NCI Imaging Data Commons

Andrey Fedorov*[1], William J.R. Longabaugh[2], David Pot[3], David A. Clunie[4], Steve Pieper[5], Hugo J.W.L. Aerts[6,7,8], André Homeyer[9], Rob Lewis[10], Afshin Akbarzadeh[1], Dennis Bontempi[6], William Clifford[2], Markus D. Herrmann[11], Henning Höfener[9], Igor Octaviano[10], Chad Osborne[3], Suzanne Paquette[2], James Petts[12], Davide Punzo[10], Madelyn Reyes[3], Daniela P. Schacherer[9], Mi Tian[2], George White[2], Erik Ziegler[10], Ilya Shmulevich[2], Todd Pihl[13], Ulrike Wagner[13], Keyvan Farahani[14], Ron Kikinis[1`]

[1]*Brigham and Women's Hospital, Department of Radiology, Boston, MA, USA*
[2]*Institute for Systems Biology, Seattle, WA, USA*
[3]*General Dynamics, Bethesda, MD, USA*
[4]*PixelMed Publishing LLC, Bangor, PA, USA*
[5]*Isomics Inc, Cambridge, MA, USA*
[6]*Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, United States of America*
[7]*Departments of Radiation Oncology & Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, United States of America.*
[8]*Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands*
[9]*Fraunhofer MEVIS, Department of Quantitative Pathology, Bremen, Germany*
[10]*Radical Imaging, Boston, MA*
[11]*Massachusetts General Hospital, Department of Pathology, Boston, MA, USA*
[12]*Ovela Solutions LTD, London, UK*
[13]*Frederick National Laboratory for Cancer Research, Frederick, MD, USA*
[14]*National Cancer Institute, Bethesda, MD, USA*

Corresponding author: Andrey Fedorov, fedorov@bwh.harvard.edu, Brigham and Women's Hospital, 1249 Boylston St #344, Boston, MA 02215. Phone: 617-525-6258.

Running title: NCI Imaging Data Commons

Abbreviations

NCI - National Cancer Institute

CRDC - Cancer Research Data Commons

IDC - Imaging Data Commons

FAIR - Findable Accessible Interoperable Reusable

TCIA - The Cancer Imaging Archive

DICOM - Digital Imaging and Communications in Medicine

GCP - Google Cloud Platform

ISB-CGC - Institute for Systems Biology Cancer Gateway in the Cloud

SBG-CGC - Seven Bridges Genomics Cancer Genomics Cloud

SQL - Structured Query Language

OHIF - Open Health Imaging Foundation

BRIDG - Biomedical Research Integrated Domain Group

HL7 - Health Level 7

FISMA - Federal Information Security Modernization Act

PHI - Protected Health Information

TCGA - The Cancer Genome Atlas

VM - Virtual Machine

API - Application Programming Interface

URL - Uniform Resource Locator

GPU - Graphics Processing Unit

CDA - Cancer Data Aggregator

STRIDES - Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability

NIH - National Institutes of Health

HTAN - Human Tumor Atlas Network

COVID-19 - Coronavirus Disease 2019

DICOM SEG - DICOM Segmentation

DICOM SR - DICOM Structured Reporting

DICOM RTSS - DICOM Radiotherapy Structure Set

SNOMED - Systematized Nomenclature of Medicine

NCIt - NCI Thesaurus

CCDH - Center for Cancer Data Harmonization

GUID - Global Unique Identifier

GA4GH - Global Alliance for Genomics in Health

DRS - Data Repository Service

Conflict of interest statement

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Task Order No. HHSN26110071 under Contract No. HHSN2612015000031.

Abstract

The National Cancer Institute (NCI) Cancer Research Data Commons (CRDC) aims to establish a national cloud-based data science infrastructure. Imaging Data Commons (IDC) is a new component of CRDC supported by the Cancer Moonshot™. The goal of IDC is to enable a broad spectrum of cancer researchers, with and without imaging expertise, to easily access and explore the value of de-identified imaging data and to support integrated analyses with non-imaging data. We achieve this goal by co-locating versatile imaging collections with cloud-based computing resources and data exploration, visualization, and analysis tools. The IDC pilot was released in October 2020 and is being continuously populated with radiology and histopathology collections. IDC provides access to curated imaging collections, accompanied by documentation, a user forum, and a growing number of analysis use cases that aim to demonstrate the value of a data commons framework applied to cancer imaging research.

Significance: This study introduces NCI Imaging Data Commons, a new repository of the NCI Cancer Research Data Commons, which will support cancer imaging research on the cloud.

Introduction

Scalable on-demand access to managed configurable cloud resources offers unprecedented opportunities in supporting cancer research. The cloud-computing paradigm of co-locating large multifaceted datasets with the compute resources and bringing tools to the data instead of downloading the data for analysis, has the potential to address numerous challenges associated with big data research (e.g., storage and bandwidth constraints, and reproducibility of the analysis). The National Cancer Institute (NCI) Cancer Research Data Commons (CRDC), a component of a national cancer data ecosystem [1], is a cloud-based environment that aims to realize the promise of the cloud [2,3]. Primary components of CRDC include cloud-based domain-specific data repositories [4] and analysis-focused cloud resources [5–7]. The NCI Imaging Data Commons (IDC) is a new data repository of CRDC that co-locates imaging data with the compute resources and analysis tools within the CRDC cloud environment and provides researchers with access to I) cancer image collections, II) infrastructure for exploration of metadata and imaging data, and III) interfaces to other components of CRDC enabling integrated analysis across various data types contained in CRDC (i.e., matching genomic and proteomic data).

Following the guiding principles of CRDC, IDC builds on the strengths of the established efforts to collect and share FAIR (Findable Accessible Interoperable Reusable) [8] imaging data, and especially that of The Cancer Imaging Archive (TCIA) [9]. While TCIA has been successful in supporting researchers that utilize the traditional approach of downloading image data for analysis using local resources, IDC aims to make public TCIA collections available within, and tightly integrated with, the CRDC cloud environment, expanding the scope over time to include data from sources other than TCIA. To organize imaging data collected at multiple sites and by different modalities, IDC uses an extensible and documented standards-based approach to enable search operations and interoperability with analysis tools. IDC relies on the DICOM (Digital Imaging and Communications in Medicine) standard [10] for the definition of the data model and interfaces for accessing data, and for harmonizing the representation of data and metadata.

The role of IDC extends beyond establishing an infrastructure for cloud-based cancer imaging research. We are actively developing use cases demonstrating how this infrastructure can be utilized efficiently for research tasks that would be more difficult to achieve “on premises”. All of the code developed by the project is being shared under non-restrictive open source licenses, and much of the code has been contributed back to established libraries and toolkits as a way to contribute further to the scientific community.

In this report we introduce IDC, describing its overall architecture and components as well as the current status and the priorities of the project.

Methods

Cloud platform

We chose to implement IDC using a combination of commercially available tools and capabilities provided by the Google Cloud Platform (GCP) and its Healthcare API, together with a range of open source components, as shown in Figure 1. The choice of GCP was motivated by our desire to expediently deliver robust industry-grade infrastructure and ensure its integration with the existing components of CRDC. GCP implements a range of capabilities to support administration and security of the system, and provides a continuously evolving set of tools for scalable analysis of big data. Being one of the major cloud provider platforms, GCP is already used by the CRDC Cloud Resources: FireCloud [6], the Institute of Systems Biology Cancer Gateway in the Cloud (ISB-CGC) [5] and Seven Bridges Genomics Cancer Genomics Cloud (SBG-CGC) [7]. Our prior experience building ISB-CGC [5] allowed us to leverage its components in establishing IDC. The GCP Healthcare API provides support for "DICOM stores", which are accessible via the standard DICOMweb interface. The API includes tools for exporting DICOM metadata into BigQuery tables. BigQuery is a GCP scalable data warehouse solution based on Dremel [11], which enables high performance queries of very large tables using Structured Query Language (SQL) compliant with the SQL 2011 standard.

Portal

Similar to the already established nodes of CRDC, the IDC search portal provides an interface for exploring available data, defining cohorts of cases, and summarizing attributes of the cohort (see Videos 1 and 2). The portal supports exploration of the metadata, imaging data and image-derived data. The IDC portal shares the code base with the ISB-CGC [5] portal. The faceted search utilizes Apache Solr [12] populated from BigQuery content to reduce latency of certain types of queries (e.g., support of facet counting). In the current deployment of IDC, radiology images are displayed with the open source OHIF Viewer [13], which uses DICOMweb to access the IDC data. The OHIF Viewer is being actively developed, with the IDC project being one of many contributors. As IDC evolves to support new data types, alternative viewers specializing in viewing specific types of images may be integrated with the platform in the future. To address the need for display of brightfield and fluorescence microscopy images in DICOM format, IDC is working to leverage the Slim viewer (https://github.com/mghcomputationalpathology/slim). Like the OHIF Viewer, Slim viewer is a serverless single-page application that facilitates interactive visualization, in this case for digital slide microscopy images. Slim also supports image annotations in the web browser, relying on DICOMweb to query and dynamically retrieve image data from the DICOM store just as in the radiology case.

Data modeling

IDC will host a variety of cancer imaging data. While the initial focus is to support radiology data, IDC aims to provide similar capabilities for collections of brightfield microscopy, multi-channel immunofluorescence, and other imaging modalities. Equally important is the ability to support the results obtained by analysis of imaging data, such as annotations of image regions of interest or various descriptors of image findings. DICOM defines data models and standard information objects that cover a significant portion of the expected needs in communicating image analysis results [14–16]. It can also be extended to support new types of data, wherever possible retaining compatibility with legacy systems [17]. IDC relies on the data model defined by the DICOM standard and on the definitions of the DICOM objects to ensure their validity. DICOM is harmonized with several

other healthcare standards (e.g., BRIDG [18] and HL7 [19]) and relies on standard vocabularies and ontologies [20], thus facilitating integration of IDC imaging data with other types of data within CRDC.

Security

As a government-owned system, IDC is required to obtain and maintain data security at the Federal Information Security Modernization Act (FISMA) Low level. While FISMA Low is less demanding than higher levels, this requirement has major implications on allocation of the engineering effort for the implementation and upkeep of the security, logging and reporting procedures, and for the users interacting with the system. IDC cannot host data that contains Protected Health Information (PHI). De-identification is performed outside of IDC and is currently done through TCIA, and in the future via additional Data Coordinating Centers. De-identification procedures implemented at additional future sources of data would need to be independently vetted before the data contributed by those sources can be hosted by IDC. While no PHI data can be included in the collections hosted by IDC, IDC users wishing to combine non-public data with the public collections can do this using CRDC Cloud Resources, which have FISMA Moderate designation, or using independent cloud projects with access to IDC public resources.

Development process and governance

The IDC development is supported through a contract between Leidos Biomedical Research and The Brigham and Women's Hospital with specific deliverables. Strategic guidance is provided by the National Cancer Institute, the Frederick National Laboratory for Cancer Research, advisory boards and stakeholders. IDC embraces the main principles of Agile development methodology, including incremental development and continuous customer involvement. While IDC is not required to use only open source components, all of the code developed by IDC is being released under permissive open source licenses. Our intent is to enable reuse of the individual open source components to support replication of the relevant capabilities of IDC.

Results

The pilot of IDC was released in October 2020, and its high-level organization, relationship to the other components of CRDC, and interaction with the user flow are summarized in Figure 1. Included in the release were 28 collections of the TCIA: radiology images related to The Cancer Genome Atlas (TCGA) project and several collections prioritized to establish the capabilities of IDC in handling image-derived data (e.g., LIDC-IDRI and NSCLC-Radiomics collections). Access to the data is available from the GCP "requester pays" storage buckets (i.e., a user-provided Google billing project is required to read the data, although loading content onto a GCP VM is free). DICOM and collection-level metadata is available from the BigQuery tables and does not require a project configured with billing. The IDC portal (available at https://imaging.datacommons.cancer.gov, also see Figure 2) allows users to define cohorts based on a subset of metadata, provides graphical summaries of the cohort attributes, and integrates a customized OHIF Viewer that supports visualization of both the images and image annotations (specifically, visualization of DICOM Segmentation and Radiotherapy Structure Set is supported, including multiplanar reformatting). All of the software components developed by the IDC team are available under the dedicated GitHub organization (https://github.com/ImagingDataCommons). Improvements and new features for the OHIF Viewer are developed in its main repository or the repositories of underlying libraries.

IDC enables the following user flow (also see Figure 1). The portal's faceted search [21] user interface (UI) will typically serve as the entry point for the new users, allowing them to explore the data (both by viewing the images and searching the metadata) and build cohorts (see Video 1). Alternatively, users will be able to utilize the IDC

API, which we intend to be functionally equivalent to the IDC Portal, to form and interact with the cohorts. Metadata attributes that are not available via the IDC Portal can be explored using BigQuery or DataStudio (see Video 2 and 3). Standard SQL and BigQuery APIs are available for interrogating the metadata and fine-tuning the definition of the cohort. Users can spot-check data quality by analyzing metadata and examining data in the IDC Viewer, which can be done either through the portal, or by configuring the viewer URL directly to show specific imaging studies. Data quality checks can utilize existing, continuously evolving general purpose cloud-based tools, such as Colab Notebooks (cloud-hosted GPU-enabled virtual machines (VMs) with the Jupyter Notebook interface) or Google DataStudio (interactive platforms for building data dashboards) (see Video 3). At the next level, the user can initialize a cloud-based instance of a VM configured with the familiar desktop-based analysis tools to experiment with customized processing and visualizations on a subset of cases (see Video 4). Once the analysis workflow is established, it can be applied at scale to the entire cohort utilizing either general-purpose pipelining tools [22], or the CRDC Cloud Resources [5–7]. Ability to identify matching data in other repositories of CRDC is being provided by the Cancer Data Aggregator (CDA) [23] APIs currently under development.

Support and engagement of IDC users is a major priority for the project. To support user training and outreach, IDC is accompanied by online documentation, examples of Colab Notebooks (including those contributed by IDC users) and DataStudio dashboards interacting with the IDC-hosted data, as well as video tutorials (see Videos 1-4 included with this article). Further use cases demonstrating implementation of radiomics and pathomics analysis pipelines integrated with the IDC data are currently under development. Users can participate in the IDC online forum based on the Discourse platform. Complete analysis use cases that demonstrate the capabilities of IDC to support imaging research needs are being developed, with the first such use case replicating an earlier study by Hosny et al [24] already available (see Video 4).

Prospective IDC users can apply for free GCP credits to experiment with the resource and develop confidence with the cloud-based analysis. Experienced investigators can participate in the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative, which provides cloud training resources and discounted credits to all NIH-funded investigators and is intended to support production use of CRDC.

Discussion

We described the design and implementation of IDC and summarized capabilities of the IDC pilot available to the cancer research community. Early examples already show promise for the utility of cloud-hosted public imaging collections co-located with the compute resources and a growing number of tools to support data analysis. The IDC Portal supports exploration and cohort building from cloud-based data. The IDC Viewer provides unique and growing capabilities in supporting visualization of image annotations. Combined with BigQuery, IDC offers the unprecedented ability to access and explore DICOM metadata for public imaging collections from the IDC-maintained tables. The analysis use cases that accompany IDC illustrate the ease of access to the data from cloud-based tools and the potential to enable sharing of fully reproducible analysis pipelines to accompany academic manuscripts.

IDC is under active development to further enhance both the capabilities of IDC itself and its integration with the other components of CRDC. Immediate priorities for the development of IDC are the data versioning strategy (motivated by the updates to the released collections due to addition of new data, correction of errors, or mitigation of PHI leaks) and subsequent ingestion and periodic updates towards inclusion of all the public TCIA radiology collections. Support for digital pathology is also planned for the production release currently scheduled for Fall 2021. Existing public collections of digital pathology images, which are typically shared using vendor-

specific formats, will be converted into a DICOM representation to better support metadata search and visualization. The datasets hosted by IDC will not be limited to human data, nor to the modalities currently available from TCIA. We expect IDC to host pre-clinical (mouse) and canine imaging data, as well as various types of images that will be shared by the NCI Human Tumor Atlas Network (HTAN) [25]. IDC will also include relevant non-cancer imaging collections, as prioritized by the NCI stakeholders, such as the recently announced COVID-19 collections released by TCIA.

Alongside replication of the imaging collections, IDC supports inclusion of image-derived data (e.g., annotations, measurements and regions of interest) and accompanying clinical data. Harmonization of clinical data is being done in coordination with the CRDC Center for Cancer Data Harmonization (CCDH) [26] and the CDA teams. Harmonization of image-derived data is a major undertaking in the IDC data intake process. Common coded and structured data representation in standard formats (DICOM SR, SEG and RTSS) using standard coded concepts for fields and value sets (SNOMED, NCIt) is critical to enable metadata search across collections, to provide a consistent interface to the data for visualization and analysis tools, and for semantic interoperability between CRDC nodes. We are actively working on these harmonization tasks both for the retrospective collections and for prospective submission of analysis results to TCIA.

IDC will be relying on global unique identifiers (GUIDs) to support persistent referencing of the data. The CRDC Data Commons Framework is in the process of implementing the relevant parts of the Global Alliance for Genomics in Health (GA4GH) [27] Data Repository Service (DRS) API [28] to support GUIDs for the data bundles at the selected levels of the DICOM hierarchical data model.

IDC is in its early days. There are numerous questions relating to costs of conducting imaging research in the cloud and limiting the risk of runaway processes. Repositories of reusable image analysis tools that are easily accessible from cloud workflows (with the relevant existing platforms including Dockstore [29] and ModelHub.AI [30]) need to be established. Integrative analysis of data across CRDC nodes needs to be enabled. We hope to engage the future users of IDC, as well as contributors and maintainers of emerging repositories of cancer research tools, through venues such as the IDC online forum (https://discourse.canceridc.dev/). Working together, we can answer these questions and develop new components of the CRDC ecosystem to support a broad range of cancer imaging research use cases. With the pilot release, we introduce an early example of the capabilities and the potential for applying the data commons concepts to the imaging space.

Acknowledgments

We are acknowledging the support of NCI Communications in refining the video materials accompanying this submission. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Task Order No. HHSN26110071 under Contract No. HHSN2612015000031. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

[1] Jaffee EM, Dang CV, Agus DB, Alexander BM, Anderson KC, Ashworth A, et al. Future cancer research priorities in the USA: a Lancet Oncology Commission. Lancet Oncol 2017;18:e653–706. https://doi.org/10.1016/S1470-2045(17)30698-8.
[2] 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;18:10–20. https://doi.org/10.1109/MCSE.2016.92.

[3] Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front Cell Dev Biol 2017;5:83. https://doi.org/10.3389/fcell.2017.00083.
[4] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood 2017;130:453–9.
[5] Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, et al. The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. Cancer Res 2017;77:e7–10. https://doi.org/10.1158/0008-5472.CAN-17-0617.
[6] Birger C, Hanna M, Salinas E, Neff J, Saksena G, Livitz D, et al. FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs. bioRxiv 2017:209494. https://doi.org/10.1101/209494.
[7] Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res 2017;77:e3–6. https://doi.org/10.1158/0008-5472.CAN-17-0387.
[8] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
[9] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.
[10] 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;4:199–212. https://doi.org/10.1136/jamia.1997.0040199.
[11] Melnik S, Gubarev A, Long JJ, Romer G, Shivakumar S, Tolton M, et al. Dremel: interactive analysis of web-scale datasets. Proceedings VLDB Endowment 2010;3:330–9. https://doi.org/10.14778/1920841.1920886.
[12] Shahi D. Apache Solr: A Practical Approach to Enterprise Search. Apress, Berkeley, CA; 2015. https://doi.org/10.1007/978-1-4842-1070-3.
[13] Ziegler E, Urban T, Brown D, Petts J, Pieper SD, Lewis R, 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;4:336–45. https://doi.org/10.1200/CCI.19.00131.
[14] Fedorov A, Clunie D, Ulrich E, Bauer C, Wahle A, Brown B, 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. https://doi.org/10.7717/peerj.2057.
[15] Herrmann MD, Clunie DA, Fedorov A, Doyle SW, Pieper S, Klepeis V, et al. Implementing the DICOM standard for digital pathology. J Pathol Inform 2018;9:37. https://doi.org/10.4103/jpi.jpi_42_18.
[16] Fedorov A, Beichel R, Kalpathy-Cramer J, Clunie D, Onken M, Riesmeier J, et al. Quantitative Imaging Informatics for Cancer Research. JCO Clin Cancer Inform 2020;4:444–53. https://doi.org/10.1200/CCI.19.00165.
[17] Clunie DA. Dual-Personality DICOM-TIFF for whole slide images: A migration technique for legacy software. J Pathol Inform 2019;10:12. https://doi.org/10.4103/jpi.jpi_93_18.
[18] Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, et al. BRIDG: a domain information model for translational and clinical protocol-driven research. J Am Med Inform Assoc 2017;24:882–90. https://doi.org/10.1093/jamia/ocx004.
[19] Indrajit IK, Verma BS. DICOM, HL7 and IHE: A basic primer on Healthcare Standards for Radiologists. Indian J Radiol Imaging 2007;17:66. https://doi.org/10.4103/0971-3026.33610.
[20] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662–6.
[21] Russell-Rose T, Tate T. Faceted Search. Designing the Search Experience, Elsevier; 2013, p. 167–218. https://doi.org/10.1016/b978-0-12-396981-1.00007-0.
[22] Larsonneur E, Mercier J, Wiart N, Floch EL, Delhomme O, Meyer V. Evaluating Workflow Management Systems: A Bioinformatics Use Case. 2018 IEEE International Conference on Bioinformatics and

Biomedicine (BIBM), ieeexplore.ieee.org; 2018, p. 2773–5. https://doi.org/10.1109/BIBM.2018.8621141.
[23] Cancer Data Aggregator n.d. https://datacommons.cancer.gov/cancer-data-aggregator (accessed May 17, 2021).
[24] Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med 2018;15:e1002711. https://doi.org/10.1371/journal.pmed.1002711.
[25] Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell 2020;181:236–49. https://doi.org/10.1016/j.cell.2020.03.053.
[26] NCI Center for Cancer Data Harmonization (CCDH) n.d. https://harmonization.datacommons.cancer.gov/ (accessed May 17, 2021).
[27] Terry SF. The global alliance for genomics & health. Genet Test Mol Biomarkers 2014;18:375–6. https://doi.org/10.1089/gtmb.2014.1555.
[28] GA4GH Data Repository Service n.d. https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.0.0/docs/ (accessed May 17, 2021).
[29] O'Connor BD, Yuen D, Chung V, Duncan AG, Liu XK, Patricia J, et al. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. F1000Res 2017;6:52. https://doi.org/10.12688/f1000research.10137.1.
[30] Hosny A, Schwier M, Berger C, Örnek EP, Turan M, Tran PV, et al. ModelHub.AI: Dissemination Platform for Deep Learning Models. arXiv [csLG] 2019.

Figures

Figure 1: High-level diagram of relevant components of the Imaging Data Commons and related entities, and their relation to the steps of the envisioned CRDC user flow with the emphasis on imaging applications. *Green boxes correspond to the envisioned user flow. IDC Extract Transform Load (ETL) process maintains the content of the data collected by external entities (e.g., TCIA) co-located with the various cloud-based tools, such as those maintained by Cloud Resources or by the Google Cloud Platform. The data can be accessed using both the interactive components (e.g., IDC Portal and Viewer) and programmatic APIs.*

Figure 2: Elements of IDC Portal user interface. *Left: front page of the IDC Portal for the pilot (pre-production) release of the platform, available at https://imaging.datacommons.cancer.gov. Right: example of filters available for defining cohort based one of the attributes describing segmentation results available in IDC.*

Videos

Video 1: Introduction to IDC Portal. This video guides the user over the main capabilities of the IDC Portal including exploration of the data collections available within IDC, using metadata to select relevant subsets of data, and visualization of the images and annotations.

Video 2: Introduction to IDC Cohorts. In this video we discuss the basic operations with IDC cohorts: how to define and save a cohort, and how to replicate the files corresponding to the cohort on a cloud virtual machine for subsequent analysis.

Video 3: Custom DataStudio dashboards and IDC. Google DataStudio is a free tool for creating interactive data dashboards and reports. This video covers the steps needed to prepare a customizable dashboard for an IDC cohort.

Video 4: A case study integrating image analysis pipeline with IDC. This video contains an overview of a case study replicating analysis published earlier using an interactive Python notebook hosted on the Google Colab platform, and interacting with IDC for data selection and visualization.

Cancer Research Data Commons (CRDC)

Cancer Data Aggregator

(Non-Imaging) Data Repositories

- Genomics Data Commons
- Proteomics Data Commons
- Integrated Canine Data Commons

Cloud Resources

- Broad FireCloud
- SevenBridges Genomic Cloud
- ISB Cancer Gateway in the Cloud

Identify available imaging data, assess quallity → Identify matching non-imaging data → Define cohorts → Perform exploratory analysis → Scale up analysis → Share analysis results and tools

Imaging Data Commons (IDC)

- **API**
- **Portal**
- **Viewer**

- **GCP Storage** DICOM files
- **GCP BigQuery** DICOM metadata tables
- **GCP Healthcare API** DICOM stores

Curated public imaging collections

Extract/Transform/Load (ETL)

Google Cloud Platform (GCP)

- **Colab Notebooks**
- **DataStudio**
- **Compute Engine**
- **AI Notebooks**

- **Data Commons Framework**
- **Data Coordinating Centers**
- **The Cancer Imaging Archive (TCIA)**

NIH NATIONAL CANCER INSTITUTE
Imaging Data Commons

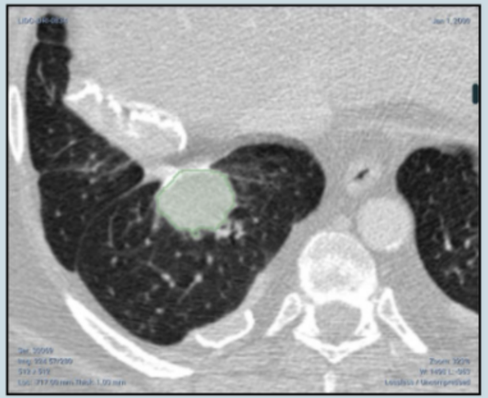
Collections | Exploration | Discourse | Documentation | Sign In

Get started today! Contact us about setting up your own Google Cloud Platform Project with free cloud credits

Collections | Exploration

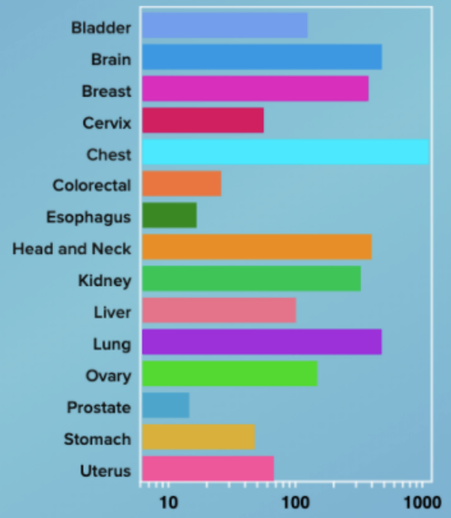
Author Manuscript Published OnlineFirst on June 15, 2021; DOI: 10.1158/0008-5472.CAN-21-0950
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

RADIOLOGY



Computed Tomography (CT) | Magnetic Resonance (MR) | Positron Emission Tomography (PET)

Cases by Major Primary Site



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

Pilot Application Version: *canceridc.202103101841.9084d0d*

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

NCI Imaging Data Commons is supported by the contract number 19X037Q from Leidos Biomedical Research under Task Order HHSN26100071 from NCI.

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

AACR American Association for Cancer Research

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

NCI Imaging Data Commons

Andrey Fedorov, William J.R. Longabaugh, David Pot, et al.

Cancer Res Published OnlineFirst June 15, 2021.