Indexing Medical Image Collections in a Biomedical Data Fabric: A Vision for a Federated Cancer Imaging Data Repository

Contributing authors: Sarah Beth Bdoyan MSPH CPHQ, Alexander J. Towbin MD, Paul Kinahan PhD, Despina Kontos PhD, Mark Rosen MD PhD, Brian Bialecki CIIP, James Gimpel RT, Charles Apgar, BSE/MBA

Summary

Large centralized public data and imaging repositories have served the medical and research communities for decades. While artificial intelligence (AI) presents a new motivation for researchers to study this data, it has also exposed challenges associated with centralized public data approaches. These challenges require technical innovation to address the growing concerns related to topics such as patient privacy, institutional autonomy, and information security. Through work with the Advanced Research Projects Agency for Health (ARPA-H) Biomedical Data Fabric (BDF) Toolbox within the Medical Imaging and Data Resource Center (MIDRC), the American College of Radiology (ACR) assembled a group of medical imaging experts to review the attributes of existing centralized cancer repositories in terms of data quality, accessibility, and usefulness and to pilot one innovative solution using data from an NCI sponsored clinical trial. Specifically, we demonstrate the feasibility of creating and applying a methodology to index medical imaging studies collected by the ACR in support of EA1141 *Abbreviated breast MRI for screening women with dense breasts*. This approach can be extended to enable imaging studies to remain on premise at local institutions, while central storage of imaging metadata would provide researchers the ability to rapidly create cohorts for their unique study objectives.

Background

Well-known centralized cancer image repositories include The Cancer Imaging Archive (TCIA) as well as the Imaging Data Commons (IDC) . These public repositories have been proactive in their approaches to ensuring access to completely deidentified imaging data. Clinical trials sponsored by the National Cancer Institute (NCI) offer the broadest opportunities for later data use given the high standards for informed consent and the federal requirements for data sharing. In comparison, other data registries have encountered resistance from potential data contributing sites arising from multiple factors including: cost and effort to collect the data, concerns over exposure of patient protected health information (PHI), potential use of the data for commercial purpose, and potential competing uses of the same data.

Public repositories arising from NIH-sponsored research generally represent highly curated data and are associated with definitive outcomes. Additionally, many of the studies included in these central repositories represent rare diseases and may represent some of the largest available collections of such patients. Many such repositories have been constructed to collect laboratory analytics, such as pathology and genetic assays, and other modes of imaging, such as medical photography or endoscopic video. These multi-specialty, multi-modality repositories offer the potential to enable precision medicine through advanced integrated diagnostic assessment using techniques such as radiomics, pathomics, and genomics.

While central data repositories have been successfully created, multiple factors have limited their impact. Namely, only a small percentage of clinical trials with published primary endpoint data have a corresponding set of imaging studies available in the public domain. Key challenges associated with public repositories of medical images include: (1) length of time to data availability for research use, (2) limited volumes of studies, (3) limited control over the quality of imaging, (4) lack of harmonization/standardization, (5) limited scope of imaging studies, (6) lack of annotation, (7) absence of digital imaging from other medical specialties (e.g. pathology), and (8) multiple technical limitations to the datasets themselves that reduce utility and limit use case applicability. Additionally, many repositories were created before the advent of artificial intelligence (AI). Thus, the data is not optimized for AI development and validation.

1. Length of time to Data Availability

Historically, imaging studies collected through the NCI's National Clinical Trials Network (NCTN) are restricted for many years before becoming available to the public. Standard practice is to embargo imaging studies and associated data from public access until after the primary aim of the trial has been published. This process allows the researchers engaged on the trial to perform their intended analysis and publish their findings, before other researchers can access the data. Depending upon the trial's endpoint and the length of time required to collect outcomes data, access to imaging studies may be restricted for several years. Once the data is approved for release, deidentification of imaging studies adds additional time before access can be permitted. Compounding the situation is the dynamic nature of medical imaging, where advances in methods and technology can lead to significant evolution over relatively short timeframes. Such rapid timeframes in technical advances and AI development can directly impact the relevance of research hypotheses and the future utility of collected data. Specifically, if data takes years to become accessible, it runs the risk of becoming outdated and irrelevant. For example, researchers conducting the National Lung Screening Trial (NLST) collected a massive and highly curated dataset on the lung cancer screening population. Public release of the data was not possible until 10 years after the study completed enrollment. Thus, at the time of data release, imaging data may have already matured 15 years since acquisition. During that timeframe major changes in imaging technology have occurred. These changes have included differences in imaging parameters, differences in image reconstruction technique, and differences in scanner technology thereby limiting the usefulness for some AI opportunities. Understanding that AI developed on CT scans which are no longer considered to be standard of care could be suboptimal, the NCI has charged the ACR with the creation of a contemporary lung cancer screening dataset. This contemporary dataset is expected to positively impact the development and validation of AI for use in lung cancer screening. While this is a laudable goal, barriers to success continue to challenge the research team for the same reasons outlined previously; cost and effort at the site, patient privacy and deidentification, and potential commercial benefit.

2. Limited Volume of Studies

Today, most AI algorithms require large, diverse datasets for development. Many public repositories struggle to amass sufficiently large volumes of imaging studies and lack sufficient consistency to easily aggregate data across trials as evidenced through the varying definitions of data elements across trials and lack of consistent utilization of data harmonization standards. For instance, glioblastoma is an aggressive malignant brain tumor that affects adults. Each year, more than 12,000 patients are diagnosed with this tumor in the United States . Glioblastoma is the most common type of brain tumor stored in TCIA. However, while this tumor is well represented in the repository, the data is limited to 9 datasets incorporating 1741 subjects (range from 39 to 630 subjects per dataset). As the dataset is segmented, it can become more limited. For example, when viewed by modality, all 9 datasets include MR, 6 had CT, and only one included PET imaging. Consequently, there are limited opportunities to validate AI using this data without supplementing the dataset from other sources.

3. Variable Quality of Imaging

NCI clinical trials promote highly rigorous data collection strategies and a high degree of data curation which ultimately results in a uniquely robust set of data for the specified aims of the protocol. New imaging modalities or new uses of established imaging methods are generally treated with the same level of rigor. However, most of the NCI-sponsored clinical trials call for the collection of standard-of-care imaging and invoke little to no additional measures to assure the quality of the imaging studies. In order to meet accrual goals, the NCTN groups seek to enroll patients from as many organizations as possible. This approach allows the group to study rare diseases and achieve targeted enrollment goals, but also results in considerable data heterogeneity. Specifically, the imaging data can vary in terms of vendor, modality, protocol, contrast bolus timing (where applicable), imaging planes, scan technique, etc. Such extreme heterogeneity creates challenges to discovering new insight. However, one benefit arising from such heterogeneity is that if an insight is made, it is likely to be generalizable.

4. Lack of Harmonization/Standardization

In an effort to promote harmonized approaches to imaging, Radiologic imaging has, for over 30 years, adopted the Digital Imaging and Communications in Medicine (DICOM) standard. Other specialties have not yet fully adopted this standard, including imaging-rich specialties like ophthalmology, pathology, and dermatology. The use of the DICOM standard has led to harmonization of imaging study header content and directly supported analytics of aggregated data sets. Yet there remains tremendous variability within various header elements, some of which are as fundamental as identification of the type of study. Such variability can lead to errors when cohorts are created and may influence the performance of an AI algorithm under development and testing.

As an example of the need for harmonization, the ACR examined a sample of 5,815 DICOM studies from various facilities that participated in EA1141 *Abbreviated breast MRI for screening women with dense breasts*, a clinical trial collecting only screening mammograms and Abbreviated MR. Out of the 5,815 studies, there were 289 unique procedure descriptions, and of the 73,450 series there were 1,419 unique series descriptions.

This variability creates additional burden for either the curator of the data commons or for the end user if not performed by the data commons itself, as the end user needs to ensure they are able to clearly discern uniformity or differences in the imaging study performed. An additional consideration is that some items are not in the DICOM header; for example, the type of contrast used could also impact the image quality and/or enhancement characteristics. The Logical Observation Identifiers Names and Codes (LOINC) provides a standard for procedure descriptions . Additional work occurring within MIDRC in the Data Quality and Harmonization subcommittee is currently focusing on making a MIDRC-LOINC

Mapping Table available for data ingestion pathways within the ARPA-H BDF projects. As this is completed, future work will focus on utilizing natural language processing and AI to normalize key pieces of metadata such as procedure description, series description, anatomic region sequence, and image orientation. The impact of this work will further enhance this group's efforts as it places the data commons at the forefront of harmonization, ensuring consistently implemented harmonization for datasets within a particular framework.

5. Limited Scope of Imaging Studies

Central repositories are typically static snapshots of the medical imaging that was collected in the context of the clinical trial. NCI clinical trial development efforts navigate a series of tradeoffs to create the most compelling scientific study while remaining within stringent budget limitations. Most trial designs impose constraints on image collection and analysis due to the limitations in funding and concerns over competing use of resources and effort at the participating institutions. For example, in the case of the Molecular Analysis for Therapy Choice (MATCH) trial, image collection was excluded initially. Over time, image collection was permitted on a voluntary basis and without any ability to query missing imaging. As a result, the MATCH trial has very limited value to the communities interested in advancing radiomics or imaging AI despite the robust genomic and clinical data. This limited value is compounded by the lack of image annotations to highlight tumor contours, lack of radiology reports, and the small number of subjects within each study arm.

Additionally, trials themselves suffer from a lack of imaging across various specialties. For example, patients with colon cancer may not and frequently do not have colonoscopy studies linked with their clinical data; and eye imaging is not present for patients presenting with orbital tumors. One obvious solution to this dilemma would be to expand the role of medical imaging and encourage inclusion on a broader scale. However, a more cost-effective approach to increase the volume of data available and improve data access is needed. One potential approach to solve this problem would be to register trial subjects in a database, allow local sites to store the clinical data, and enable central indexing of anonymized data so that the data could be identified for future retrieval and use. Such an approach could allow for access to a more complete imaging record without increasing the burden associated with image transfer, processing, and storage.

6. Lack of Annotation

Identifying datasets which will support AI development and validation is a difficult challenge. But once initial identification has been achieved, the AI developer is faced with additional steps to make the studies fully "AI Ready". "AI Ready studies" refer to studies which have undergone processing so as to make them ready for the AI developer to use; this generally entails some level of annotation to mark the clinically relevant region of interest. Yet the definition of fully annotated studies frequently varies based on the AI use case under consideration and can be debated among experts. For our purposes, we define fully annotated studies as those in which significant image features have been labeled and marked by a radiologist as appropriate to the AI use case. Annotation of the imaging studies is one of the most critical steps and typically requires extensive time and effort to complete this task. Creating annotated studies is typically not part of the funded effort within the clinical trials groups, and annotation of imaging data available in public repositories such as TCIA is inconsistent. As the field continues to expand and evolve, the very definition of annotations should be challenged and refined and will need to leverage innovative annotation methods such as transfer learning, self-supervised learning, or semi-supervised learning. Moving beyond traditional annotation methods and definitions will require additional infrastructure in centralized repositories. Extending these methods to enable performance at the local institution which acquired the imaging study will also be critical to the success of federated approaches.

7. Absence of Digital Pathology

Medical imaging AI relies heavily on key clinical data and reference standards, such as confirmation of the presence or absence of disease. Pathology slides have served as a highly valued source of such data, yielding important information about the cells in the tumor and in the surrounding space. Generating, processing and accessing these slides is cumbersome and costly. Some of this data can be extracted from reports or other clinical data which is also typically collected in the context of clinical trials, but represents burden on participating clinical trial sites and it may be even harder to locate and retrieve when developing real world data (RWD) registries. There are, however, significant opportunities to bridge this need by incorporating digital pathology within the context of clinical trials. The inclusion of digitized images of pathology slides is becoming increasingly routine, and the ability to move digital pathology images is already done easily by organizations such as the Imaging and Radiation Oncology Core (IROC) using ACR TRIAD. Within central repositories, digitized images of whole pathology slides generate large file sizes, which may prove problematic for central repositories and will certainly increase cost of retrieval and analytics. Therefore, it is important to note that digitized images of pathology slides do not necessarily refer to whole slide scanning and could refer to a secondary capture image, which is not as preferred but still has utility. Another current challenge in digital pathology is that pathology has not fully adopted DICOM as a standard. Instead, most sites use a proprietary file format, which creates additional challenge for both central repositories and AI development. While this discussion is limited to cancer, these concepts apply equally to other imaging use cases in cardiology, dermatology, ophthalmology, and obstetrics.

8. Technical Limitations

Central repositories also struggle with technical limitations that may limit the use cases that can benefit from access to the data and, by extension, directly affects the utility of the datasets for AI development. Such limitations include data labeling, data compatibility, and interoperability of repositories. Data labeling may be lacking (such as the case with contrast agent use not being included in the DICOM header) or may be applied inconsistently (such as patient age assignment). Data compatibility issues may arise when changing file format of images, such as during conversion of DICOM to Neuroimaging Informatics Technology Initiative format (NIfTI), or archiving DICOM Grayscale Softcopy Presentation State (GSPS) annotation objects to central storage. And, while Radiology is fortunate to have a multitude of imaging storage, presentation and analytic platforms to choose from, there may be challenges to integrate graphics processing units (GPU) with repositories and other analytic tools in order to apply AI. Without the technical infrastructure to make facile connections across repositories, the research able to be conducted with the available datasets remains limited: an especially painful challenge in a world focusing more and more on multi-modal data for improved patient outcomes and intelligent care pathways.

The MIDRC Initiative: Medical Image Indexing

The Medical Imaging and Data Resource Center (MIDRC), through funding from ARPA-H, has initiated a process by which a reference library of medical imaging studies could be generated across multiple central repositories using a common indexing schema. In theory, this process could lead to easier identification and access to imaging studies which could be sourced from multiple repositories and used on a common AI use case.

The MIDRC medical image indexing methodology captures key data elements from the DICOM header and populates a searchable index. Data elements include study ID, series ID, anatomy imaged, modality, and associated repository (e.g.: TCIA, IDC, MIDRC, etc.).

In consideration of the challenges and limitations noted previously with NCI clinical trial data repositories, ACR conducted a pilot study which extended the index model to include additional data elements which our working group believed would add important additional context for a researcher looking to create specific patient cohorts. We have applied the indexing methodology to a clinical trial performed under the NCI's clinical trials program (ECOG-ACRIN (EA) 1141- *Comparison of Abbreviated Breast MRI and Digital Breast Tomosynthesis in Breast Cancer Screening in Women with Dense Breasts*) and then harmonized Study ID terminology using ACRCommon (see Figure 1).

Level	Data Element
Program	Name
Project	Organ
	Focus Area
	Name
Facility	Name - Internal Attribute PHI
	ID
	Location - (State)
Patient	ACR ID
	Datavant ID
	Age
	Age Factor - (Years, Months or Days)
	Sex
	Race
	Ethnicity
	[Space within Schema for Future Use Case Tags]
Imaging Study	Study UID
	Accession Number - Internal Attribute PHI
	Year
	Patient Year of Birth
	Patient Sex
	Study Description
	[Space within Schema for Future Use Case Tags]
[Image series] – Common Elements – CT, MR, Nuclear	Manufacturer
	Manufacturer Model
	Software Version
	Modality
	[Sop Class]

	Body part Examined
	Contrast (y/n)
	Series UID
	Series Description
	[Space within Schema for Future Use Case Tags]
[MR Series] - Unique Elements	Magnetic Field Strength
	Receive Coil Name
	Transmit Coil Name
	[Slice Thickness]
	[Scanning Sequence]
	[Sequence Variant]
	[Scan Options]
	[Acquisition Type]
	[Space within Schema for Future Use Case Tags]
Nuclear Series – Unique Elements	Radiopharmaceutical

Figure 1: Imaging study indexing schema applied to EA1141

As part of this project, a fully deidentified version of the resulting index will be provided to MIDRC as a proof of concept supporting the development of the infrastructure which represents another “node” in the medical image index library which represents NCI Clinical Trials. The NCI clinical trial imaging index process could provide additional unique benefit to the research community. First, it allows for the identification of imaging studies which are not yet published in public repositories such as TCIA, allowing researchers the opportunity to pursue earlier access to these imaging studies. Secondly, the methodology could be extended to be performed at the imaging site and applied to additional imaging studies being performed on the same subject, thereby expanding the library of identified imaging studies to include imaging performed but not submitted to the sponsoring clinical trial group. Access to such images, in consideration of local consent processes and IRB rules, may be difficult to achieve if the images were to be collected. The ACR is therefore developing our informatics platform to enable AI algorithms to be pushed to participating imaging centers where the algorithm can then be run locally against the imaging studies and the findings can be collected and aggregated.

Image Indexing Methodology Process

ACR's indexing methodology was developed based on the sample use case of EA1141 - *Comparison of Abbreviated Breast MRI and Digital Breast Tomosynthesis in Breast Cancer Screening in Women with Dense Breasts,* a study completed through ACR's ECOG-ACRIN framework. Tools used to ingest the data and create the index included ACR's proprietary systems but the methodology could be hosted in other operating environments.

Through this process our team encountered multiple challenges. The primary challenge arises from the need to optimize the indexing characteristics so that it can be easily employed (suggesting selection of fewer variables) while still providing sufficient granularity to facilitate search and cohort creation (suggesting inclusion of more variables). The team also recognized the importance of capturing patient age and creating temporal links between data points (such as during COVID). Because date of birth (and other dates) are PHI, they are typically eliminated or altered to protect patient privacy. As a resolution to this problem, we collected two age points: age at time of imaging study and age at clinical trial enrollment. Our third major challenge stems from the degree of variability in DICOM header elements, particularly as it related to study description and imaging series description. Use of ACRCommon enabled dramatic reduction in variability and increase in the likelihood that users will correctly select imaging studies which should be included in the defined cohort. It is thus the recommendation of this group that future harmonization should take place even at the tagging and annotation level to provide true utility to the index. The data harmonization process supporting cohort selection project within MIDRC through implementation of the MIDRC-LOINC Mapping Table can further benefit this area and support this recommendation. Furthermore, the use of standardized nomenclature such as LOINC for procedure codes, SNOMED CT for anatomic terms, and other standards where applicable would further enhance the indexing approach.

Next Steps

As detailed in our project proposal to ARPA-H entitled "Harmonizing and Indexing Metadata Across Cancer Imaging Repositories: NCI Clinical Trial Program Image Indexing", the ACR will continue toward 2 objectives: (1) transfer knowledge gained to date to IROC leadership for further consideration within the NCI Clinical Trials program; and (2) apply lessons learned in this pilot to a real-world dataset.

This project was conducted as a pilot to demonstrate the ability to create a medical image indexing methodology and then apply that methodology in the NCI clinical trials operating environment. The project demonstrated that medical imaging studies being collected as part of the NCI clinical trials program can be indexed and therefore increase the discoverability of imaging studies being performed. This project also demonstrates the ability to link the indexed medical records to a national reference library being established by MIDRC, such that researchers could leverage one common gateway to identifying imaging studies that can meet their research or AI purposes.

Imaging study collection for the NCI clinical trials program is performed by the Imaging and Radiation Oncology Core (IROC) and uses TRIAD as the image collection platform. IROC is therefore ideally positioned to extend this pilot project to pragmatic uses which support the research objectives of the sponsoring clinical trials groups. This paper is being provided to IROC leadership with the request that they consider:

1. the potential to incorporate a medical imaging indexing methodology across IROC as part of the image collection process;
2. changes to the indexing method which would enhance practicality and value to users and the public; and
3. the potential to extend the indexing method to include imaging studies acquired by sites on the clinical trial subjects but not required to be collected as part of the clinical trial.

While it is informative to increase the transparency of medical imaging studies being performed on subjects enrolled on an NCI trial, the ultimate value can only be achieved if access to the studies can be achieved by the researcher/AI developer. That challenge is best managed by IROC and the leadership of the clinical trials groups in cooperation with sites themselves. Questions IROC might consider include: Can access to images collected on published clinical trials but not submitted to TCIA be made easier? Can access to images collected on active clinical trials be permitted to enable learning while the study is ongoing? Can access to images performed on clinical trial patients but NOT collected by IROC be permitted?

While there may be significant value to expanding upon the pilot within the NCI clinical trials program, the ACR also envisions robust opportunities to pursue application of this concept in a real world data setting. So, as another pathway forward, the ACR intends to apply this proven methodology to a real-world setting across a network of sites, thereby creating a distributed “virtual repository” of medical imaging studies. We then intend to engage this site network to evaluate AI algorithms which can be transferred to the participating site and tested locally on the identified cohort of imaging studies. AI algorithm performance data can then be aggregated and used to inform the AI vendor and used to educate the user community and the FDA.

Conclusion

The transformation of medical care brought about by the advent of artificial intelligence will only continue to grow in scope, influence and value. Much as clinicians are needing to evolve with the growth of AI methods and usage, so too should our approaches to making medical imaging studies available through large publicly available central repositories and novel, distributed approaches. As one critical step to bridge many of the challenges currently faced by these repositories, we propose leveraging a federated approach for images with a centrally stored metadata index.

References

[1] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. 2013. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository." *Journal of Digital Imaging* (Springer Science and Business Media LLC) 26 (6): 1045-1057. doi:https://doi.org/10.1007/s10278-013-9622-7.

[2] Fedorov, A., Longabaugh, W. J. R., Pot, D., Clunie, D. A., Pieper, S., Aerts, H. J. W. L., Homeyer, A., Lewis, R., Akbarzadeh, A., Bontempi, D., Clifford, W., Herrmann, M. D., Höfener, H., Octaviano, I., Osborne, C., Paquette, S., Petts, J., Punzo, D.,. 2021. "NCI Imaging Data Commons." *Cancer Research* 81: 4188-4193. doi:http://dx.doi.org/10.1158/0008-5472.CAN-21-0950.

[3] American Brain Tumor Association. n.d. *Glioblastoma (GBM).* Accessed May 2024. https://www.abta.org/tumor_types/glioblastoma-gbm/#:~:text=On%20average%2C%20more%20than%2012%2C000,year%20in%20the%20United%20States.

[4] Bialecki, B, Gimpel, J. 2024. Study of imaging data conducted on EA1141 *Abbreviated breast MRI for screening women with dense breasts*.

[5] Regenstrief. n.d. *LOINC/RSNA Radiology Playbook User Guide.* Accessed May 2024. https://loinc.org/kb/users-guide/loinc-rsna-radiology-playbook-user-guide/.