Introduction to Special Issue on Datasets hosted in
The Cancer Imaging Archive (TCIA)

Justin Kirby[1], Fred Prior[2], Nicholas Petrick[3], Lubomir Hadjiski[4], Keyvan Farahani[5], Karen Drukker[6], Jayashree Kalpathy-Cramer[7], Carri Glide-Hurst[8], Issam El Naqa[9*]

[1]Frederick National Laboratory for Cancer Research, Cancer Imaging Informatics Lab, National Institute of Health, Frederick, MD 2170, USA

[2]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock AR 72205, USA

[3]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA

[4]Department of Radiology, University of Michigan, Ann Arbor, MI 48109, USA

[5]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD 20814, USA

[6]Department of Radiology, Chicago, IL 60637, USA.

[7]Department of Radiology, Charlestown, MA 02129, USA

[8]Department of Radiation Oncology, University of Wisconsin, Madison, WI 53792, USA

[9]Department of Machine Learning, Moffitt Cancer Center, Tampa, FL 33612, USA

Corresponding author email:
*Issam.elnaqa@moffitt.org

No conflict of interest associated with this publication.

Introduction

Public datasets play a key role in enabling the medical research community to validate and build upon each other works using data acquired outside of their home institutions. This is especially critical for stimulating studies utilizing quantitative data analysis (radiomics) or artificial intelligence/machine learning (AI/ML) approaches, for which validation and generalizability on independent external cohorts are essential for acceptance and future clinical translation. Recognizing this fact, the Journal of Medical Physics has introduced a new category of article submissions known as *Medical Physics Dataset Articles* (MPDAs)[1]. MPDAs help facilitate the use of valuable open-access datasets by granting authors the opportunity to publish detailed scientific or clinical descriptions of their data with unique digital object identifiers (DOIs) for future citations. Unlike traditional manuscripts, these articles would focus on reproducibility and the dataset's potential use cases and details of how it was acquired, curated, and published.

This special issue was organized in partnership with The Cancer Imaging Archive (TCIA)[2]. TCIA is an official image repository of the National Cancer Institute (NCI), and the preferred digital repository for sharing cancer-related datasets described by the MPDA readership[3]. Its mission is to provide proper de-identification and hosting services to relieve individual researchers of the legal and technical complexities of sharing patient datasets. Image datasets are organized as "collections"; typically focused on a common disease (e.g., lung cancer), image modality (MRI, CT, digital histopathology, etc.) or research focus (e.g., quantitative imaging). TCIA is currently home to 126 datasets[4] collected as part of numerous NCI-funded clinical trials and data sharing initiatives[5,6] as well as datasets proposed by investigators in the broader research community[7].

In many cases the submitter(s) of TCIA datasets may include radiology or pathology annotations, image classifications, segmentations, radiomics features, or derived/reprocessed images. However, there are often cases where those who access the data on TCIA may perform their own analyses, which can result in additional image labels. In order to further support the enrichment of existing datasets with these additional labels, TCIA has begun accepting proposals for third party "Analysis Results" based on existing image collections. Sharing such analyses is critical, not only to enhancing medical studies reproducibility and reusability, but also to providing significant value to the data science community in the form of labeled image sets for training new AI/ML algorithms and other automated analysis approaches. Currently TCIA contains 28 such datasets[8] several of which were submitted in relation to this special issue.

The aim of this special issue is to highlight valuable examples of MPDAs and publicly available datasets that can be reused for future research endeavors and utilized for addressing emerging scientific or clinical questions.

In "Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations" by Cardenas, et al.[9] describe a T2-weighted MRI dataset of 55 head and neck cancer patients that can be used to evaluate the accuracy of auto-segmentation systems delineating organs at risk (OAR) through comparisons to expert manual segmentations. The dataset can further complement existing CT datasets, where MR soft tissue discrimination can further aid results for treatment planning, for instance.

In “FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, Interobserver, Lung1 and Head-Neck1 TCIA collections” by Kalendralis, et al.[10] describe updated clinical data, radiomics features and Digital Imaging and Communications in Medicine (DICOM) headers from 4 datasets analyzed as part of their Nature Communications radiomics study[11] in order to support repeatability, reproducibility, generalizability and transparency in radiomics research, which can be used as useful benchmark for future CT radiomics studies.

In “DICOM Re-encoding of Volumetrically Annotated Lung Imaging Data Consortium (LIDC) Nodules” by Fedorov, et al.[12] describe annotations for lung nodules from 875 of the subjects collected by the Lung Imaging Data Consortium and Image Database Resource Initiative (LIDC) converted into standard DICOM objects to simplify reuse of the data with the readily available open-source tools, and to improve adherence to FAIR (Findable, Accessible, Interoperable, Reusable) principles[13].

In “PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines” by Kiser, et al.[14] describe a dataset of thoracic cavity segmentations and discrete pleural effusion segmentations annotated on 402 CT scans acquired from patients with non-small cell lung cancer (NSCLC). These data can be used for developing image analysis pipelines such as lung structure segmentation, lesion detection, and radiomics feature extraction. Combining these pleural effusion segmentations with the gross tumor volume segmentations already available from the “NSCLC Radiomics” dataset, which will also enable investigation of radiomics profile differences between effusion and primary tumors.

In “Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset” by Pati, et al.[15] describe their analyses and resulting data from 31 IvyGAP subjects including multi-institutional expert annotations for tumor sub-compartments, radiomic features, and the associated reproducibility meta-analysis to facilitate developing image-based biomarkers for prognostic/predictive applications in patients with glioblastoma.

In "CT images with expert manual contours of thoracic cancer for benchmarking auto-segmentation accuracy" by Yang, et al.[16] describe a well-curated computed tomography (CT) dataset of high-quality manually drawn contours from 60 patients with thoracic cancer that can be used to evaluate the accuracy of thoracic normal tissue auto-segmentation systems.

In "MRQy: An Open-Source Tool for Quality Control of MR Imaging Data" by Sadri, et al.[17] describe how they used MRQy, an open-source quality control tool to analyze TCIA collections with data that was submitted from multiple sites. The results can be used for: (a) interrogating MRI cohorts for site- or equipment-based differences, and (b) quantifying the impact of MRI artifacts on relative image quality. This information can help determine how to correct for these variations prior to model development and assess future harmonization techniques.

In summary, this special issue and its related datasets will serve as a valuable resource to help develop benchmarks for a wide variety of imaging applications including image processing, quality assurance, diagnostic, prognostic, and radiomics approaches using rich, annotated CT and/or MR datasets. This will further strengthen the value of these datasets, their utility and potential impact in the field of medical physics with the overarching goal of encouraging the creation of new public datasets through MPDA/TCIA and their dissemination in the field.

References

1. Williamson, J.F., Das, S.K., Goodsitt, M.S. and Deasy, J.O. (2017), Introducing the Medical Physics Dataset Article. Med. Phys., 44: 349-350. doi:10.1002/mp.12003
2. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013 Dec;26(6):1045-57. doi: 10.1007/s10278-013-9622-7. PMID: 23884657; PMCID: PMC3824915.
3. URL: https://www.aapm.org/pubs/MPJ/policies/details.asp?id=465&type=MP.

4. URL: https://www.cancerimagingarchive.net/collections/
5. URL: https://www.cancerimagingarchive.net/imaging-proteogenomics/
6. URL: https://wiki.cancerimagingarchive.net/x/BQHDAg
7. URL: https://www.cancerimagingarchive.net/primary-data/
8. URL: https://www.cancerimagingarchive.net/tcia-analysis-results/
9. Cardenas CE, Mohamed ASR, Yang J, Gooding M, Veeraraghavan H, Kalpathy-Cramer J, Ng SP, Ding Y, Wang J, Lai SY, Fuller CD, Sharp G. Head and neck cancer patient images for determining auto-segmentation accuracy in T2-weighted magnetic resonance imaging through expert manual segmentations. Med Phys. 2020 Jun;47(5):2317-2322. doi: 10.1002/mp.13942. PMID: 32418343; PMCID: PMC7322982.
10. Kalendralis P, Shi Z, Traverso A, Choudhury A, Sloep M, Zhovannik I, Starmans MPA, Grittner D, Feltens P, Monshouwer R, Klein S, Fijten R, Aerts H, Dekker A, van Soest J, Wee L. FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections. Med Phys. 2020 Jun 10. doi: 10.1002/mp.14322. Epub ahead of print. PMID: 32521049.
11. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun.434 2014;5(1):4006. doi:10.1038/ncomms5006
12. Fedorov, A., Hancock, M., Clunie, D., Brochhausen, M., Bona, J., Kirby, J., Freymann, J., Pieper, S., J. W. L. Aerts, H., Kikinis, R. and Prior, F. (2020), DICOM re-encoding of volumetrically annotated Lung Imaging Database Consortium (LIDC) nodules. Med. Phys. doi:10.1002/mp.14445
13. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18
14. Kiser, K.J., Ahmed, S., Stieb, S., Mohamed, A.S.R., Elhalawani, H., Park, P.Y.S., Doyle, N.S., Wang, B.J., Barman, A., Li, Z., Zheng, W.J., Fuller, C.D. and Giancardo, L. (2020), PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines. Med. Phys.. doi:10.1002/mp.14424
15. Pati, S., Verma, R., Akbari, H., Bilello, M., Hill, V.B., Sako, C., Correa, R., Beig, N., Venet, L., Thakur, S., Serai, P., Min Ha, S., Blake, G.D., Taki Shinohara, R., Tiwari, P.

and Bakas, S. (2020), Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset. Medical Physics. doi:10.1002/mp.14556

16. Yang, J., Veeraraghavan, H., van Elmpt, W., Dekker, A., Gooding, M. and Sharp, G. (2020), CT images with expert manual contours of thoracic cancer for benchmarking auto-segmentation accuracy. Med. Phys., 47: 3250-3255. doi:10.1002/mp.14107
17. Sadri, A.R., Janowczyk, A., Zhou, R., Verma, R., Beig, N., Antunes, J., Madabhushi, A., Tiwari, P. and Viswanath, S.E. (2020), Technical Note: MRQy — An open-source tool for quality control of MR imaging data, Medical Physics, doi: 10.1002/mp.14593.