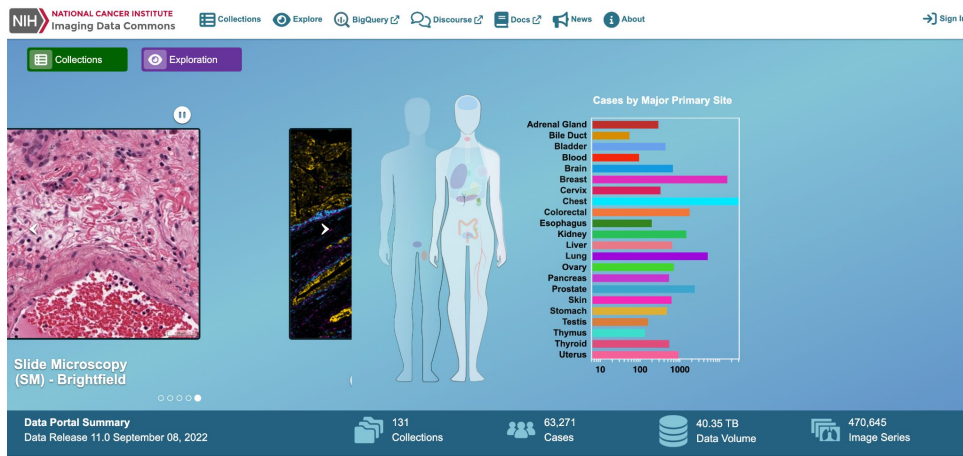


NCI Imaging Data Commons

*Andrey Fedorov, PhD, on behalf of the IDC team
Brigham and Women's Hospital, Mass General Brigham, Boston
28 November 2022*

National Cancer Institute (NCI) Imaging Data Commons (IDC)

NCI Imaging Data Commons (IDC) is a cloud-based repository of publicly available cancer imaging data co-located with the analysis and exploration tools and resources.



- Public DICOM imaging datasets
- Images, image-derived and image-related (clinical) data
- Radiology, digital pathology, and more
- Tools for search, visualization, exploration of data

Data Portal Summary
Data Release 12.0 October 11, 2022

128
Collections

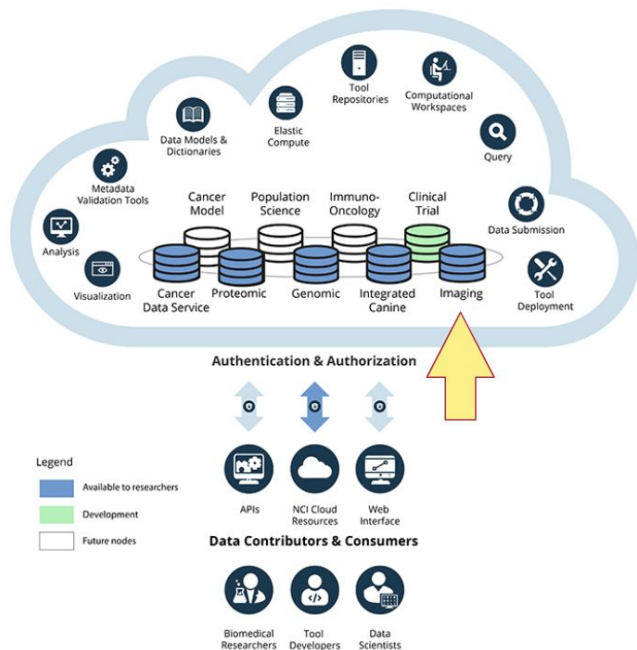
63,316
Cases

40.37 TB
Data Volume

470,850
Image Series

<https://imaging.datacommons.cancer.gov>

NCI Imaging Data Commons (IDC)



IDC is a component within the broader NCI Cancer Research Data Commons (CRDC) infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data.

NCI Imaging Data Commons in numbers

Collections
131

DOIs
124

Size on disk
44.4T

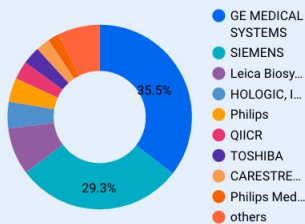
Cases
63,271

Studies
137,978

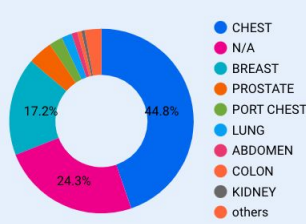
Series
470,645

Instances
41.9M

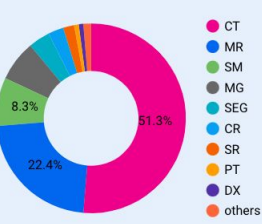
Distinct StudyDate
8,296



Manufacturer
87



BodyPartExamined
84



Modality
21

Cancer types

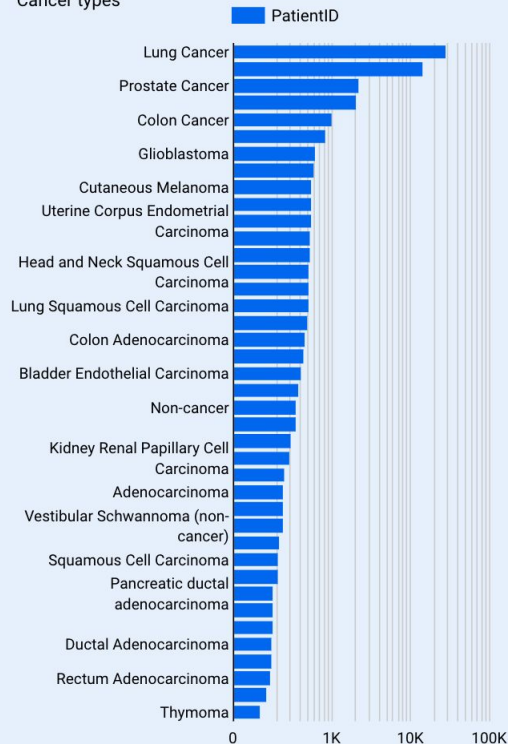


Image collections accompanied by annotations

Collections
38

Cases
6,480

Studies
10,233

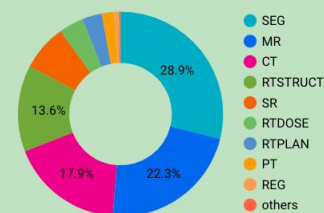
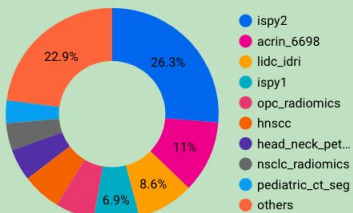


Image collections accompanied by clinical data

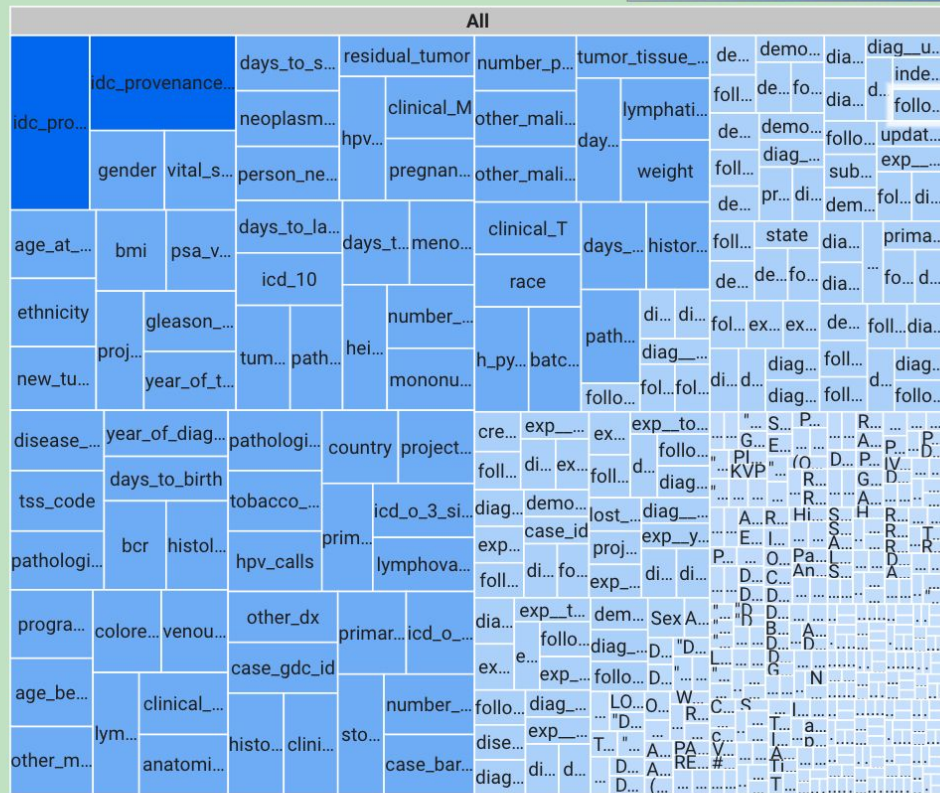
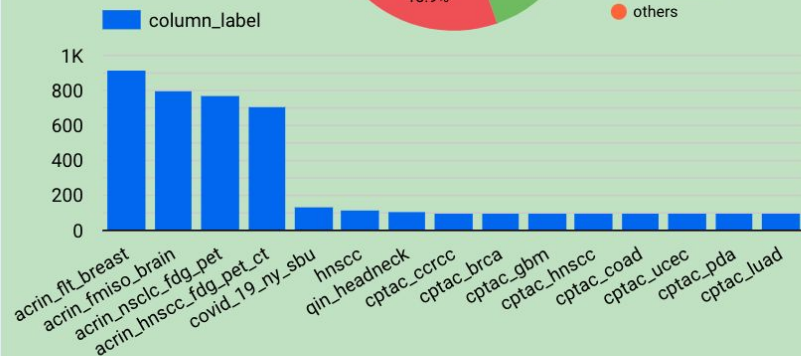
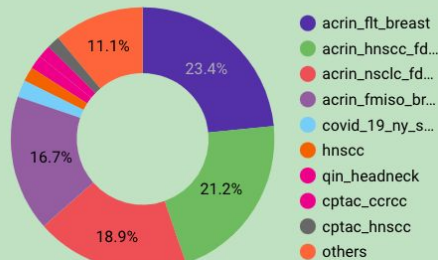
Collections
78

Tables
189

Columns
5,371

Patients
20,935

Right: Summary of the number of distinct columns (dictionary values) per collection.



Left: Count of distinct dictionary values per collection. Right: Distinct dictionary values across all collections with corresponding clinical data (mouse over individual items to see full column label text). Size of the rectangle corresponds to the number of collections where specific column label was encountered.



Explore

Broad range of imaging data across scales, modalities, cancer types.



Subset

Extensive metadata index powered by highly scalable cloud search.



Analyze

Self-contained reproducible end-to-end cloud-based workflows.



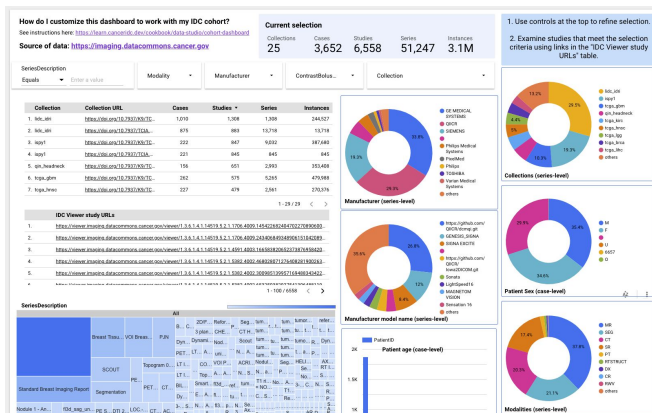
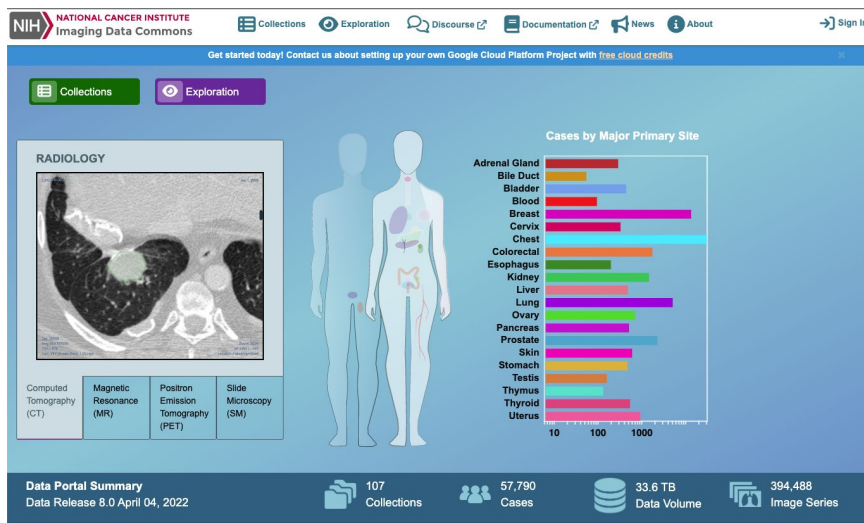
Share

Persistent data cohorts for reproducibility and transparency.

What you get

Without downloading data or installing any software on your computer:

- **Search and explore data**



```

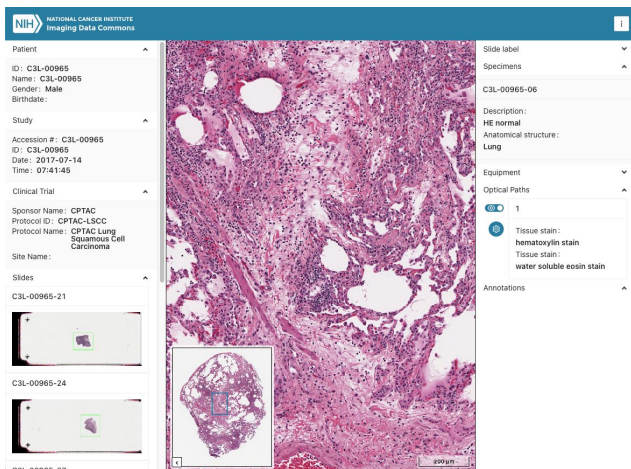
1 WITH
2   nlst_instances_per_series AS (
3     SELECT
4       StudyInstanceUID,
5       SeriesInstanceUID,
6       COUNT(DISTINCT(SOPInstanceUID)) AS num_instances,
7       COUNT(DISTINCT(ARRAY_TO_STRING(ImagePositionPatient,"/"))) AS p
8       COUNT(DISTINCT(ARRAY_TO_STRING(PixelSpacing,"/"))) AS pixel_spa
9       COUNT(DISTINCT(ARRAY_TO_STRING(ImageOrientationPatient,"/"))) A
10      MIN(SAFE_CAST(SliceThickness AS float64)) AS min_SliceThickness
11      MAX(SAFE_CAST(SliceThickness AS float64)) AS max_SliceThickness
12      MIN(SAFE_CAST(ImagePositionPatient[SAFE_OFFSET(2)] AS float64))
13      MAX(SAFE_CAST(ImagePositionPatient[SAFE_OFFSET(2)] AS float64))
14      STRING_AGG(DISTINCT(SAFE_CAST("LOCALIZER" IN UNNEST(ImageType)
15 FROM
16   'bioguerv-public-data.idc.current.dicom.all'

```

What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- **Visualize images and derived data**



Patient

ID: HTA7_926
Name: HTA7_926
Gender: Male
Birthdate:

Study

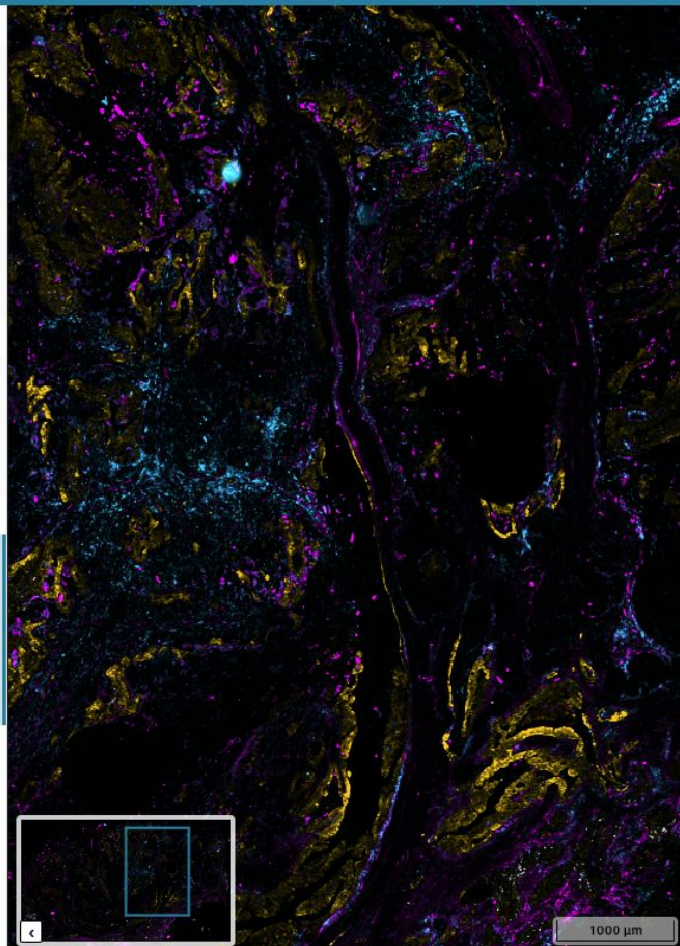
Accession #: HTA7_926
ID: HTA7_926
Date: 1970-01-01
Time: 00:00:00

Clinical Trial

Sponsor Name: NCI
Protocol ID: HTAN-HMS
Protocol Name:
Site Name: Harvard Medical School
Time Point ID: Initial Diagnosis

Slides

HTA7_926_1000



☐ Illumination wavelength: 555 nm
☐ Tissue stain: Pan Cytokeratin Monoclonal AB (AE1/AE3), eFluor 570, eBioscience

☒ 14: CD45 ☐
☐ Illumination wavelength: 555 nm
☐ Tissue stain: PE anti-human CD45

☒ 26: Vimentin ☐
☐ Illumination wavelength: 555 nm
☐ Tissue stain: Vimentin (D21H3) XP Rabbit mAb (Alexa Fluor 555 Conjugate)

☒ 34: Antigen Ki67 (2) ☐
☐ Illumination wavelength: 555 nm
☐ Tissue stain: Ki-67 Monoclonal Antibody (20Raj1), eFluor 570, eBioscience

Presentation States

Pan-cytokeratin-Viment...

Annotations

What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- **Define and save cohorts**

```
SELECT
  StudyInstanceUID,
  gcs_url
FROM
  `bigquery-public-data.idc_current.dicom_all`
WHERE
  Modality = "CT"
  AND collection_id = "n1st"
  AND SAFE_CAST(SliceThickness AS float64) < 1
```

Filter Definition

COLLECTION IN (NSCLC-Radiomics) AND **SEGMENTATION TYPE** IN (Heart)

NIH NATIONAL CANCER INSTITUTE
Imaging Data Commons

Cohorts

Export Manifests

Delete

Showing 1 to 10 of 17 entries Show 10 entries

Previous 1 2 Next

Search:

<input type="checkbox"/>	Cohort ID	Name	Case Count	Study Count	Series Count	Data Version	Version Compare
<input type="checkbox"/>	245	All of SM	456	2218	2218	IDC Data Release Version 3.0 2021-07-21	
<input type="checkbox"/>	244	Chest CT	1320	1598	17046	IDC Data Release Version 3.0 2021-07-21	
<input type="checkbox"/>	230	4 cases v3	4	4	9	IDC Data Release Version 3.0 2021-07-21	
<input type="checkbox"/>	213	4 cases kidney tumor seg	4	4	9	IDC Data Release Version 2.0 2021-04-07 (Inactive)	Compare
<input type="checkbox"/>	175	nodules	875	883	14691	IDC Data Release Version 1.0 2020-10-06 (Inactive)	Compare

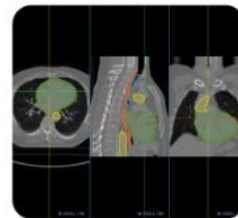
What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- **Develop/share analysis notebooks**



BigQuery Query

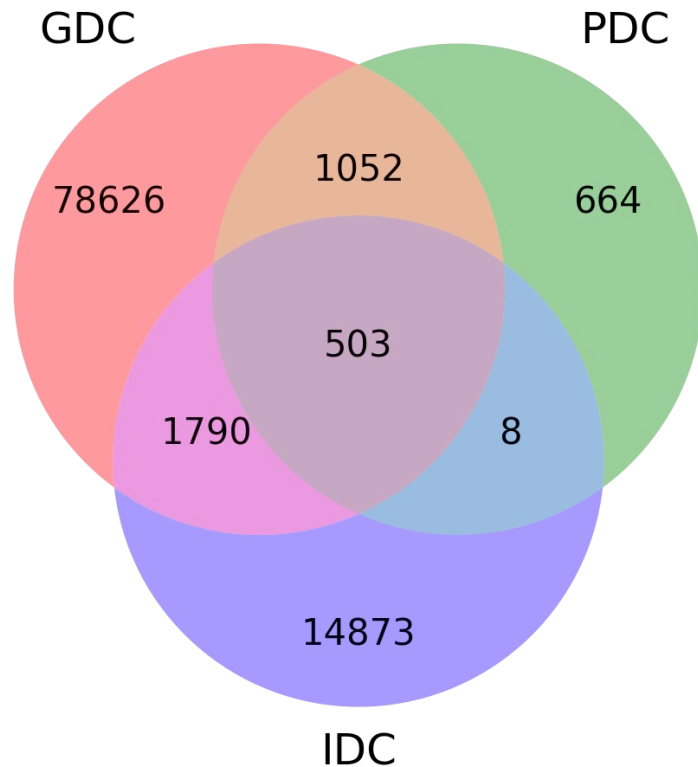
A screenshot of a web-based notebook interface titled 'IDC segmentation primer'. The interface includes a sidebar with a 'Table of contents' and a main content area. The main content area displays the title 'Experimenting with nnU-Net Segmentation of IDC Data' and provides a short link to the notebook: <https://bit.ly/idc-seg-primer>. It also mentions that the notebook is based on the notebook initially presented at RSNA 2021 Deep Learning Lab tutorial series. The 'Learning Objectives' section lists three goals: 1. Understand basic capabilities of IDC. 2. Explore relevant functionality of IDC to support data exploration using Google BigQuery, cohort definition, and retrieval of the data. 3. Learn how to analyze and visualize the data retrieved from IDC on an example of segmentation of abdominal organs at risk.

What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- **Find matching non-imaging data**

<https://datacommons.cancer.gov/cancer-data-aggregator>



Number of cases that have matching data between IDC, Genomic Data Commons (GDC) and Proteomic Data Commons (PDC)

Figure courtesy Fabian Seidl, ISB-CGC

What you get

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- Find matching non-imaging data

Can I just download IDC data from the cloud and move on?

- Yes!



Downloading data

ⓘ You will need to complete prerequisites described in [Getting started with GCP](#) in order to be able to follow the instructions below!

IDC does not have an interactive point-and-click download application! If you want to download data from IDC you will need to use command line interface (Terminal on Mac/Linux or Command prompt on Windows).

Download of data from IDC is a 2-step process covered on this page:

- **Step 1:** create the manifest - the list of files defined by the Google Storage `gs://` URLs;
- **Step 2:** given that list of files, download files to your computer or to a cloud VM.

If you are analyzing IDC data in Google Colab, check out our [Colab cookbook notebook](#) that includes examples of how to query and download IDC data!

<https://learn.canceridc.dev/data/downloading-data>

What you get

FOR FREE

Without downloading data or installing any software on your computer:

- Search and explore data
- Visualize images and derived data
- Define and save cohorts
- Develop/share analysis notebooks
- Find matching non-imaging data



Can I just download IDC data from the cloud and move on?

- **Yes!**
- **no out of cloud egress fees**