Parent [PW40](#) project page:
https://projectweek.na-mic.org/PW40_2024_GranCanaria/Projects/InvestigatingTheRoleOfPublicImagingDataInResearch/

Discord channel: https://discord.gg/H8eh8zTe

Survey organization

Below is just an initial dump of ideas to get us started - I think the challenge will be in coming up with a reasonably sized list of questions that are informative but not overwhelming - it will be easy to come up with a survey they will not be feasible to fill out!

WIP Form:
https://docs.google.com/forms/d/e/1FAIpQLSdaprbZ_113UXgu-p7G2J-R1MwQ-n2KhXqAiD8_wYy0pZK1-Q/viewform?usp=sf_link

About: The purpose of this survey is to evaluate the current trends of using public data and public data repositories, and to try to deduce recommendations for improvement and future development of those. Thank you for taking the time to respond to this survey.

1. About me:
    a. My role: select from choices: [student, postdoc, industry, PI, clinician, software developer, other]
    b. Education: select from choices: [graduate degree (MS or PhD), BS, MD, other]
    c. Years working in medical imaging research: select from choices [beginner, 1-5 years, >5 years]
    d. Experience working with imaging data: scale 1-5
    e. Experience working with DICOM: scale 1-5
    f. I have access to clinical imaging data at my institution/organization
    g. I can initiate new studies that involve collection of clinical data to fit my needs
    h. I have access to clinical experts to provide labeling
    i. …
2. What do I do with image data
    a. My typical project requires: select all that apply [images, manual expert annotations of the regions of interest, any annotations of regions of interest as long as I am confident in the results, image-derived features, clinical data (demographics, treatment, diseases, etc), data from other non-imaging domains (e.g., genomics)]
    b. Did you use public imaging data in your last project? Yes/No
    c. Ballpark how many of your past projects used public data? <select from ranges>
    d. What characteristics of a public data repository did you use or found useful: select all that apply (or arrange in the order of priorities, if Google Forms has this feature): [interactive portal to search and subset the data, ability to visualize images in the portal, ability to visualize annotations, API for programmatic access to the search interface, rich metadata for data selection, interactive (GUI) interface to download the data, programmatic (API/batch/command line) tools to download the data]

e. Thinking of your recent project that utilized public data, what did you do with the data as part of your project: [prepared expert annotation, applied existing AI tools, extracted radiomics features, tested the algorithm I am developing, other]
f. Do you use any of the following in your work: check all that apply [cloud-based virtual machines, AWS, Google Cloud Platform, Google Colab, AWS Sage Maker, …]

3. Where do I go to find public image data
a. Which sources of public data you know about: TCIA, ADNI, Zenodo, Dataverse, Figshare, IDC, …
b. Which sources of public data did you use in the past: TCIA, ADNI, IDC, datasets referenced in the papers I read…
c. …

4. Sharing of Publishing of the data
a. Have you shared any of the analysis results you generated for the existing public datasets you analyzed? [yes/no]
b. Have you shared any of the images from your institution publicly? Yes/No - Skip if the rest of the section if the answer is No
c. Are you comfortable and are you allowed to share images in DICOM format that you de-identified at your institution publicly? [yes; yes, but I would prefer to have de-id confirmed by an entity such as TCIA; no, I need de-id assistance by an entity such as TCIA]
d. Did you share your dataset as public or restricted access? Public/Restricted
e. What mechanism did you use for sharing? [TCIA, Dropbox/Google drive or similar, General-purpose repositories (e.g., Zenodo, Figshare), institutional repository, your choice]
f. What are the applicable mechanisms that are important for you to get credit for sharing your data: [peer-reviewed manuscript about the data, citations of my dataset, citations of my manuscript about the dataset, I share data because it is important and getting credit is secondary me]

Relevant materials to consider

TCIA paper citations:
https://scholar.google.com/scholar?cites=1657363728974225254&as_sdt=40000005&sciodt=0,22&hl=en

Woznicki, P., Laqua, F. C., Al-Haj, A., Bley, T. & Baeßler, B. Addressing challenges in radiomics research: systematic review and repository of open-access cancer imaging datasets. Insights Imaging 14, 1–13 (2023).

Dulaney, A. & Virostko, J. Disparities in the demographic composition of the cancer imaging archive. *Radiol. Imaging Cancer* **6,** (2024). - some discussion of the uses of data from TCIA

TCGA paper: 6.3K citations: Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013 Oct;45(10):1113-20.

An inventory of repositories … ⊞ TF3.2_perfusion_repositories_v00

Data collected by one of the IDC user studies

Study Demographics

Nine users who had a need for imaging data sets participated in this usability assessment. Users were not required to have prior experience with the IDC to participate, though a couple had interacted with the site prior to this session.

Most had used similar repositories like TCIA or local data repositories accessible through their university systems. Over half of the users in this study had data science expertise (as represented by Kat) and less than half did not have a data science background (as represented by Robert). See the table below for a full breakdown.

Occupation	• Cancer Biologist = 1 • Clinician/Scientist = 2 • Biomedical Engineer = 1 • Medical Physics Researcher/Medical Physicist = 2 • Software / Tool Developer = 2 • Image Analysis = 1
Role	• Post Doc / Staff Scientist = 1 • Principal Investigator = 6 • Software Developer = 2
Education	• Graduate degree (masters or PhD) = 7 • Post Baccalaureate Experience (before Masters) = 1 • Professional degree (MD, DDS, JD, etc.) = 1
Data Comfort	• Intermediate = 6 • Expert = 3
IDC Prior Use	• Have never heard of it = 1 • Heard of it, but not used = 6 • I have tried it/played around with it = 1 • Used it a few times = 1
IDC Attitude (if used) *Net Promotor Score = 0 (on a scale from -100 to 100)*	• Promotor (9 or 10 on scale) = 1 • Passive (6 to 8 on scale) = 1

App.dimensions.com is useful in finding citations and looking at the dynamics