35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

TCIA: An Information Resource to Enable Open Science*

Fred W. Prior, *Senior Member, IEEE*, Ken Clark, Paul Commean, John Freymann, Carl Jaffe, Justin Kirby, Stephen Moore, *Member, IEEE,* Kirk Smith, Lawrence Tarbox, Bruce Vendt, and Guillermo Marquez

***Abstract*—Reusable, publicly available data is a pillar of open science. The Cancer Imaging Archive (TCIA) is an open image archive service supporting cancer research. TCIA collects, de-identifies, curates and manages rich collections of oncology image data. Image data sets have been contributed by 28 institutions and additional image collections are underway. Since June of 2011, more than 2,000 users have registered to search and access data from this freely available resource. TCIA encourages and supports cancer-related open science communities by hosting and managing the image archive, providing project wiki space and searchable metadata repositories. The success of TCIA is measured by the number of active research projects it enables (>40) and the number of scientific publications and presentations that are produced using data from TCIA collections (39).**

I. INTRODUCTION

The volume of scientific data doubles each year with single experiments now generating petabytes of data annually [1]. Data-driven research and decision-making, though broadly recognized as critical, suffer a gap between potential and realization due, in part, to the challenge of effectively managing the exploding volume of data [2, 3].

NIH research funding for genomics and medical imaging, two Big Data disciplines, has shifted to a paradigm supporting large public databases and encouraging funded researchers to publicly share their data in hopes of using open-data to stimulate open-science collaboration. Genomics has spawned numerous knowledge-sharing databases (model organisms, nucleotide, protein, structure, taxonomy) [4, 5]. Imaging projects such as the Bioinformatics Research Network (BIRN) [6] and recently the Human Connectome Project [7] are accumulating vast amounts of image data in order to accelerate our understanding of brain structure and function and have firmly established medical imaging in the realm of Big Data based science. In cancer imaging, the National Cancer Institute (NCI) has funded The Cancer Imaging Archive (TCIA), described here, as a public repository of cancer images and related clinical data for the express purpose of enabling open science research [8].

II. OPEN SCIENCE AND OPEN DATA

The concept of open science is perhaps most generally assumed to mean the free sharing of tools, data and results among scientists; a process that began with the Renaissance. In more recent literature the term *open science* has become somewhat nebulous and has been used to encompass a wide variety of concepts [9, 10] including:

- Using Open Source software in scientific research;
- Making data and tools available to the public to enhance basic science education;
- Making scientific results available in Open Access journals;
- Finding innovative solutions to scientific problems via crowd sourcing;
- Using Open Source software to capture and manage Open Data to encourage and support research and education;
- Creating Research Communities around an Open Data resource.

TCIA utilizes open source software to create and support research communities around an open access information resource. TCIA data was originally collected for clinical diagnosis or a specific research project but is now being offered to the research community to enable new lines of research.

*Research supported by the National Cancer Institute under Contract No. HHSN261200800001E, and Washington University subcontract 10XS220.

Fred Prior is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (phone: 314-747-0331; fax: 314-362-6971; e-mail: priorf@mir.wustl.edu).

Ken Clark is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: clarkk@mir.wustl.edu).

Paul Commean is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: commeanp@mir.wustl.edu).

John Freymann is with SAIC-Frederick, Inc., Frederick, MD 21702 USA (e-mail: freymannj@mail.nih.gov).

Carl Jaffe is with the Department of Radiology, Boston University School of Medicine, Boston, MA USA (e-mail: carljaffe@gmail.com).

Justin Kirby is with SAIC-Frederick, Inc., Frederick, MD 21702 USA (e-mail: kirbyju@mail.nih.gov).

Stephen Moore is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: moores@mir.wustl.edu).

Kirk Smith is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: smithki@mir.wustl.edu).

Lawrence Tarbox is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: tarboxl@mir.wustl.edu).

Bruce Vendt is with the Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: vendtb@mir.wustl.edu).

Guillermo Marquez is with the National Cancer Institute, Bethesda, MD 20892 USA (e-mail: marquezg@mail.nih.gov).

III. THE CANCER IMAGING ARCHIVE (TCIA)

A. TCIA's Multi-component Architecture

Figure 1 illustrates the various ways by which images and non-image data are added to TCIA, stored in TCIA, and harvested from TCIA. Images may be provided as completed collections from ongoing supported research projects or from completed clinical trials. Inbound images, de-identified at their contributing source, are deposited with an intake server until they have been curated, after which they are placed with TCIA's public server, either among general-access (fully public) collections or among limited-access collections, with placement determined by NCI. While most collections are publicly available, about 5% are limited-access for groups of investigators needing to share images but not quite ready to release their images to the public. Image metadata may be extracted from the images and deposited with clinical-trial non-image data, in the TCIA Clinical Data and Metadata Repository. Some image collections arrive with annotation and markup objects either in DICOM format [11] or study specific format [12]. Ongoing research projects may add annotations, created by Annotation and Image Markup (AIM) compliant applications [13], to the TCIA Annotation Repository [14] or project metadata to the TCIA wiki. All users have read-access to the Public Image Repository, the Annotation Repository, wiki, and the Clinical Data and Metadata Repository. Users with project specific privileges, including those connected with supported research projects, may harvest images and data from the limited-access portions of the repositories and wiki and contribute (write privileges) to the Annotation Repository and the wiki.

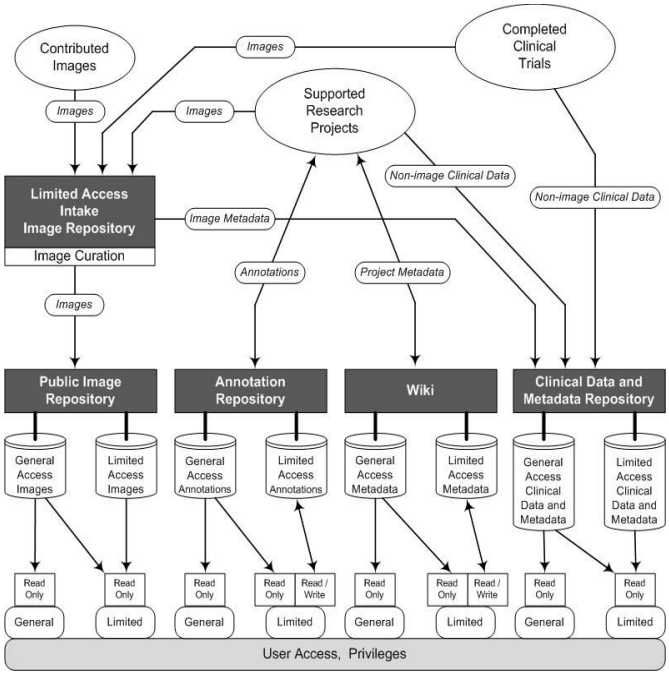

Figure 1. TCIA collects multiple types of de-identified data documenting supported research projects and completed clinical trials, and makes these data available to enable on-going research.

B. Contributed Images

TCIA is a managed archive of contributed radiology images of cancer in DICOM format. TCIA supports the de-identification, submission and curation of image data so that they can be made publicly available in a HIPAA compliant form while maximizing their scientific value. Image data are de-identified with open-source software [15] configured and provided to the contributor for the transmission of images to TCIA's intake server. Arriving images are visually inspected for image corruption and visible protected health information (PHI), while image headers are automatically scanned for potential PHI. Preparation of de-identification scripts tailored to individual image collections and in-coming image quality control require significant effort and attention to detail. These efforts are essential, however, to high-quality curation [3], the activity of organizing biological information such that they are easily digestible by both humans and their computers. Upon such effort rests the efficient proof or disproof of hypotheses put forward by image-consuming researchers in hopes of biological discovery.

TCIA groups images into collections. A collection typically includes studies (groups of images and associated study data) from several human subjects. In some collections, there may be only one study per subject. In other collections, subjects may have been followed over time, in which case there will be multiple studies per subject. The subjects typically have in common a particular disease and/or particular anatomical site (e.g., lung, brain). Collections are labeled so that a TCIA user can easily identify the related research project and cancer type (e.g., TCGA-GBM) or imaging modality and anatomy imaged (e.g., Prostate-MRI).

C. Image Retrieval

Images passing quality control are posted to a public server from which anyone with a TCIA account (free) may view and download images. The primary image management application is the open-source National Biomedical Imaging Archive (NBIA) [16]. NBIA presents the user with over ninety DICOM tags upon which to refine queries on the image data. Once an investigator has selected desired images, the images may be downloaded immediately or the investigator may save links to the images as a *shared list*; a list of image series stored in the NBIA database. The investigator may recall a *shared list* at any future time and download the associated images. The investigator may also inform collaborators who could then log into NBIA and access the specified *shared list* in order to download the same image set, thus enabling the collaboration with a simple mechanism for sharing images.

How does a researcher know what data are relevant to his research and how does one search for these data? Typically, one would be directed to the TCIA home page to find "For Researchers," specific links for: gaining access to the images, image collections, related publications, and research projects. The *how* of searching is well described in the TCIA User Guide, available from the main system menu.

A public TCIA wiki space provides detailed information for most collections. Multi-site collections include links to the project in which the providers are participating. As users

enquire about certain kinds of images, the answers are captured on a public-faced wiki page. The wiki gives data contributors a platform to describe the scope and intent of their image collection and to provide metadata and/or ways for users to contact them. The wiki supports research groups by summarizing the work of participants and posting conference abstracts and publications. The public space also provides access to user guides.

The TCIA Support Center services users via email and direct links from the TCIA web site. All user issues are documented and tracked using an open-source trouble-ticket program for problems in these areas: (1) normal user questions concerning account creation and credentialing, (2) use of the NBIA application, (3) direction to documentation on the collections.

IV. TCIA DATA COLLECTIONS

An NCI Cancer Imaging Program advisory group prioritizes new TCIA image collection candidates based on the extent to which the data comply with the following objectives:

- NCI grant / contract award data sharing requirements;
- Analysis of imaging features to be used as biomarkers;
- Creation of correlative signatures for multi-platform biomarkers;
- Creation of algorithms for detection of cancer;
- Testing and validating quantitative analysis techniques;
- Unique characteristics for clinical training.

TCIA image collections represent cancers affecting a variety of organs (brain, breast, head/neck, lung, colon, prostate, kidney) from a variety of imaging modalities (computed tomography, magnetic resonance, mammography, X-ray, positron-emission tomography, radiation treatment planning). There are also a few phantom collections available for algorithm and measurement process verification. Image collections are typically from completed studies, as TCIA does not manage ongoing clinical trials. Table 1 summarizes the number of images (e.g. single CT axial slice) in the TCIA image collections by anatomy and imaging modality. It includes over 20 million chest CT images belonging to the limited-access National Lung Screening Trial (NLST) [17, 18] collection.

Most collections have associated clinical and/or image metadata, which can be accessed via TCIA wiki pages. The NLST collection utilizes a Query Tool that allows an investigator to pose user-created queries against the non-image data collected during the trial and trial results (e.g., demographics, image-screening results, smoking history, medical history, work history, cancer diagnosis and tracking) and/or the imaging data extracted from the DICOM header (e.g., study year, kVp, mAs, pitch, series description, series instance UID). Once satisfied with the results of a query, the results can be saved to a text file, and/or a *shared list*, or the images can be downloaded from TCIA. In addition, the queries may be saved for later recall or for finer tuning.

While the Query Tool was developed with NLST data, it is now being deployed for use with other research groups with TCIA images and associated non-image data, thus allowing researchers to query non-image data and, among other things, choose images for downloading by invoking the TCIA image-download function from the Query Tool.

Table 1. TCIA image collections by anatomic region (number of images for each imaging modality and the types of cancer imaged).

Anatomic Region	DX	CT	MR	PT	Cancer Type(s)
Brain		6,482	959,401		Glioma, Glioblastoma Multiforme
Breast		6,980	257,062	5,492	Breast Invasive Carcinoma
Lung/ Chest	569	21,424,099		123,744	Adenocarcinoma, Squamous cell carcinoma, Bronchioloalveolar carcinoma, Large-cell carcinoma, non-small-cell carcinoma, small cell carcinoma, carcinoid
Colon		941,771			Adenocarcenoma
Head/Neck		83,915		118,133	Squamous cell carcinoma
Kidney		50,852	25,630		Renal Clear cell carcinoma
Prostate		13,534	81,132	27,870	Adenocarcinoma
Imaging Modalities: **DX** -digital x-ray, **CT** - Computed Tomography, **MR** - Magnetic Resonance Imaging, **PT** - Positron Emission Tomography					

V. TCIA ENABLED RESEARCH

As an open-access archive linked to extensive meta-data, cross-disciplinary researchers can use TCIA to test biomedical hypotheses and develop analytic techniques. TCIA provides the international research community with free access to imaging data sets that have in the past been prohibitively costly or impossible to generate. Cancer researchers can use these data to test new hypotheses and develop new analysis techniques to advance the scientific understanding of cancer. Engineers and software developers can build new analysis tools and techniques using this data as test material for developing and validating algorithms. Educators can use it as a teaching tool for introducing students to medical imaging technology and cancer phenotypes. In addition, a number of active research communities have developed around specific TCIA collections. Table 2 lists the currently active communities and the associated TCIA collections.

TCIA is actively developing collections of image data from cases where genomic, clinical and histopathology data are available on The Cancer Genome Atlas [5] website, providing a unique resource for researchers in the relatively new field of imaging phenotype to genotype analysis. TCGA researchers are collecting tissue samples (brain, breast, gastrointestinal, head and neck, hematologic, skin, thoracic, and urologic) and are mapping the genetic changes in 20 cancers. The TCGA Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA while associated radiology images are available through TCIA. To date 16 active research projects are on going based on the data available from the TCGA Data Portal and TCIA. TCIA enabled researchers are advancing the use of image and genomics data in the fight against breast, brain, lung and renal cancers [19-21].

The Quantitative Imaging Network (QIN)[16] has contributed brain, breast, head-neck, and prostate cancer

images. More than 16 active QIN research projects utilize TCIA data and many of these projects maintain limited access collections on TCIA to support the development and validation of quantitative imaging-derived biomarkers.

Table 2. Collaborative research groups that are enabled by the TCIA resource.

Community	Collaborative Projects	Active Researchers	TCIA Collections Utilized
TCGA Glioma Phenotype Group	11	>20	TCGA-GBM TCGA-LGG
TCGA Breast Phenotype Group	4	>12	TCGA-BRCA
TCGA Renal Phenotype Group	1	>13	TCGA-KIRC
Quantitative Imaging Network	>16	>190	QIN Breast QIN Phantom QIN Lung QIN Prostate
National Lung Screening Trial Related Groups	8	25	NLST

The National Lung Screening Trial was a decade-long multi-center trial to determine whether screening for lung cancer with low-dose helical computed tomography (CT) reduces mortality from lung cancer in high-risk individuals relative to screening with chest radiography. Approximately 54,000 participants were enrolled between August 2002 and April 2004. The primary outcome of the trial was the finding that lung cancer mortality was reduced by 20% in the CT arm of the trial [18]. This extensive data set is now available as a limited access collection with access permission granted by NCI [22]. Eight research groups are currently utilizing this resource.

A key metric of the value of TCIA is the dissemination of scientific research results that rely on the TCIA resource. Since the Cancer Imaging Archive went on-line in 2011, TCIA enabled research initiatives have produced 6 peer reviewed publications (with more in review) and 33 scientific presentations [23] with more in preparation as the work is ongoing and new projects and collections are continually being added.

VI. CONCLUSIONS

The Cancer Imaging Archive is an investment in Open Science by the National Cancer Institute and allows Open Access to cancer images, trial data, and mechanisms for collaborative research. TCIA is not primarily technology focused but rather a service, designed to give access to image collections to the broadest possible research community. Open Science initiatives such as TCIA are producing substantial scientific impact. Open science communities have formed around TCIA data collections and are gaining traction as evidenced by a steadily increasing output of abstracts, presentations and publications.

REFERENCES

[1] A. Szalay and J. Gray, "2020 Computing: Science in an exponential world," *Nature,* vol. 440, pp. 413-414, 2006.

[2] C. Lynch, "Big data: How do your data grow?," *Nature,* vol. 455, pp. 28-29, 2008.

[3] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, *et al.*, "Big data: The future of biocuration," *Nature,* vol. 455, pp. 47-50, 2008.

[4] E. Birney, A. Bateman, M. E. Clamp, and T. J. Hubbard, "Mining the draft human genome," *Nature,* vol. 409, pp. 827-828, 2001.

[5] T. Hampton, "Cancer Genome Atlas," *JAMA: The Journal of the American Medical Association,* vol. 296, pp. 1958-1958, 2006.

[6] J. S. Grethe, C. Baru, A. Gupta, M. James, B. Ludaescher, M. E. Martone, *et al.*, "Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease," *Studies in health technology and informatics,* vol. 112, pp. 100-110, 2005.

[7] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, *et al.*, "The human connectome project: a data acquisition perspective," *Neuroimage,* 2012.

[8] C. C. Jaffe, "Imaging and Genomics: Is There a Synergy?," *Radiology,* vol. 264, pp. 329-331, August 2012 2012.

[9] M. Woelfle, P. Olliaro, and M. H. Todd, "Open science is a research accelerator," *Nat Chem,* vol. 3, pp. 745-748, 2011.

[10] J. C. Molloy, "The open knowledge foundation: open data means better science," *PLoS Biology,* vol. 9, p. e1001195, 2011.

[11] D. A. Clunie, "DICOM structured reporting and cancer clinical trials results," *Cancer informatics,* vol. 4, p. 33, 2007.

[12] M. F. McNitt-Gray, S. G. Armato III, C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, *et al.*, "The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation," *Academic Radiology,* vol. 14, p. 1464, 2007.

[13] D. S. Channin, P. Mongkolwat, V. Kleper, K. Sepukar, and D. L. Rubin, "The cabig™ annotation and image markup project," *Journal of Digital Imaging,* vol. 23, pp. 217-225, 2010.

[14] F. Wang, T. Pan, A. Sharma, and J. Saltz, "Managing and querying image annotation and markup in XML," in *Proceedings of SPIE*, 2010, p. 762805.

[15] J. Freymann, J. Kirby, J. Perry, D. Clunie, and C. Jaffe, "Image Data Sharing for Biomedical Research - Meeting HIPAA Requirements for De-identification," *Journal of Digital Imaging,* pp. 1-11, 2011.

[16] L. P. Clarke, B. S. Croft, R. Nordstrom, H. Zhang, G. Kelloff, and J. Tatum, "Quantitative imaging for evaluation of response to cancer therapy," *Translational Oncology,* vol. 2, p. 195, 2009.

[17] K. Clark, D. Gierada, G. Marquez, S. Moore, D. Maffitt, J. Moulton, *et al.*, "Collecting 48,00 CT Exams for the Lung Screening Study of the National Lung Screening Trial," *Journal of Digital Imaging,* vol. 22, pp. 667-680, Decmeber 2009 2009.

[18] D. Aberle, A. Adams, C. Berg, W. Black, J. Clapp, R. Fagerstrom, *et al.*, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *The New England journal of medicine,* vol. 365, p. 395, 2011.

[19] R. Jain, L. Poisson, J. Narang, D. Gutman, L. Scarpace, S. N. Hwang, *et al.*, "Genomic Mapping and Survival Prediction in Glioblastoma: Molecular Subclassification Strengthened by Hemodynamic Imaging Biomarkers," *Radiology,* 2012.

[20] P. O. Zinn, P. Sathyan, B. Mahajan, J. Bruyere, M. Hegi, S. Majumder, *et al.*, "A Novel Volume-Age-KPS (VAK) Glioblastoma Classification Identifies a Prognostic Cognate microRNA-Gene Signature," *PloS one,* vol. 7, p. e41522, 2012.

[21] P. O. Zinn, B. Majadan, P. Sathyan, S. K. Singh, S. Majumder, F. A. Jolesz, *et al.*, "Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme," *PloS one,* vol. 6, p. e25451, 2011.

[22] NCI. (2013, February 3, 2013). *CDAS Cancer Data Access System*. Available: https://biometry.nci.nih.gov/cdas/

[23] TCIA. (2013, January 18, 2013). *For Researchers; Related Publications*. Available: http://cancerimagingarchive.net/publications.html