lung Screening Trial, TB = terabyte, TCGA = The Cancer Genome Atlas.

data. We envision the following broad categories of activities and interactions with the data that IDC can enable, as summarized in Figure 4.

Explore.—Conversion of the data into DICOM representation enables basic uniformity of the data model and metadata. This in turn makes it possible to use consistent selection criteria while interacting with the data. Datasets sharing common metadata expressing relationships can be cross-linked into graphs representing, for example, which input images and human annotations were used as input to ML algorithms. The IDC web portal application provides the entry-level interface to enable exploration of IDC data. Users seeking more detailed data can use the SQL interface, which provides complete access to both the collection-level and DICOM metadata describing the files hosted by IDC and to the clinical data accompanying those collections. Images and accompanying annotations can be visualized using hosted instances of Open Health Imaging Foundation (OHIF; *https://ohif.org/*) (radiology images) and Slim *(https://github.com/ImagingDataCommons/slim)* (digital pathology and fluorescence images) viewers (30, 31).

Subset.—IDC enables unambiguous referencing of individual items and cohorts using unique identifiers. IDC data are versioned, and such references will remain valid and will point to the same files, unaffected by the updates to the IDC content. Using either the IDC portal or SQL queries, comprehensive metadata-based selection criteria filters can be used to define cohorts or subsets of data for a specific analysis task. Application of a query filter to a specific version of IDC data can be used to precisely and reproducibly define the list of files corresponding to the selection.

Analyze.—Standard representation of the data and its colocation with the scalable cloud-based computational resources have the potential to lower the barriers for the analysis of IDC data. Data loading and preprocessing workflows can be standardized and leverage existing libraries implementing DICOM support and can be applied to any standard dataset. Today, all major cloud-computing vendors provide integration of Jupyter notebooks (32), under different products, hosted on seamlessly provisioned cloud virtual machines.

Combined with efficient access to the cloud-based image data, such notebook environments can be used to quickly apply existing tools and pipelines to selected IDC datasets and define the workflow for a larger cohort analysis. Highly scalable computational resources for the analysis of those cohorts can be provisioned directly from the cloud providers or by using platforms, such as Terra *(https://terra.bio/)* and Seven Bridges Cancer Genomics Cloud *(https://www.cancergenomicscloud.org/)*, that implement additional layers of abstraction (33, 34).

Share.—The definition of the cohort that was selected from IDC and used in an analysis can be unambiguously recorded and shared to help achieve transparency. Analysis workflows can be shared in a form that will enable recipients to

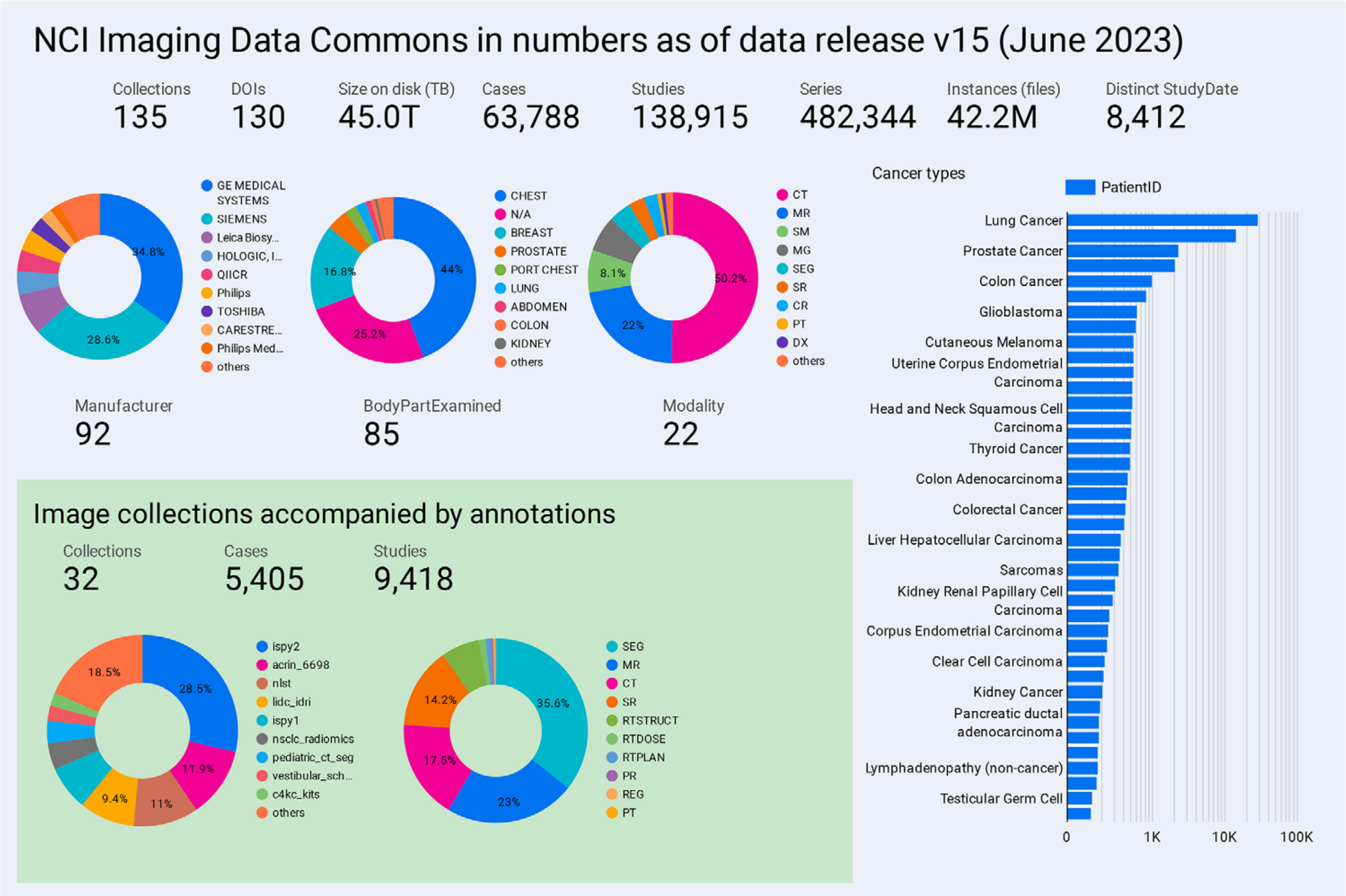


Figure 2. Chart shows a summary of the data available in IDC as of data release version 15 (June 2023). Note that the size on disk reported is in terabytes (TB) (10^{12} bytes). An interactive version of this summary dashboard is publicly available (21).

re-execute the workflow in a reproducible manner, with the cloud environments simplifying the provisioning of the precise environment (both in terms of the virtualized hardware and software required). Visualization of the data used in the analysis and exploration of the accompanying metadata can be accomplished with minimal effort by the recipient using the infrastructure maintained by IDC through simply sharing a web link. The aforementioned resources aim to complement and improve the quality, rigor, accessibility, and reproducibility of the traditional academic publications. Importantly, outputs produced in the process of analyzing IDC data can be harmonized into appropriate DICOM objects and contributed back to IDC, enriching its content and allowing more rapid development of the analysis tools.

Intended Users

We envision the primary group of IDC users to be biomedical computational scientists interested in cancer research who have a technical background in computer science, informatics, or related fields. The content and capabilities of IDC expect the user to have at least some understanding of biomedical imaging. Increasing availability of image-derived features in IDC (eg, shape and intensity texture features characterizing tissue patterns or morphology) aims to make it easier to perform hypothesis exploration and multiomics analysis of IDC data by users with less imaging expertise.

There are many uses of the data IDC contains and IDC infrastructure that could be of interest for the broader audience. Early-stage scientists and students, especially those who are not part of the established groups positioned to have access to large institutional repositories, should be able to identify data that may be relevant to their areas of interest and evaluate existing state-of-the-art analysis tools, prototype, collaborate, and share intermediate findings. Articles presenting novel analysis tools or imaging-based findings can be accompanied by computational notebooks or demonstrations of the developed tools, if not complete containerized analysis workflows. This can be of benefit to academic researchers as well as publishers seeking to make publications more accessible and attractive.

Developers of commercial solutions can evaluate proprietary tools and benchmark them against state-of-the-art open-source solutions or use the data in demonstrations and pilot projects. Funders can consider recommending the use of IDC as a persistent repository holding images and analysis results for broader dissemination. IDC can also simplify the task of evaluating continuously evolving AI tools for practicing radiologists with interest in imaging research. Benchmarking of such

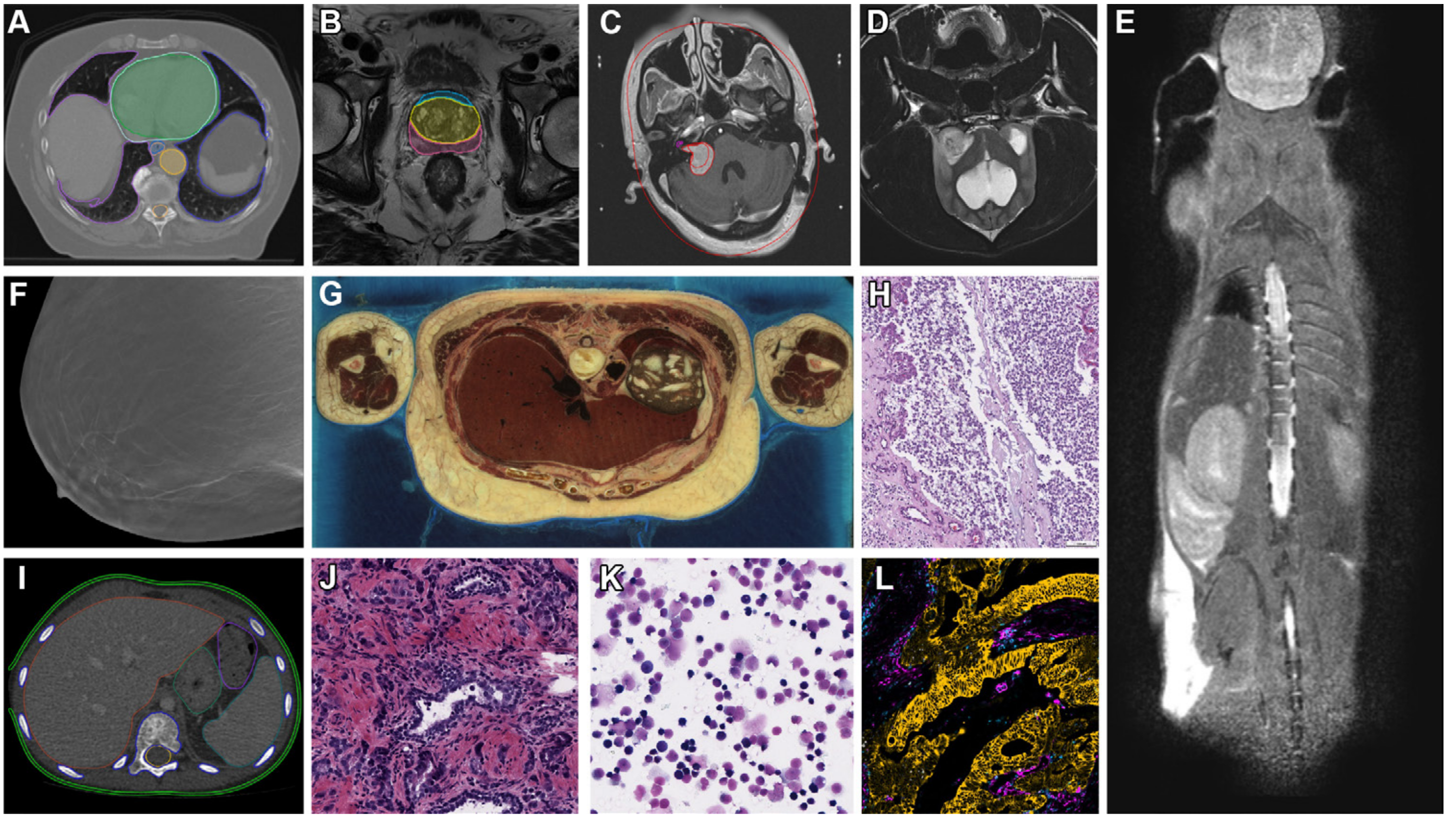


Figure 3. Representative images and image annotations available in IDC. **(A)** NSCLC-Radiomics (22) CT image shows lung cancer with manually annotated regions of interest and nnU-Net-BPR-Annotations (23) (AI-annotated regions of interest). *NSCLC* = non–small cell lung cancer. **(B)** PROSTATEx (24) MR image shows PROSTATEx-Seg-Zones (25), expert-annotated prostate anatomy zones. *Seg* = segmentation. **(C)** Vestibular-Schwannoma-SEG (26) MR image shows schwannoma with manually annotated regions of interest. **(D)** ICDC-Glioma (27) MR image shows canine glioma. *ICDC* = Integrated Canine Data Commons. **(E)** PDMR-997537-175-T MR image shows a mouse adenocarcinoma colon xenograft. *PDMR* = Patient-Derived Models Repository. **(F)** Breast-Cancer-Screening-DBT (28) tomosynthesis image. *DBT* = digital breast tomosynthesis. **(G)** NLM-Visible-Human-Project (20) cryomacrotome anatomic image. **(H)** ICDC-Glioma (27) canine hematoxylin and eosin (H-E) stain digital pathology photomicrograph. **(I)** Pediatric-CT-SEG (29) pediatric CT image with expert-annotated organ contours. **(J)** TCGA-PRAD (5) H-E stain digital pathology photomicrograph shows prostate cancer. *TCGA-PRAD* = The Cancer Genome Atlas Prostate Adenocarcinoma. **(K)** CPTAC-AML (6) H-E stain digital pathology photomicrograph shows acute myeloid leukemia. *CPTAC-AML* = Clinical Proteomic Tumor Analysis Consortium Acute Myeloid Leukemia. **(L)** HTAN-HMS (7) multichannel fluorescence image with pan-cytokeratin, CD45, vimentin, and Ki67 channels selected. *HTAN* = Human Tumor Atlas Network.

tools made readily available by using the cloud resources against public datasets can streamline selection of robust tools before their further evaluation on the internal datasets. Availability of annotations and ongoing work to enrich existing collections with AI-derived annotations, measurements, and features can be of further interest to clinical users, including radiologists, pathologists, and other specialists.

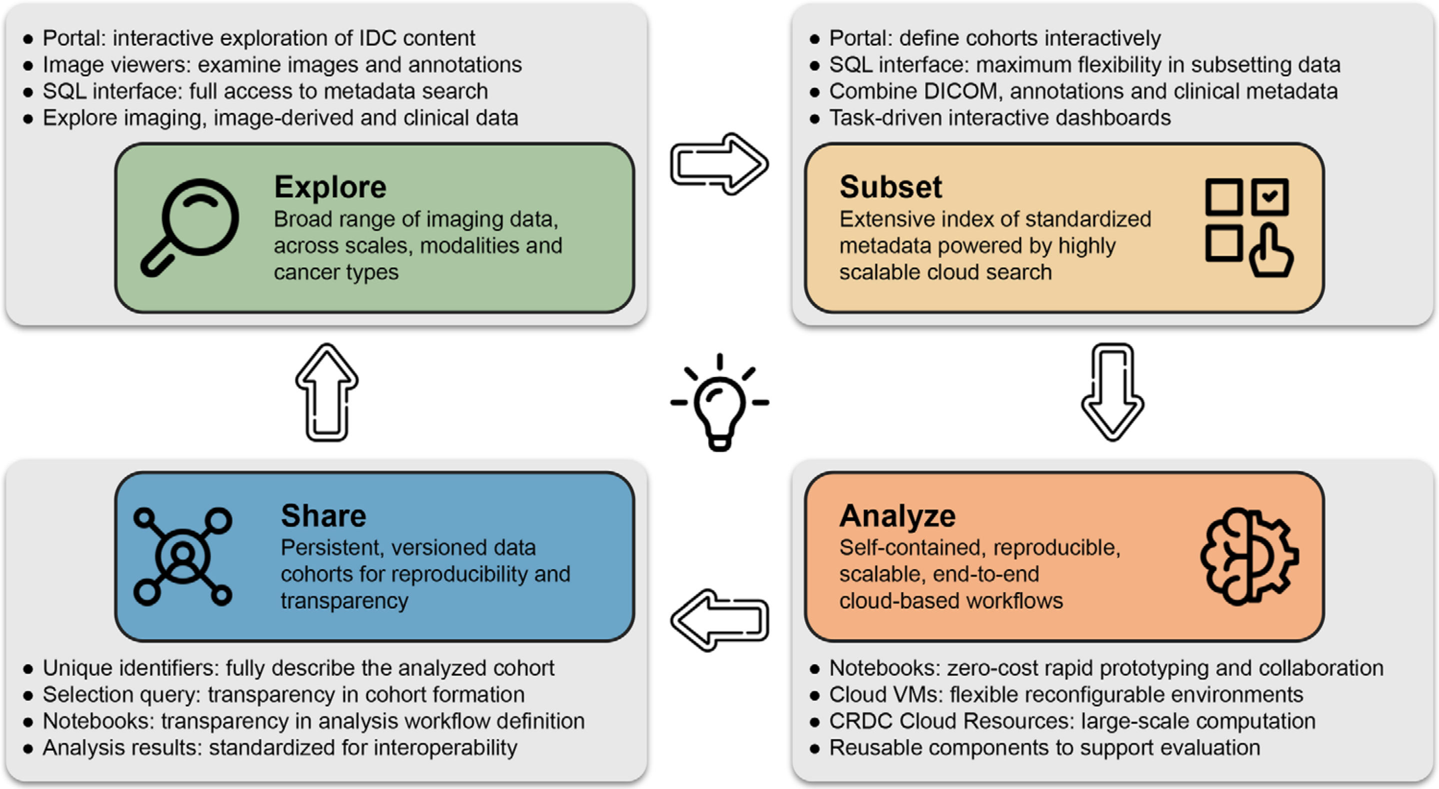
Use Cases

In this section, we illustrate the capabilities of IDC discussed previously with their application to address specific needs in the context of biomedical imaging research, while enabling transparency, reproducibility, and scalability of the analyses. While those activities are somewhat interrelated, we discuss them to draw the attention of different communities of cancer researchers and technology developers.

Best Practices for Data Provenance in Research Reports

Insufficient or unknown provenance of the datasets used while developing ML imaging tools plagues many, if not most, research studies (35), in turn jeopardizing transparency of those studies and reducing their reproducibility. Even with the best intent, data provenance reporting is fraught with pitfalls in the absence of an easily accessible, machine-readable, and detailed description of the dataset.

IDC offers a practical means to address the FAIR principles while describing a dataset used in training or benchmarking an analysis tool. Every single file in IDC is assigned a unique and persistent identifier, with the revisions of the file assigned new identifiers and tracked by IDC versioning. Accompanying metadata is available in a standard representation and is searchable by using a standardized communication protocol, allowing anyone to assess its heterogeneity and the presence of biases. The data and metadata are maintained as close to the source representation as possible. Availability of such metadata can reduce the effort and increase the transparency of data curation, allowing, for example, identification of images that are not suitable for use with a specific AI model without downloading the data (Fig 5). When converting into a DICOM representation, accompanying metadata is harmonized and, where possible, enriched by including additional attributes describing acquisition, as an example. A dataset defined as a list of IDC unique identifiers,



- Portal: interactive exploration of IDC content
- Image viewers: examine images and annotations
- SQL interface: full access to metadata search
- Explore imaging, image-derived and clinical data

Explore
Broad range of imaging data, across scales, modalities and cancer types

- Portal: define cohorts interactively
- SQL interface: maximum flexibility in subsetting data
- Combine DICOM, annotations and clinical metadata
- Task-driven interactive dashboards

Subset
Extensive index of standardized metadata powered by highly scalable cloud search

Analyze
Self-contained, reproducible, scalable, end-to-end cloud-based workflows

- Notebooks: zero-cost rapid prototyping and collaboration
- Cloud VMs: flexible reconfigurable environments
- CRDC Cloud Resources: large-scale computation
- Reusable components to support evaluation

Share
Persistent, versioned data cohorts for reproducibility and transparency

- Unique identifiers: fully describe the analyzed cohort
- Selection query: transparency in cohort formation
- Notebooks: transparency in analysis workflow definition
- Analysis results: standardized for interoperability

Figure 4. Conceptual summary of the capabilities provided by IDC and flowchart of the interactions of the target user with the platform. Each of the gray panels highlights the specific components available within IDC to support the corresponding capabilities. *VM* = virtual machine.

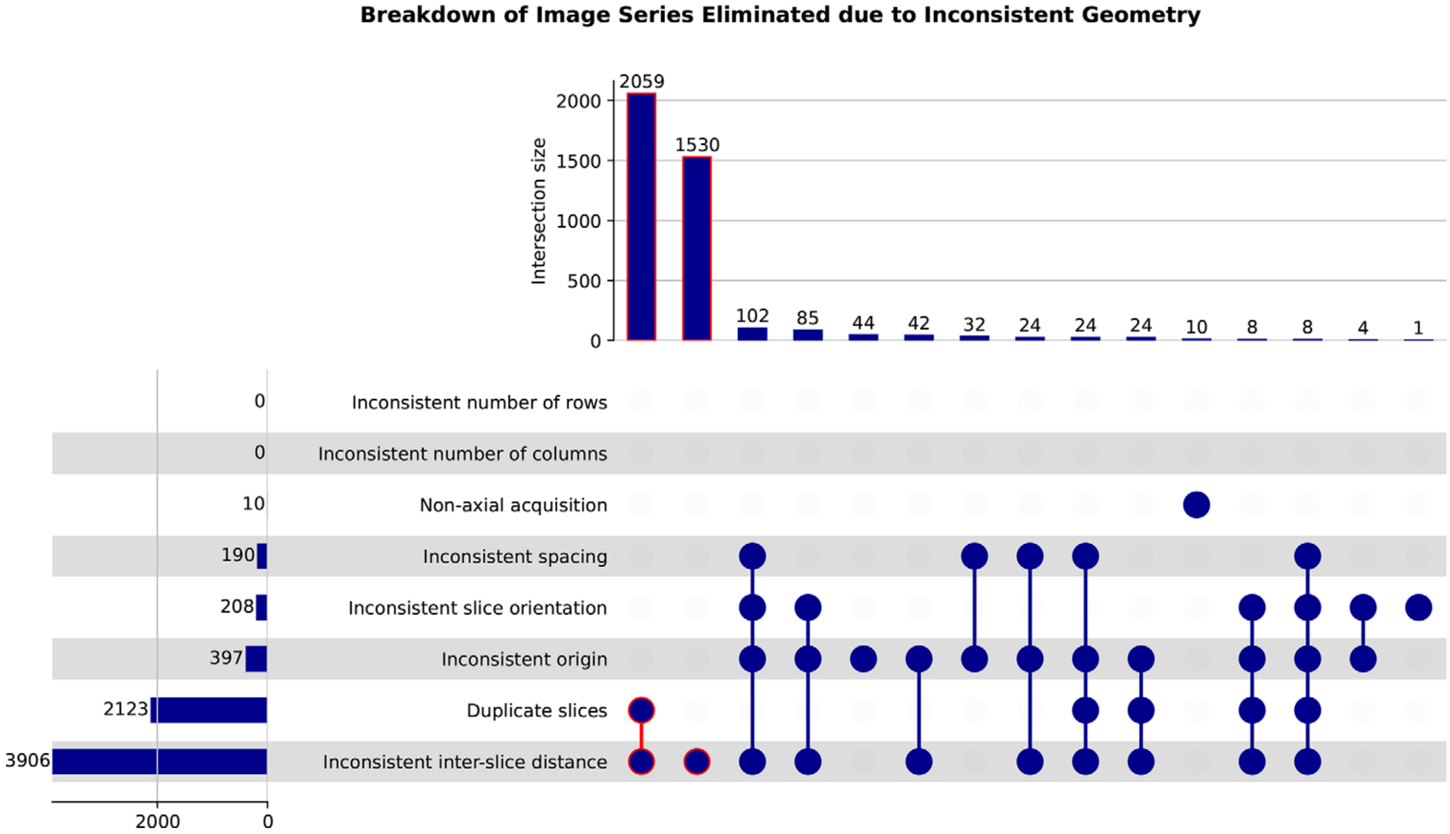


Figure 5. UpSet plots show the distributions of DICOM series from the National Lung Screening Trial (NLST) collection that were identified to have inconsistent geometry. Definition of the rules to identify these series was done by using an SQL statement against the DICOM metadata available in IDC. Items outlined in red correspond to the DICOM series groups constituting the largest portion of those that have inconsistent geometry.

accompanied by the selection query used to produce it, is a persistent solution for dataset reporting.

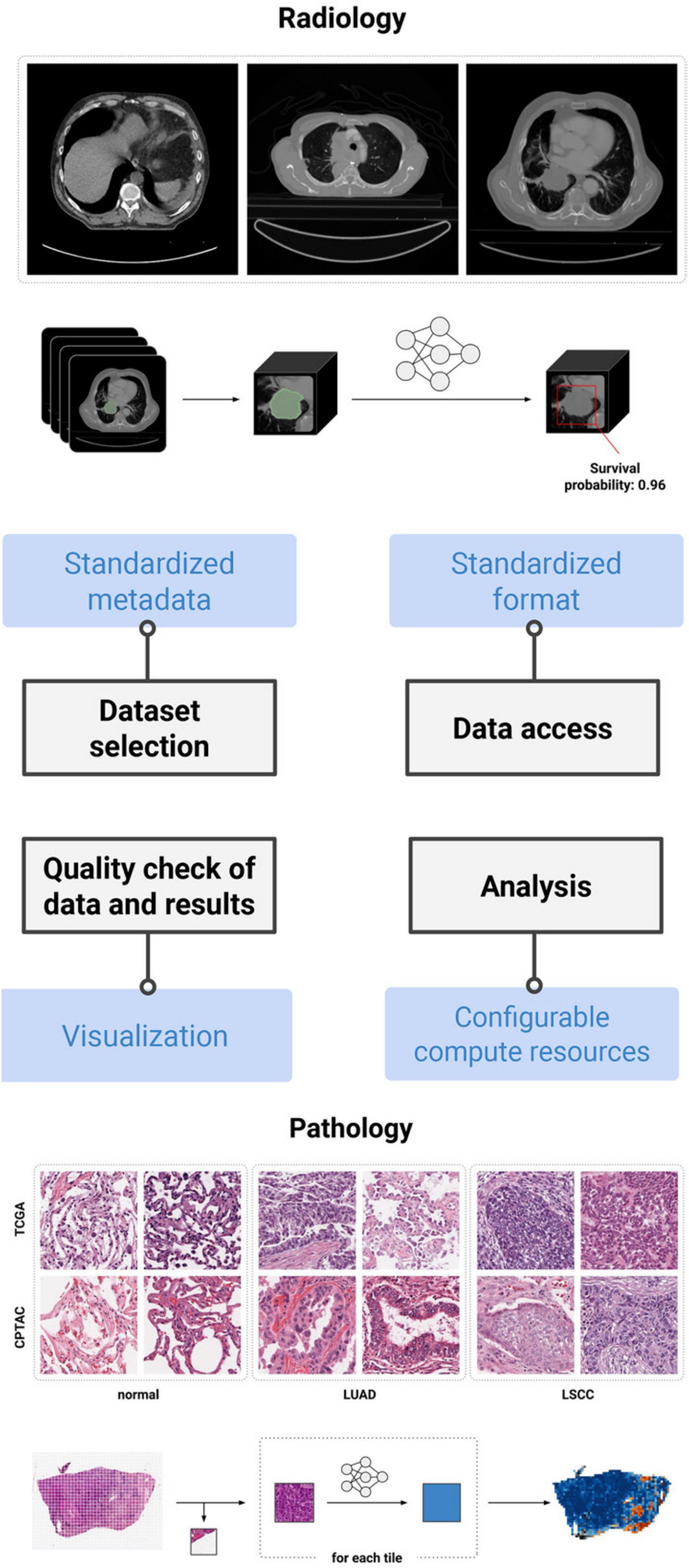


Figure 6. Highlights of IDC features and their contribution to the development of reproducible pipelines. To date, two reproducible analysis studies have been conducted by the IDC team: automated prognosis of lung cancer mortality based on standard-of-care CT (40) and automated classification of lung cancer histopathologic images (39). Both are accompanied by notebooks allowing for reproduction if the study is using the cloud-provisioned compute resources and data available at IDC. *LSCC* = lung squamous cell carcinoma, *LUAD* = lung adenocarcinoma.

Reproducibility of Scientific Studies

Reproducibility, replicability, and repeatability (36–38) prove to be challenging when developing AI tools for medical imaging applications (15). In our experience, IDC and cloud infrastructure can simplify the process of achieving each (39). The hardware configuration and software stack can be described in a manner that allows anyone to instantiate a virtual machine replicating the original deployment environment. When development or evaluation of an analysis method uses public data from IDC, cloud resources combined with versioned code repositories and the aforementioned dataset descriptions can effectively complement the traditional article describing the study.

In situations where analysis involves nonpublic datasets, it may be possible to find a sufficiently similar dataset in IDC that can be used to provide reproducible demonstrations. An example might be to demonstrate the utility of segmentation tools developed on institutional data when applied to IDC public data. Recent studies investigated the utility of IDC in improving reproducibility of specific AI studies in radiology (40) and digital pathology (39) (Fig 6).

Accessibility and Transparency of Image Analysis Tools

We define *accessibility* of an image analysis tool as the ability of a user, with the minimal effort, to apply this tool to a demonstration dataset where the tool is expected to perform well and further to any dataset that meets the input requirements. Accessibility of a tool can be dramatically improved if the article describing the tool and its documentation are accompanied by working examples that incorporate all of the steps needed to deploy and configure the tool, preprocess the data, perform the analysis on a representative dataset, visualize the results, etc. Such minimal working examples are similar to the concept of a *software vignette* first introduced in the Bioconductor project (41).

IDC can serve as the source of readily accessible and searchable public image data that can support such notebooks in medical imaging research. As an example, IDC tutorial materials include a notebook that demonstrates the end-to-end process of performing inference on a chest CT examination using the nnU-Net Task055 SegTHOR (Segmentation of THoracic Organs at Risk) segmentation model (42). Such a minimal example is useful well beyond a mere demonstration of functionality. By making minor adjustments to the selection query, users can experiment with applying the tool to datasets from a variety of collections and manufacturers, gathering hands-on experience with generalizability of the tool. The example operates on the data in the DICOM format, as collected by the imaging equipment used in the clinic, which makes it easier to experiment with the same tool on nonpublic institutional data. This creates opportunities to dramatically improve the scholarly publications by accompanying them with working examples that are relatively easy to create and even easier for readers of the article to reuse.

Enrichment of Public Imaging Datasets

Analysis of medical imaging data and, in particular, integration of imaging data with other sources of biologic data often involves intermediate processing operations, such as segmenting a region of interest or biomarker quantification. Image annotations, which can be pixel-level labeling of the region of interest or image-level assignment of a label, can also be instrumental in making imaging data more searchable and reusable. In such usage scenarios, it is of significant value to have datasets that are accompanied by the annotations and other image-derived items produced by experts and established automated analysis tools. The application of existing tools on public data could be considered by some as straightforward in principle. In practice, access to specialized hardware and domain expertise is required to deploy and configure the tool, address variations in the input data, perform analysis in a time- and cost-efficient manner, conduct quality checks, and document the process and the result.

As described in a recent study (23), IDC can support development of the analysis pipelines aimed at enriching existing datasets (Fig 7). Resulting pipelines can be developed following the examples mentioned in the previous sections to facilitate their transparency and reproducibility. Availability of the resulting annotations along with the images should simplify the use of data by those without imaging expertise (eg, by supporting studies investigating correlation of segmentation-derived features, such as volume, with nonimaging endpoints), support exploration and hypothesis generation, and enable integration of imaging phenotypes with the complementary genomics, proteomics, and other sources of data within CRDC.

Benchmarking of the Analysis Tools

Over the past decade, biomedical image analysis challenges emerged as a mechanism to assess state-of-the-art solutions to relevant computational problems of clinical relevance. Unfortunately, most of the challenges do not require the participants to *publicly* share the easily accessible documented analysis tools or the results produced by those tools in the course of the challenge. Access to the datasets used in the challenges is often restricted to the participants in the challenge and may not be archived for persistent access. These issues limit transparency and reproducibility of the analyses performed using participating tools. Faced with a practical question of understanding and using state-of-the-art tools, the leaderboard of a challenge may be of limited value to their prospective users.

Although, as of today, IDC does not support sequestration of data to enable algorithm challenges, it contains a growing number of annotations to help with the benchmarking efforts applied to either open-source or proprietary analysis tools. To demonstrate the potential of IDC in this area, we have been investigating open-source tools available for segmenting the prostate gland anatomy from MRI. While this topic has been the subject of extensive algorithmic development efforts for decades (44,45), a readily accessible solution to this common preprocessing task remains elusive. Utilizing expert annotations of prostate gland anatomy accompanying several IDC collections, we evaluated two open-source tools: Medical Open Network for AI (MONAI; 46) implementation of the model proposed by Adams et al (47) and nnU-Net (42) that notably showed the best performance in the Medical Imaging Decathlon challenge (48). Implementation of our ongoing benchmarking procedure is publicly available (49), and while it confirms some of the earlier findings, it also identifies limitations and problematic cases for each of the two algorithms, stimulating further refinement of the technology (Fig 8).

Development of Enabling Technology

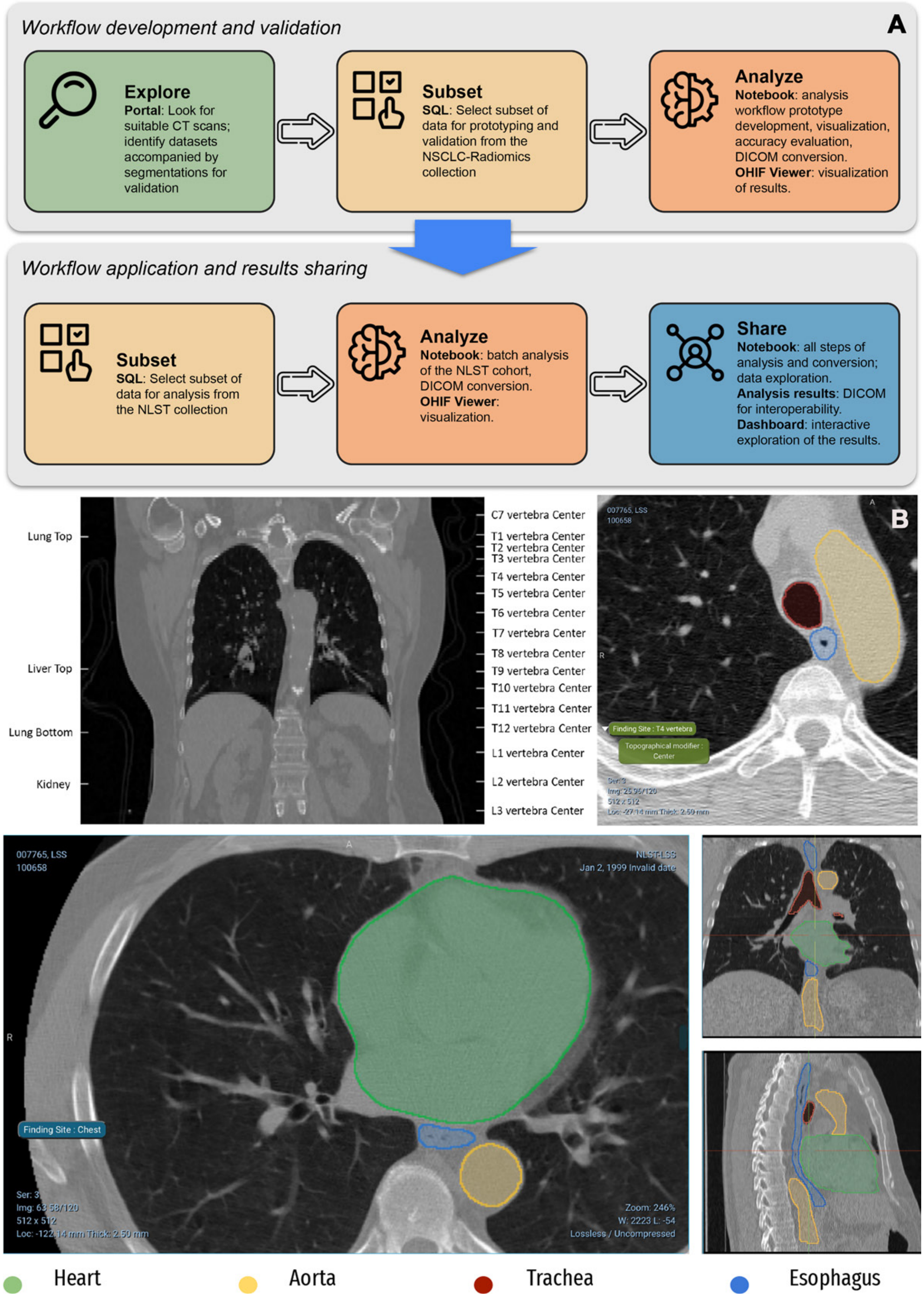
Computational analysis of biomedical images typically involves multiple processing steps, a variety of libraries and tools providing capabilities ranging from the visualization of images to automatic segmentation of anatomic organs, and interoperable communication of the analysis results. Development and continuous refinement of those individual components hinges on the availability of readily accessible samples of data that are representative of the complexity and variety of data encountered in practice. Permissive licenses accompanying data hosted by IDC, along with the standard interfaces to select and access representative samples based on the needs of the specific tool, support the use of IDC data for such tasks—both in the context of academic and commercial activities.

To list a few examples, data available in IDC was instrumental to improve and support the development of the open-source OHIF radiology (30) and Slim microscopy viewers (31); Bio-Formats (*https://www.openmicroscopy.org/bio-formats/*) (51), OpenSlide *(https://openslide.org/)* (52), and highdicom *(https://github.com/ImagingDataCommons/highdicom)* (53) libraries; refinements to the dcm2niix tool *(https://github.com/rordenlab/dcm2niix)*, which converts from DICOM to Neuroimaging Informatics Technology Initiative (NiFTI) format (54); and commercial implementations of the DICOM standard, to name a few relevant efforts. In a recent development, Kulkarni et al (55) investigated the application of large language models (LLMs) to simplify IDC image search using the rich metadata curated by IDC.

Large-Scale Analysis of Biomedical Data

Analysis of biomedical imaging data poses significant computational challenges, primarily due to the sheer size of both the individual images and collective cohorts and graphical processing unit hardware requirements imposed by modern AI tools. Cloud-based solutions have the ability to provide scalable computational resources on demand, with the promise of reduced costs enabled by the "pay only for what you use" model (33,56,57). In practice, adopting cloud-based solutions is not easy; there is limited evidence of proven cost benefit, while there is a potential for budget overruns in the absence of robust cost-control mechanisms.

Within CRDC, to help mitigate the aforementioned challenges, the Broad Institute FireCloud *(https://portal.firecloud.org/)* and Seven Bridges Cancer Genomics Cloud platforms offer additional services complementing the feature set provided by the cloud vendors, simplifying the use of the cloud. Experience using these higher-level platforms for imaging research is currently very limited. It is within the IDC mission to develop a better understanding of the best practices



Workflow development and validation

Explore
Portal: Look for suitable CT scans; identify datasets accompanied by segmentations for validation

Subset
SQL: Select subset of data for prototyping and validation from the NSCLC-Radiomics collection

Analyze
Notebook: analysis workflow prototype development, visualization, accuracy evaluation, DICOM conversion.
OHIF Viewer: visualization of results.

Workflow application and results sharing

Subset
SQL: Select subset of data for analysis from the NLST collection

Analyze
Notebook: batch analysis of the NLST cohort, DICOM conversion.
OHIF Viewer: visualization.

Share
Notebook: all steps of analysis and conversion; data exploration.
Analysis results: DICOM for interoperability.
Dashboard: interactive exploration of the results.

Figure 7. AI-generated annotations of CT images using BodyPartRegression (43) and the nnU-Net Task055 SegTHOR segmentation model (42), with further details on the application of these algorithms to the NSCLC-Radiomics and NLST collections of IDC discussed by Krishnaswamy et al (23). **(A)** Top panel of the flowchart shows the process of development and interaction with the individual components of IDC. The bottom panel of the flowchart shows products generated in the process of analysis. **(B)** Top CT images containing anatomic landmarks corresponding to the centers of vertebra and several other structures (top left) are identified and labeled automatically (top right). Bottom CT images show volumetric segmentations of the pixels corresponding to the heart, esophagus, aorta, and trachea.

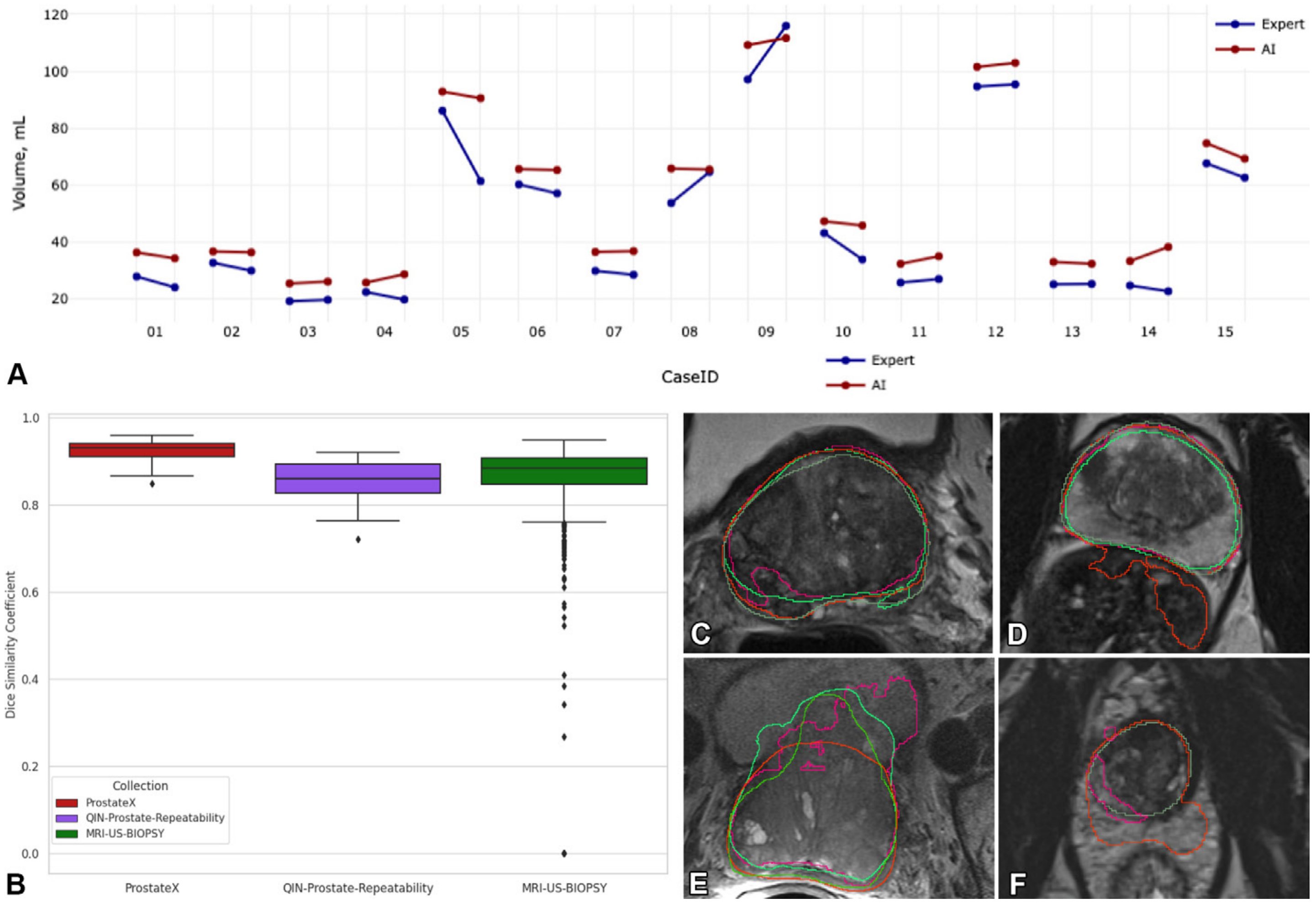


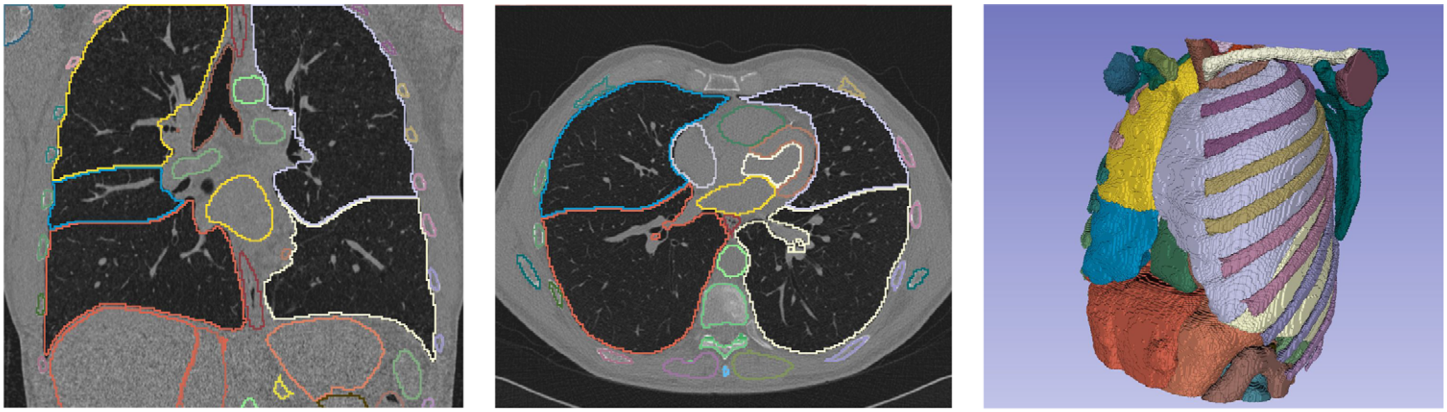
Figure 8. Highlights of an ongoing study evaluating the nnU-Net (42) and Prostate158 (47) algorithms applied to the prostate gland segmentation task. **(A)** Graph shows a comparison of the prostate gland volume calculated from the segmentation produced by AI (red) and the expert (blue) on the QIN-Prostate-Repeatability collection (50) available in IDC. Two measurements corresponding to the same case identification number were derived from the segmentations obtained from the two imaging studies obtained within 2 weeks, precluding biologic changes in the anatomy. Prostate volume is in most cases overestimated by AI as compared with the expert segmentations. **(B)** Dice similarity coefficient (DSC) graph shows that the distributions are visually different across the PROSTATEx, QIN-Prostate-Repeatability, and MRI-US-Biopsy collections of IDC. **(C–F)** Sample MR images from the PROSTATEx **(C, D)**, QIN-Prostate-Repeatability **(E)**, and MRI-US-Biopsy **(F)** collections show segmentation results produced by different AI algorithms and the manual outlines of the prostate (dark green).

for using the cloud for time- and cost-efficient analysis of large cohorts. One such related initiative currently underway is focused on automatic segmentation of more than 100 anatomic structures for the entire NLST cohort using the TotalSegmentator algorithm (48) deployed via CRDC cloud resources. Our initial results demonstrate that the use of these higher-level platforms offers remarkable advantages in terms of processing time at a rather modest cost for large image cohorts, as shown in Figure 9.

Discussion and Future Work

The NCI CRDC aims to implement a holistic approach to data curation, management, and collaborative analysis within the cancer data ecosystem. In turn, IDC should always be considered as a component within CRDC, which is intended to support all of these activities in cancer imaging research beyond serving as a data archive. Dedicated efforts of IDC to harmonize imaging and image-derived data into a uniform DICOM representation, along with the development of the supporting tools to enable such harmonization and the use of resulting data, differentiate IDC from the existing repositories, such as TCIA (9) or Medical Imaging and Data Resource Center (MIDRC) (61). Unlike project-driven repositories, where data collection procedures are formalized and defined ahead of time, such as UK BioBank (62), our goal is to accommodate a growing range of heterogeneous data collection efforts and projects, whether initially well or poorly conditioned. The data stored in IDC are available in DICOM representation as close as possible to that generated by medical imaging devices to support a broad range of uses, which is conceptually different from the approaches implemented in such repositories as Radiopaedia (63) or MedMNIST (64), where data organization and representation are targeting a very specific application.

Ingestion and harmonization of new collections of images and/or annotations and analysis results to grow the IDC data offerings remain a priority for IDC. Refinement of the procedures and development and improvement of open-source



Cohort size, DICOM series	Slices per series, median (inter-quartile range) [min-max]	Total size of the files processed	Virtual Machines used in parallel	Wall clock time	Total compute time	Total cost	Cost per series
1037	157 (39) [101-342]	89 GB	87	4 hrs 54 min	4 days 1 hr 45 min	$6.37	$0.00614
9880	174 (172) [50-801]	1.09 TB	824	3 hrs 52 min	41 days 4 hr 35 min	$71.63	$0.00725
126068	155 (47) [50-801]	10.11 TB	10508	8 hrs 7 min	785 days 13 hr 1 min	$806.20	$0.00640

Figure 9. Summary of the results of a preliminary study evaluating the time- and cost-efficient scalable application of the TotalSegmentator algorithm (58) to the IDC NLST collection using CRDC resources. For each of the analyzed cases in the three cohorts of sizes 1037, 9880, and 126 068 in a CT series, the algorithm was used to segment up to 104 anatomic structures (depending on the coverage of the anatomy in a given imaging examination), followed by the extraction of the shape and first-order radiomics features for each of the segmented regions using the pyradiomics library (59). Coronal and axial CT images (top left and center, respectively) and a surface rendering of the segmentations generated using 3D Slicer *(https://slicer.org)* software (top left) show sample visualizations of the analysis (60). Bottom table summarizes the key parameters and observed performance of the two experiments. The total compute time corresponds to the time needed to perform computation sequentially. In the case of the 126 068 series analysis (red box), scaling of the processing to use 10 508 cloud-based virtual machines in parallel reduced the processing time from the estimated more than 785 days by using a single virtual machine to about 8 hours. The costs are expected to be even lower for the researchers eligible to access the discounts provided by the National Institutes of Health Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative.

tools for conversion of user-submitted nonstandard data into DICOM representation are an active area of work. This is particularly important for submissions of whole slide digital pathology collections, where commercial adoption and deployment of DICOM in acquisition devices is in the relatively early stages (65,66).

Ease of access to the state-of-the-art AI analysis tools, and streamlining the process of applying those to the data in IDC, is another direction of ongoing development. MHub (67) is an emerging NCI-funded repository of self-contained deep learning models pretrained for a wide variety of applications and is being developed in coordination with IDC. Models curated as part of MHub are designed to be DICOM-enabled, which should ease the use of IDC and simplify contribution of the analysis results back to IDC.

There are numerous ongoing directions of work within the IDC project to address known and future challenges. One such challenge in using IDC and CRDC is the lack of experience within the community in utilizing large-scale cloud-computing resources for medical imaging data analysis tasks. Our early experience analyzing the NLST collection (at the moment, summarized only in Figure 9) demonstrates the potential for using the cloud. More work is required to document those large-scale analysis use cases and develop educational materials for the community. IDC data intake and curation currently require significant resources to review de-identification procedures implemented by the submitters, collect the accompanying metadata, and convert the data submitted into standard DICOM representation. It is expected that some of those tasks will be supported by the CRDC Data Hub, a dedicated resource within CRDC to support data submitters. Currently, IDC is limited to hosting only those collections that are available without restrictions. Subject to priorities and resources availability, we are considering adding support for limited access collections to accommodate data sequestration or limited time embargo on data release.

Conclusion

IDC is an established “home” for findable, accessible, interoperable, and reusable (10) cancer imaging data within the national cancer data ecosystem. IDC is continuously evolving with the goal to better meet the needs of a broad community. Concerted focus on the conversion of images and image-derived data into DICOM representation empowers data exploration and enables interoperability. Balanced use of established commercial products with open-source solutions, interconnected by standard interfaces, allows us to provide value and performance, while preserving sufficient agility to address the evolving needs of the research community. Emphasis on the development of tools, use cases to demonstrate the utility of uniform data representation, and cloud-based analysis all aim to ease adoption and help define best practices. Integration with other data within the CRDC opens opportunities for multiomics studies incorporating imaging data to further empower the research community to accelerate breakthroughs in cancer detection, diagnosis, and treatment.

Author affiliations.—From the Department of Radiology, Brigham and Women's Hospital and Harvard Medical School, 399 Revolution Dr, Somerville, MA 02145 (A.F., D.K., V.K.T., C.C., R.K.); Institute for Systems Biology, Seattle, Wash (W.J.R.L., D.L.G.); General Dynamics Information Technology, Rockville, Md (D.P.); PixelMed Publishing, Bangor, Pa (D.A.C.); Isomics, Cambridge, Mass (S.D.P.); Departments of Radiology (C.B.) and Pathology (M.D.H.), Massachusetts General Hospital and Harvard Medical School, Boston, Mass; Fraunhofer MEVIS, Bremen, Germany (A.H., D.P.S.); Radical Imaging, Boston, Mass (R.L.); Artificial Intelligence in Medicine Program, Mass General Brigham, Harvard Medical School, Boston, Mass (H.J.W.L.A., D.B.); Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, the Netherlands (H.J.W.L.A., D.B.); Frederick National Laboratory for Cancer Research, Rockville, Md (T.P., U.W.); and National Cancer Institute, Bethesda, Md (K.F., E.K.). Presented as an education exhibit at the 2022 RSNA Annual Meeting. Received June 28, 2023; revision requested July 28 and received August 10; accepted September 12. **Address correspondence** to A.F. (email: *andrey.fedorov@gmail.com*).

Funding.—Supported by the National Cancer Institute, National Institutes of Health, under Task Order No. HHSN26110071 under Contract No. HHSN261201500003l.

Disclosures of conflicts of interest.—**D.A.C.** Consulting fees from NEMA, HCTS, IBIS, Enlitic, MDDX, Flywheel, maiData, AgMednet, Philips, Brigham and Women's Hospital, University of Leeds, Mahidol University, Washington University in St Louis, and the Mayo Foundation; support for attending meetings from the National Cancer Institute and Brigham and Women's Hospital. **M.D.H.** Support for attending meetings and/or travel from the College of American Pathologists and Digital Pathology Association; patents pending (US Patent Apps 17/083,761, 17/050,435, 17/761,561); leadership or fiduciary role for Digital and Computational Pathology Committee, College of American Pathology Board of Directors, Digital Pathology Association; stock holder in and employed by Roche. **R.L.** Subcontractor on an OHIF grant to Mass General Hospital supporting OHIF development. **H.J.W.L.A.** Consulting fees from Onc.Ai, Sphera, Love Health. **T.P.** Grant and support for attending meetings/travel from the Frederick National Laboratory for Cancer Research. All other authors, the editor, and the reviewers have disclosed no relevant relationships.

References

1. Daye D, Wiggins WF, Lungren MP, et al. Implementation of Clinical Artificial Intelligence in Radiology: Who Decides and How? Radiology 2022;305(3):555–563.
2. Syeda-Mahmood T. Role of Big Data and Machine Learning in Diagnostic Decision Support in Radiology. J Am Coll Radiol 2018;15(3 Pt B):569–576.
3. Holzinger A, Haibe-Kains B, Jurisica I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. Eur J Nucl Med Mol Imaging 2019;46(13):2722–2730.
4. Clunie DA, Flanders A, Taylor A, et al. Report of the Medical Image De-Identification (MIDI) Task Group: Best Practices and Recommendations. arXiv 2303.10473 [preprint] http://arxiv.org/abs/2303.10473. Updated April 1, 2023. Accessed March 21, 2023.
5. Weinstein JN, Collisson EA, Mills GB, et al; Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–1120.
6. Edwards NJ, Oberti M, Thangudu RR, et al. The CPTAC Data Portal: A resource for cancer proteomics research. J Proteome Res 2015;14(6):2707–2713.
7. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, et al; Human Tumor Atlas Network. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell 2020;181(2):236–249.
8. Aberle DR, Berg CD, Black WC, et al; National Lung Screening Trial Research Team. The National Lung Screening Trial: overview and study design. Radiology 2011;258(1):243–253.
9. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26(6):1045–1057.
10. Jaffee EM, Dang CV, Agus DB, et al. Future cancer research priorities in the USA: a Lancet Oncology Commission. Lancet Oncol 2017;18(11):e653–e706.
11. Grossman RL. Ten lessons for data sharing with a data commons. Sci Data 2023;10(1):120.
12. Fedorov A, Longabaugh WJR, Pot D, et al. NCI Imaging Data Commons. Cancer Res 2021;81(16):4188–4193.
13. Bidgood WD Jr, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc 1997;4(3):199–212.
14. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.
15. Haibe-Kains B, Adam GA, Hosny A, et al; Massive Analysis Quality Control (MAQC) Society Board of Directors. Transparency and reproducibility in artificial intelligence. Nature 2020;586(7829):E14–E16.
16. Renard F, Guedria S, Palma N, Vuillerme N. Variability and reproducibility in deep learning for medical image segmentation. Sci Rep 2020;10(1):13724.
17. Herrmann MD, Clunie DA, Fedorov A, et al. Implementing the DICOM standard for digital pathology. J Pathol Inform 2018;9(1):37.
18. Fedorov A, Clunie D, Ulrich E, et al. DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. PeerJ 2016;4:e2057.
19. IDC User Guide. https://learn.canceridc.dev/data/data-release-notes. Last modified September 2023. Accessed August 10, 2023.
20. Ackerman MJ. The Visible Human Project: a resource for anatomical visualization. Stud Health Technol Inform 1998;52(Pt 2):1030–1032.
21. Google PDP Imaging Data Commons Dataset Dashboard. https://lookerstudio.google.com/reporting/04cf5976-4ea0-4fee-a749-8bfd162f2e87/page/p_s7mk6eybqc. Lasted updated October 20, 2023. Accessed August 10, 2023.
22. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006 [Published correction appears in Nat Commun 2014;5:4644.]
23. Krishnaswamy D, Bontempi D, Thiriveedhi V, et al. Enrichment of the NLST and NSCLC-Radiomics computed tomography collections with AI-derived annotations. arXiv 2306.00150 [preprint] http://arxiv.org/abs/2306.00150. Published May 31, 2023. Accessed June 2, 2023.
24. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging 2014;33(5):1083–1092.
25. Meyer A, Rakr M, Schindele D, et al. Towards patient-individual PI-rads v2 sector map: Cnn for automatic segmentation of prostatic zones from T2-weighted MRI. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019; 696–700. https://ieeexplore.ieee.org/abstract/document/8759572. Accessed February 18, 2023.
26. Shapey J, Kujawa A, Dorent R, et al. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. Sci Data 2021;8(1):286.
27. Amin SB, Anderson KJ, Boudreau CE, et al. Comparative Molecular Life History of Spontaneous Canine and Human Gliomas. Cancer Cell 2020;37(2):243–257.e7.
28. Buda M, Saha A, Walsh R, et al. A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images. JAMA Netw Open 2021;4(8):e2119100.
29. Jordan P, Adamson PM, Bhattbhatt V, et al. Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours. Med Phys 2022;49(5):3523–3528.
30. Ziegler E, Urban T, Brown D, et al. Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research. JCO Clin Cancer Inform 2020;4(4):336–345.
31. Gorman C, Punzo D, Octaviano I, et al. Interoperable slide microscopy viewer and annotation tool for imaging data science and computational pathology. Nat Commun 2023;14(1): 1572.
32. Perkel JM. Why Jupyter is data scientists' computational notebook of choice. Nature 2018;563(7729):145–146.
33. Birger C, Hanna M, Salinas E, et al. FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs. bioRxiv 209494 [preprint] https://www.biorxiv.org/content/10.1101/209494v1. Published November 3, 2017. Accessed January 5, 2018.

34. Lau JW, Lehnert E, Sethi A, et al; Seven Bridges CGC Team. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Res 2017;77(21):e3–e6 [Published correction appears in Cancer Res 2018;78(17):5179.].
35. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199–217.
36. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008;31(4):466–475.
37. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front Neuroinform 2018;11:76.
38. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018;102(4):1143–1158.
39. Schacherer DP, Herrmann MD, Clunie DA, et al. The NCI Imaging Data Commons as a platform for reproducible research in computational pathology. Comput Methods Programs Biomed 2023;107839.
40. Bontempi D, Nuernberg L, Krishnaswamy D, et al. Transparent and reproducible AI-based medical imaging pipelines using the cloud. Research Square 3142996 [preprint] https://www.researchsquare.com/article/rs-3142996/latest.pdf. Published July 21, 2023. Accessed July 31, 2023.
41. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.
42. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–211.
43. Schuhegger S. Body Part Regression for CT Images. arXiv 2110.09148 [preprint] http://arxiv.org/abs/2110.09148. Published October 18, 2021. Accessed August 10, 2023.
44. Ghose S, Oliver A, Martí R, et al. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Comput Methods Programs Biomed. Elsevier Ireland Ltd, 2012.
45. Sunoqrot MRS, Saha A, Hosseinzadeh M, Elschot M, Huisman H. Artificial intelligence for prostate MRI: open datasets, available applications, and grand challenges. Eur Radiol Exp 2022;6(1):35.
46. Jorge Cardoso M, Li W, Brown R, et al. MONAI: An open-source framework for deep learning in healthcare. arXiv 2211.02701 [preprint] https://arxiv.org/abs/2211.02701. Published November 4, 2022. Accessed August 10, 2023.
47. Adams LC, Makowski MR, Engel G, et al. Prostate158 - An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. Comput Biol Med 2022;148:105817.
48. Antonelli M, Reinke A, Bakas S, et al. The Medical Segmentation Decathlon. Nat Commun 2022;13(1):4128.
49. ImagingDataCommons/idc-prostate-mri-analysis. Github. https://github.com/ImagingDataCommons/idc-prostate-mri-analysis. Accessed June 23, 2023.
50. Fedorov A, Schwier M, Clunie D, et al. An annotated test-retest collection of prostate multiparametric MRI. Sci Data 2018;5(1):180281.
51. Moore J, Linkert M, Blackburn C, et al. OMERO and Bio-Formats 5: flexible access to large bioimaging datasets at scale. In: Ourselin S, Styner MA, editors. Medical Imaging 2015: Image Processing; 2015:37–42.
52. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. J Pathol Inform 2013;4(1):27.
53. Bridge CP, Gorman C, Pieper S, et al. Highdicom: a Python Library for Standardized Encoding of Image Annotations and Machine Learning Model Outputs in Pathology and Radiology. J Digit Imaging 2022;35(6):1719–1737.
54. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J Neurosci Methods 2016;264:47–56.
55. Kulkarni P, Kanhere A, Yi PH, Parekh VS. Text2Cohort: Democratizing the NCI Imaging Data Commons with natural language cohort discovery. arXiv 2305.07637 [preprint] http://arxiv.org/abs/2305.07637. Updated May 16, 2023. Accessed May 16, 2023.
56. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A Case for Data Commons: Toward Data Science as a Service. Comput Sci Eng 2016;18(5):10–20.
57. Uray M, Hirsch E, Katzinger G, Gadermayr M. Beyond desktop computation: Challenges in scaling a GPU infrastructure. arXiv 2110.05156 [preprint] http://arxiv.org/abs/2110.05156. Published October 11, 2021. Accessed June 8, 2023.
58. Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell 2023;5(5):e230024.
59. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77(21):e104–e107.
60. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012 Nov;30(9):1323–1341.
61. Baughan N, Whitney H, Drukker K, et al. Sequestration of imaging studies in MIDRC: a multi-institutional data commons. In: Mello-Thoms CR, Taylor-Phillips S, eds. Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment. SPIE, 2022; 91–98. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12035/120350F/Sequestration-of-imaging-studies-in-MIDRC–a-multi-institutional/10.1117/12.2610239.short. Accessed August 7, 2023.
62. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020;11(1):2624.
63. Gaillard F. Radiopaedia: Building an Online Radiology Resource. European Congress of Radiology: RANZCR ASM 2011. https://epos.myesr.org/ranzcr/viewing/index.php?module=viewing_poster&task=viewsection&pi=108203&ti=339767&si=1126&searchkey=. Accessed August 7, 2023.
64. Yang J, Shi R, Wei D, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data 2023;10(1):41.
65. Gupta Y, Costa C, Pinho E, Bastião Silva L. DICOMization of Proprietary Files Obtained from Confocal, Whole-Slide, and FIB-SEM Microscope Scanners. Sensors (Basel) 2022;22(6):2322.
66. Fischer M, Schader P, Braren R, et al. DICOM Whole Slide Imaging for Computational Pathology Research in Kaapana and the Joint Imaging Platform. Bildverarbeitung für die Medizin 2022. Springer Fachmedien Wiesbaden, 2022; 273–278.
67. MHub. https://mhub.ai. Accessed June 9, 2023