Implementation of Big Imaging Data Pipeline Adhering to FAIR Data Principles for Distributed Machine Learning in Oncology

Implementation of Big Imaging Data Pipeline Adhering to FAIR Principles for Federated Machine Learning in Oncology

Article *in* IEEE Transactions on Radiation and Plasma Medical Sciences · September 2021

DOI: 10.1109/TRPMS.2021.3113860

CITATIONS
0

READS
58

15 authors, including:

Ashish Jha
Tata Memorial Centre
115 PUBLICATIONS **112** CITATIONS
SEE PROFILE

Sneha Mithun
Tata Memorial Centre
49 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Umesh Kumar Baburao Sherkhane
Tata Memorial Centre
5 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Vinay Jaiswar
Tata Memorial Centre
5 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Implementation of Big Imaging Data Pipeline Adhering to FAIR Principles for Federated Machine Learning in Oncology

Ashish Kumar Jha, Sneha Mithun, Umesh B. Sherkhane, Vinay Jaiswar, Zhenwei Shi, Petros Kalendralis, Chaitanya Kulkarni, M. S. Dinesh, R. Rajamenakshi, Gaur Sunder, Nilendu Purandare, Leonard Wee, V. Rangarajan, Johan van Soest, and Andre Dekker

***Abstract*—Cancer is a fatal disease and one of the leading causes of death worldwide. The cure rate in cancer treatment remains low; hence, cancer treatment is gradually shifting toward personalized treatment. Artificial intelligence (AI) and radiomics have been recognized as one of the potential areas of research in personalized medicine in oncology. Several researchers have identified the capabilities of AI and radiomics to characterize phenotype and there by predict the outcome of treatment in oncology. Although AI and radiomics have shown promising initial results in diagnosis and treatment in oncology, these technologies are also facing challenges of standardization and scalability. In the last few years, researchers have been trying to develop a research infrastructure for federated machine learning that increases the usability of Big Data for clinical research. These research infrastructures are based on the findable, accessible, interoperable, and reusable (i.e., FAIR) data principles. The India-Dutch "big imaging data approach for oncology in a Netherlands India collaboration" (BIONIC) is a jointly funded initiative by the Dutch Research Council (NWO) and the Indian Ministry of Electronics and Information Technology (MeitY), aiming to introduce radiomic-based research into clinical environments using federated machine learning on geographically dispersed collections of FAIR data. This article described a prototype end-to-end research infrastructure implemented through the BIONIC partnership into a leading cancer care public hospital in India.**

***Index Terms*—Artificial intelligence (AI), findable, accessible, interoperable, and reusable (FAIR) data, machine learning, natural language processing (NLP), radiomics.**

Manuscript received May 20, 2021; revised September 10, 2021; accepted September 15, 2021. Date of publication September 23, 2021; date of current version February 3, 2022. This work was supported in part by NWO in Netherlands (Sanction no.: 629.002.205) and in part by the Ministry of Electronics and Information Technology in India (Sanction no.: 13(2)-2015-CC-BT). *(Leonard Wee, V. Rangarajan, Johan van Soest, and Andre Dekker are contributed equally to this work.) (Corresponding author: Ashish Kumar Jha.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the hospital Institutional Ethics Committee (Institutional Ethics Committee-I, Tata Memorial Centre [IEC, TMC], Mumbai, India) under Application no. IEC/1017/1905/001, as a retrospective study, with waivers of informed consent from involved patients as per IEC policy of our hospital by the same Ethics Committee.

Please see the Acknowledgment section of this article for the author affiliations.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TRPMS.2021.3113860.

Digital Object Identifier 10.1109/TRPMS.2021.3113860

I. Introduction

CANCER is a fatal disease and is the second leading cause of deaths worldwide. According to GLOBOCAN, cancer accounted for about ten million deaths in 2020 [1]. Cancer treatment is complex and requires multidisciplinary collaboration [2]. Gradually, the paradigm in cancer treatment is shifting toward more personalized medicine [3]. While the personalized approach is gaining ground in oncology, artificial intelligence (AI) is also being applied to support personalized treatment [4].

There is a vast spectrum of treatment-related data, i.e., imaging, pathology, biochemical measurements and blood tests, genetics, and clinical observation, being generated every day during cancer treatment and recorded in electronic format. The data are stored in various parts of a hospital information system (HIS) [5]. Medical imaging, such as computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and single-photon emission tomography (SPECT), play an integral role in cancer diagnosis and treatment [6]. These constitute the majority proportion by volume in a typical HIS. Medical images are stored in specialized high-throughput storage devices known as picture archival and communication systems (PACS) [7].

Imaging features, along with a wide spectrum of histological and clinical features, have been explored as prognostic indicators and to predict the future outcomes of cancer treatment. "Radiomics" is a major research development in recent years that attempts to define a digital "fingerprint" of cancer [8]. Radiomics involves high-throughput mathematical extraction of structured quantitative data (i.e., features) from qualitative clinical imaging [9]. A radiomics signature, meaning a set of single features carrying either prognostic or predictive value, can augment conventional visual interpretation by human experts, and reveal deeper insights into the structure, behavior, and therapeutic response of cancer [10]–[13]. However, the vast volume of features extracted by radiomics can also be unwieldy to manage.

The implementation of AI research in oncology touches upon many problems; these include: 1) data retrieval from an HIS and syntactic standardization of data; 2) standardization of radiomics extraction; 3) incorporation

of radiomics models into clinical practice; 4) maintaining persistent and interoperable descriptions of the data; and 5) archival of data in such a way that is findable and accessible to future researchers.

Some of the aforementioned issues can be ameliorated by following FAIR data principles. Kalendralis *et al.* [14] have shown how semantic ontologies can be used to apply syntactic standards to, and provide descriptive metadata about feature extraction methods on top of, existing medical data but without editing the original source data itself. Semantic Web standards based on the resource descriptor framework (RDF) can be used to store data with persistent unique identifiers, in such a way that is completely agnostic to the underlying database schema in the original data source [15]. This, therefore, allows powerful query, filter, and joining commands in the SPARQL language that can link disparate data sources together.

Above and beyond the FAIR principles, scientific collaboration can be encouraged at the health systems and institutions level by homogenizing data collection in clinical routine procedures and allowing multi-institutional sharing of data [14].

However, actual sharing of patient-level data requires additional contractual procedures associated with data ownership, control, access, permitted usage, and protection of patient confidentiality. Such legal needs vary immensely between jurisdictions, for example, The Netherlands and India [17]. The need to share individual patient data may be side-stepped via a privacy-preserving distributed learning approach using federated (decentralized) datasets for statistical modeling [18]. Researchers have demonstrated the feasibility of federated learning over a variety of open and nonopen-source infrastructures, showing that models can be trained on large datasets, are equivalent to results on centralized data and can support radiomics model training and validation [18]–[23]. The RDF-based FAIR data representation forms the basis of distributed learning systems that were able to operate with multisite geographically dispersed data sources [18], [24]–[26].

In this work, we explain how a big imaging data processing pipeline has been implemented in Tata Memorial Hospital (TMH) in Mumbai, India, with the support of big imaging data approach for oncology in a Netherlands India collaboration (BIONIC) partners. We discuss how FAIR been used as a guiding principle, such that data are available for internal research as well as for privacy-preserving federated machine learning. While the solutions presented needed to be intrinsically adapted to TMH, the intention is to demonstrate how similar technologies could be implemented in other hospitals that make private data available for federated learning studies.

II. Materials and Methods

Overall Organization of the Implementation: The overall implementation scheme of the pipeline is shown in Fig. 1. This pipeline serves as the template for preparing data in a FAIR manner, using RDF and ontologies in the data aggregation/integration layer. Prior to this, we show how clinical, imaging, and text report data need to be individually extracted from different sections of the HIS using data stream-specific workflows, and how standardization and metadata need to be applied within each of these stream-specific workflows.

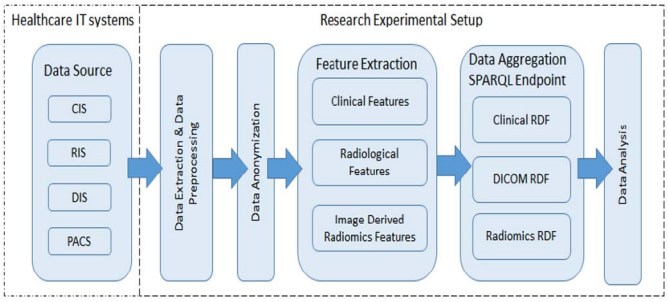

Fig. 1. Schematic providing overview of the BIONIC big imaging data processing pipeline.

TABLE I
Health IT Systems and Attributes

Information Systems	Nature of Data
Clinical Information Systems (CIS)	Demographic data, treatment data and follow-up data including clinical baseline factors such as TNM stage, pleural effusion etc.
Diagnostic Information Systems (DIS)	Pathology report, Immunohistochemistry (IHC) Report and Blood Report that are based on the blood and tissue samples
Radiological information systems (RIS)	Radiology Report and Nuclear Medicine report that includes Imaging related diagnostic findings like malignancy, non-malignancy, disease progression

Datasets: Health IT infrastructure comprises multiple software applications that are required to manage day-to-day clinical activities. The clinical data, nonimage diagnostic data, radiologist reports, and radiological images are stored in the HIS as subsystems; a clinical information system (CIS), diagnostic information system (DIS), radiological information system (RIS), and PACS, respectively. Table I shows the types of data within each of the aforementioned subsystems. Imaging data and its associated metadata are stored in a PACS using the industry-standard digital imaging and communications in the medicine (DICOM) format.

Patient Consent: For retrospective nonexperimental studies, the institutional policy was to apply for a consent waiver through the Institutional Ethics Committee (IEC). The IEC reviews the research study and gives approval (including any necessary usage conditions) for private data of hospital subjects to be harvested from the HIS.

Data Extraction: In-house scripts were developed at TMH in Python programming language to customize and control the data extraction from the HIS subsystems. The data extraction is covered in two major modules: 1) nonimaging data for the CIS, RIS, and DIS streams (see Fig. 2) and 2) imaging data from the PACS (see Fig. 3).

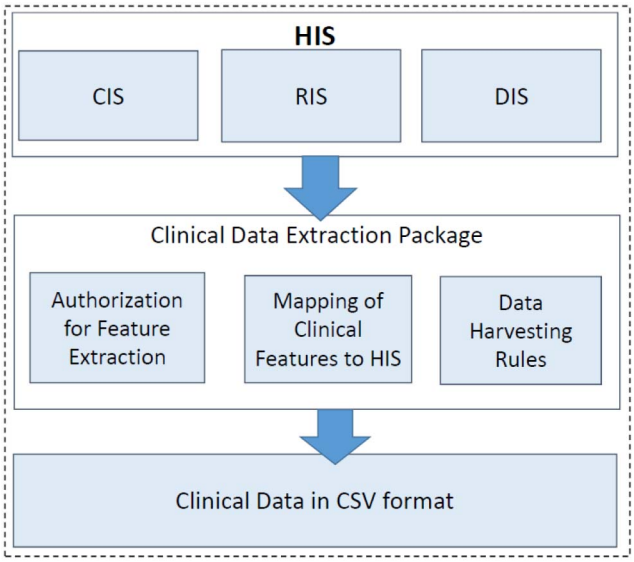

Fig. 2. Overview of the nonimage data extraction module.

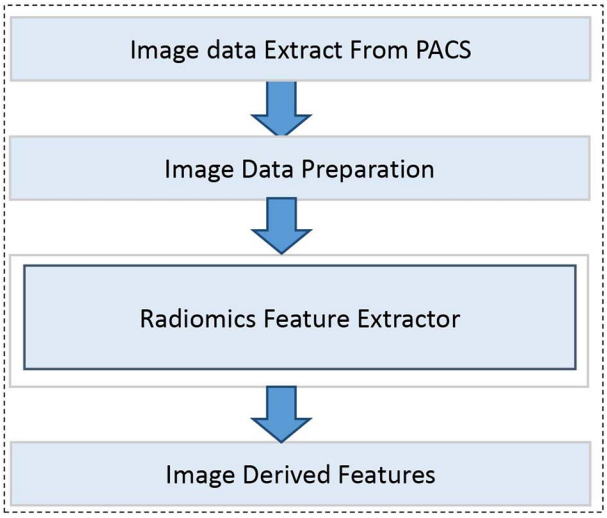

Fig. 3. Overview of the image data extraction module.

A. Nonimage Data

The nonimage data extraction workflow shown in Fig. 2 from CIS, DIS, and RIS requires a detailed understanding of the internal schema in each system. Our procedure also includes checking authorization to access the data elements (in accordance with the IEC approval). The customizable parts refer to the mapping of the data elements to specific locations in the HIS and extraction routines to retrieve values from the HIS.

The data harvested from the HIS will rarely be complete in all aspects. There is generally a high prevalence of "missing values." The customization also covers how missing values are handled. For instance, the missing values can be initially filled in as "NA" text strings. Next, specific filtering rules set by the clinical user for a given study can be applied.

1) If 20% or more of values in the data field is missing, then the data field may be omitted.
2) Or if 20% of the data fields for a given subject is missing, the subject may be omitted.
3) Or if a value is missing in a strictly mandatory data field, such as "gender" or "survival status," then the subject may be omitted.

These filtering steps are performed to ensure a reasonably high degree of data completeness coming from the harvesting process, without entirely relying on complete case analysis. Additional filters, imputation, and exclusion rules can be manually added into the workflow if required. Finally, protected health information (PHI) is obscured by using a lookup key file to replace identifiable information (supplementary material). The result of extraction of nonimage data is a deidentified plain text file in the comma-separated value (CSV) format.

B. Imaging Data

The image data extraction workflow is shown in Fig. 3. Individual subjects imaging studies comprising of CT, MRI, and/or PET need to be manually "pushed" from the PACS or retrieved using one of the integrated image management workstations provided by a vendor, e.g., Philips Intellispace Discovery (research-only build; Philips Medical System, Eindhoven, The Netherlands) or Advantage 4.6 (General Electric, Waukesha, WI, USA. The native format for medical images was retained as DICOM.

For BIONIC data preparation, we were specifically interested in the image data and the region-of-interest annotation file "RTSTRUCT." The RTSTRUCT files were generated in the vendor workstations using radiological annotation tools. The Philips Intellispace Discovery platform permits the option to connect with other tools, such as a radiomics extraction tool or a deep-learning automated segmentation algorithm. As before, PHI is obscured using the same lookup key file as for the nonimaging data.

We implemented a radiomics feature extractor in Python language based on ORAW [27] but we used the *Plastimatch*v1.9.0 [28] library to convert a gross tumor volume (GTV) region of interest from each RTSTRUCT file into a binary mask. The image and corresponding binary mask were passed to *Pyradiomics* v2.1.2 [29], [30] to compute 1093 features. Features consisted of 13 shape, 17 intensity-histogram, and 73 textural features. Intensity-histogram and textural features were recomputed after applying the Laplacian of Gaussian (LoG) filters with three widths (total 270 features) and wavelet decomposition filters at eight levels (total 720 features). Details about the features and filters are available on the online Pyradiomics documentation. Parameters of radiomics extraction were controlled via configuration (.yml) files for each of Plastimatch and Pyradiomics. The radiomics features were saved in the CSV format (Fig. 4).

Data Aggregation: The above steps have defined how clinical features and radiomics features have been separately extracted deidentified and then preprocessed as individual CSV objects. The intermediate step of saving these CSV objects allows additional quality assurance and inspection of the data for any errors or inconsistencies that had

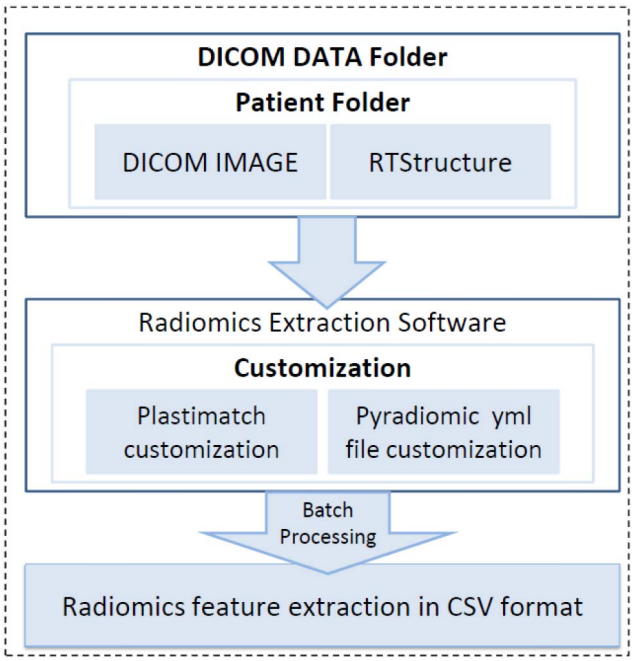

Fig. 4. Radiomics extraction pipeline.

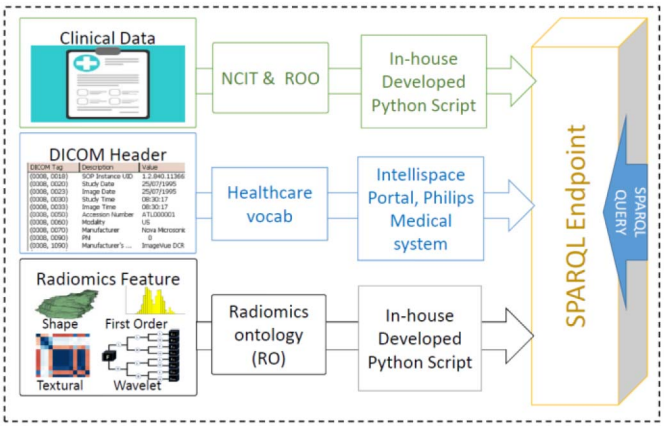

Fig. 5. Clinical, DICOM metadata, and radiomics data are converted to the RDF format and integrated as linked data in a local SPARQL repository.

escaped interception, particularly for the clinical factors. At the moment, this is being done manually by domain experts and researchers but we leave the option open for future automation (see Discussion later). The CSV format made it amenable to be copied into SPSS or other data management tools for additional cleaning and filling of missing values (where possible).

Data integration, standardization, and interoperability are performed using an ontology-guided semantic mapping procedure. Fig. 5 indicates the ontologies that are applied to the individual data streams; the radiation oncology ontology (ROO) [31], [32] and the National Cancer Institute Thesaurus (NCIT) [33] for clinical and treatment-related features, an open-source DICOM ontology [34] for DICOM imaging metadata, and the radiomics ontology (RO) [35], [36] for the radiomics features that follow the Image Biomarker Standardization Initiative recommendations [37], [38].

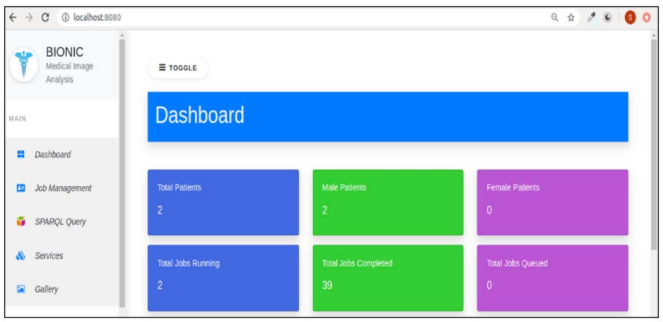

Fig. 6. Home screen of a prototype Web-based dashboard for the BIONIC project.

Each of the CSV objects was converted to the RDF format using an in-house Python script using the rdflib 5.0.0 [39] library with different target ontologies as defined above. The DICOM headers were extracted directly into RDF (without passing through an intermediate CSV) via a plug-in provided on Philips Intellispace Discovery.

The final destination of the generated RDF data was an Apache Jena Fuseki server [40] installed as a SPARQL endpoint inside the hospital IT firewall. The SPARQL query language is then used to access the RDF triples that are archived in the SPARQL endpoint [41]. The RDF triples are maintained in a persistent online graph database through a TDB triple store client application [42], which also supplies a user interface through which remote query of RDF data is possible using a SPARQL v1.1 compliant engine called ARQ [43]. SPARQL queries are entered in ARQ to retrieve the data from the RDF store [43].

III. Results

Given a large number of Python scripts, management of the data pipelines can potentially become cumbersome. In collaboration with BIONIC partner C-DAC, a prototype workflow management system has been developed that integrated some of the scripting work behind a graphical Web interface (dashboard). This was done to try to reduce the level of technical complexity for a typical clinically minded researcher.

Fig. 6 provides a cursory overview of the functionality implemented to date in the prototype BIONIC dashboard at TMH. This includes views for managing data, for task logging, and for script execution status reports. The dashboard provides interfaces to the user to inspect clinical and image-derived features directly from CSV files or DICOM folders. Backend subroutines can call on internal scripts where needed to extract radiomics features and export the data directly into RDF for immediate querying using a built-in SPARQL interface. We hasten to add that such work is still a developmental prototype at the present time; however, more functionality and scripting integration are planned in future.

Fig. 7 shows the user-friendly SPARQL query Web interface that allows filtering and joining operations on RDF data residing in the local RDF data endpoint. As an example, that illustrates the power of SPARQL queries to retrieve and join data from disparate sources, we provide an example in Textbox S1 in the supplementary material. In this demonstration query,

Fig. 7. Close up view of the SPARQL querying interface, where a researcher may enter filtering and linking queries to locate data in the RDF database.

we retrieved patients' age and biological sex from the clinical RDF, then linked it with two radiomics features in the Radiomics RDF. The join was achieved via the patients' ID and patients' CT scan identifier stored in the DICOM metadata RDF. The result of such a query is shown in Fig. S1 in the supplementary material.

IV. Discussion

The purpose of this article is to describe a medical imaging data processing pipeline linking clinical features, imaging metadata, and extracted radiomics features. The goal of the BIONIC collaboration is to help get data ready for federated machine learning and multicenter collaboration at a large scale, but without forcing anyone to share identifiable data. The lynchpin of this work is thus to integrate disparate sources of data in such a way that is agnostic to the internal schema of the databases, language, local coding, etc. Semantic ontologies and mapping scripts were central to linking hospital data together, guided by the FAIR management principles.

Though the procedures presented here need to be specifically adapted to TMH, our intention was to demonstrate how similar procedures might be implemented in other hospitals that can make vast amounts of private data more efficiently interoperable and reusable in research.

Patients' demographics or clinical details were reasonably directly findable on the local HIS. We designed the data extraction module to be customizable to harvest project-dependent data from separate subsystems within the HIS; assuming appropriate regulatory permissions have been given. This level of automation could search thousands of patients' data fields within minutes. At the present time, quality assurance and searching for missing values still unavoidably needs expert human intervention, however, the utilization of such automated data collection procedures had already significantly reduced time and effort in finding relevant data.

Several licensed and open-source software have been made widely available for radiomic extraction [9], [44]–[47], but close integration with clinical data pipelines has not been previously demonstrated in detail. This work demonstrates that is indeed feasible, with simple connector scripts, to integrate such radiomics software as "plug-and-play" modules into an imaging research workflow that can consume clinical real-world images close to its source.

While it is widely accepted that AI models require vast amounts of data for training, it has not been so well established as to how model training can be integrated with on-the-ground clinical record-keeping systems such as the HIS and its multiple subsystems [48]. Interoperability of the data and richly descriptive metadata will become ever more critical, since AI research still faces significant challenges in independent validation; AI models developed in one setting with one set of data seem rarely able to interoperate smoothly with data from an independent setting, and hence AI model performance is generally degraded during independent external validation [49]–[51]. The recommendation to pool data together or share data to centralized repositories does not adequately acknowledge the administrative and legal barriers that stand in the way with regards to sharing individual-level data. This is the foundation of our interest in rendering data in such a way as to be more interoperable in federated machine learning projects.

However, there remain some major challenges that we have not yet addressed in the present work. First, there exists extreme heterogeneity in HIS schema and design, such that the low-level extraction scripts that interface with the HIS and its components are expected to be very difficult to clone from one hospital to another; in this scenario, one size certainly does not fit all. A great effort has been expended in TMH in order to customize the data harvesting procedures to the systems and schema that are available locally. It is likely such adaptation efforts need to be repeated in every hospital, unless vendors of HIS and related electronic systems step in with some universally and strictly applied standards. Clinical medical imaging already has the global DICOM standard, which aids significantly with information retrieval and sharing. While HIS vendors are certainly investing in query and data retrieval tools that are compliant with FHIR and HL7, progress and adoption remains slow.

Second, even assuming that data can be rapidly harvested from the HIS, it remains a major effort to impose syntactic and semantic interoperability on the data, so that it can be universally understood by any person and, more importantly, by any machine algorithm. In this work, we have used the power of semantic ontologies to harmonize terminology and attach metadata to data, such that these "unique identifiers" can be used to construct filtering and linking queries over heterogeneous data sources. This design task remains highly time consuming and requires high levels of expertise, as well as the knowledge of the clinical domain. With the rapid development of the natural language processing (NLP) domain within AI and machine learning, it is expected that emerging NLP technologies can eventually be deployed to the task of the semantic labeling of data.

This is related to another limitation of the present work that remains to be addressed in detail; much of clinical and diagnostic data exists as natural free-flowing human language texts (so-called "free text"). Such narrative descriptions of disease—its diagnosis, development, and outcome—is recognized as a rich source of potentially clinically actionable information, but we have presently placed the major part of our initial efforts on structured information encoded into specific data fields in the HIS. We acknowledge therefore that our work is not yet complete, and better utilization of unstructured free-text information must be achieved in the near future by exploiting NLP.

With regard to FAIR principles, we need to consider how close we have managed to come to the FAIR principles, while

also recognizing gaps in implementation that need to be closed in the future [52].

In terms of findability, we have successfully transitioned from highly localized data retrieval from the HIS into an RDF database with persistent unique identifiers based on public domain ontologies, including identifiers for richly descriptive metadata for images and radiomics features. However, due to the privacy-sensitive nature of the data, patient data and metadata of such detailed nature cannot be placed online, even in the deidentified form. In its place, we have not yet developed an institutional protocol to make our data repository findable without betraying confidentiality.

With regards to accessibility, we have discussed our aforementioned strategy favoring privacy-preserving federated learning. Our present vision is that data accessibility matters need to be defined within formal research collaborations, each with focused clinical questions and contractual legal arrangements pertaining to the use of real-world patient data.

In the matters pertaining to interoperability and reusability, we propose that our data integration with RDF already applies a high degree of general understandability to the data, which in turn allows us to reuse patient data garnered from routine encounters and provision of standard care. At this time, specifications of lexica and ontologies used, as well as conditions of using the data, are all governed within a formal multiparty collaboration structure.

Finally, the aforementioned medical imaging data pipeline was intended to make data FAIR for internal clinical research, in addition to being the necessary preparatory steps to making data FAIR for federated international collaboration. To this end, we felt the development of a prototype dashboard for management, process monitoring, and linked data queries, and presenting this through a user-friendly internal browser-based interface, is an important step toward usability and clinical deployment, whereas the Web-based database SPARQL query is also possible for authorized collaborators using Web-based query tools.

Several state-of-the-art ETL and research data warehouse are described in the literature, which are designed for customized extraction, ontological mapping, and the automated generation of SQL statements [53]–[56]. These data warehouses also allow the storage of heterogeneous medical data as ours but unlike uses relational structured database. The ETL tools of these data warehouse extract data from the structured table of HIS, transforming them to a given target structure with the help of ontology and finally load the source system contents into a research data warehouse, whereas our ETL module is able to extract data from EHR without knowing the relational database structure and table directly through Web access and free text data mining provides more scalability to the module. Ontology mapped triples allow SPARQL query independent of local terminology.

V. Conclusion

We implemented a medical imaging data processing pipeline linking clinical factors, imaging metadata, and extracted radiomics features. This was done for a comprehensive cancer hospital in India (TMH) as a demonstration of the BIONIC Indo-Dutch research collaboration. Data integration across HIS subsystems was guided by the FAIR data principles; specifically, we relied on domain semantic ontologies for terminology standardization, knowledge representation, and internal data linkage. We harvested and then stored clinical data, DICOM imaging metadata, and extracted radiomics features as an RDF repository. A browser-based dashboard was also provided to facilitate usage and future deployment as a possible clinical research environment. This local RDF-based system is synergistic with, and readily connected to, a multicenter federated machine learning infrastructure.

Acknowledgment

The authors wish to thank the Dutch Research Council (NWO) and Indian Ministry for Electronics and Information Technology (MeitY) for providing financial support to the BIONIC project.

Ashish Kumar Jha and Sneha Mithun are with the Department of Radiation Oncology (MAASTRO), GROW School for Oncology, Maastricht University Medical Centre+, 6229 ET Maastricht, The Netherlands, also with the Department of Nuclear Medicine and Molecular Imaging, Tata Memorial Hospital, Mumbai 400012, India, and also with the Homi Bhabha National Institute, Deemed University, Mumbai 400094, India (e-mail: ashish.kumar.jha.77@gmail.com).

Umesh B. Sherkhane is with the Department of Radiation Oncology (MAASTRO), GROW School for Oncology, Maastricht University Medical Centre+, 6229 ET Maastricht, The Netherlands, and also with the Department of Nuclear Medicine and Molecular Imaging, Tata Memorial Hospital, Mumbai 400012, India.

Vinay Jaiswar is with the Department of Nuclear Medicine and Molecular Imaging, Tata Memorial Hospital, Mumbai 400012, India.

Zhenwei Shi, Petros Kalendralis, Leonard Wee, Johan van Soest, and Andre Dekker are with the Department of Radiation Oncology (MAASTRO), GROW School for Oncology, Maastricht University Medical Centre+, 6229 ET Maastricht, The Netherlands.

Chaitanya Kulkarni and M. S. Dinesh are with the Philips India Research, Philips Innovation Campus, Bengaluru 560045, India.

R. Rajamenakshi and Gaur Sunder are with the Centre for Development of Advanced Computing, Pune 411008, India.

Nilendu Purandare and V. Rangarajan are with the Department of Nuclear Medicine and Molecular Imaging, Tata Memorial Hospital, Mumbai 400012, India, and also with the Homi Bhabha National Institute, Deemed University, Mumbai 400094, India.

References

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] S. Chakraborty and T. Rahman, "The difficulties in cancer treatment," *Ecancermedicalscience*, vol. 6, p. ed16, Nov. 2012.

[3] A. A. Agyeman and R. Ofori-Asenso, "Perspective: Does personalized medicine hold the future for medicine?" *J. Pharm. Bioallied Sci.*, vol. 7, no. 3, pp. 239–244, Jul.–Sep. 2015.

[4] N. J. Schork, "Artificial intelligence and personalized medicine," *Cancer Treat Res.*, vol. 178, pp. 265–283, Oct. 2019.

[5] M. Khalifa and O. Alswailem, "Hospital information systems (HIS) acceptance and satisfaction: A case study of a tertiary care hospital," *Procedia Comput. Sci.*, vol. 63, pp. 198–204, 2015, doi: 10.1016/j.procs.2015.08.334.

[6] L. Fass, "Imaging and cancer: A review," *Mol. Oncol.*, vol. 2, no. 2, pp. 115–152, Aug. 2008.

[7] T. E. Schultheiss, L. R. Coia, E. E. Martin, H. Y. Lau, and G. E. Hanks, "Clinical applications of picture archival and communications systems in radiation oncology," *Seminars Radiat. Oncol.*, vol. 7, no. 1, pp. 39–48, Jan. 1997.

[8] H. J. W. L. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.*, vol. 5, p. 4006, Jun. 2014.

[9] P. Lambin *et al.*, "Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012.

[10] I. Fornacon-Wood, C. Faivre-Finn, J. P. B. O'Connor, and G. J. Price, "Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype," *Lung Cancer*, vol. 146, pp. 197–208, Aug. 2020.

[11] Z. Liu *et al.*, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges," *Theranostics*, vol. 9, no. 5, pp. 1303–1322, Feb. 2019.

[12] P. Lambin, E. R. Velazquez, and R. Leijenaar, "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[13] P. Lambin *et al.*, "Radiomics: The bridge between medical imaging and personalized medicine," *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, 2017.

[14] P. Kalendralis *et al.*, "FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections," *Med. Phys.*, vol. 47, no. 11, pp. 5931–5940, Jun. 2020.

[15] *Resource Description Framework (RDF)*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.w3.org/RDF/

[16] A. Zuiderwijk, R. Shinde, and W. Jeng, "What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0239283.

[17] K. Tucker *et al.*, "Protecting patient privacy when sharing patient-level data from clinical trials," *BMC Med. Res. Methodol.*, vol. 16, no. S1, p. 77, Jul. 2016.

[18] Z. Shi, I. Zhovannik, A. Traverso, F. J. W. M. Dankers, T. M. Deist, and P. Kalendralis, "Distributed radiomics as a signature validation study using the personal health train infrastructure," *Sci. Data*, vol. 6, pp. 1–8, Oct. 2019, doi: 10.1038/s41597-019-0241-0.

[19] Z. Shi *et al.*, "External validation of radiation-induced dyspnea models on esophageal cancer radiotherapy patients," *Front Oncol.*, vol. 9, p. 1411, Dec. 2019.

[20] T. M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, and S. Walsh, "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: EuroCAT," *Clin. Transl. Radiat. Oncol.*, vol. 4, pp. 24–31, Jun. 2017.

[21] M. Wolfson *et al.*, "DataSHIELD: Resolving a conflict in contemporary bioscience—Performing a pooled analysis of individual-level data without sharing the data," *Int. J. Epidemiol.*, vol. 39, no. 5, pp. 1372–1382, 2010.

[22] C.-L. Lu *et al.*, "WebDISCO: A Web service for distributed cox model learning without patient-level data sharing," *J. Amer. Med. Informat. Assoc.*, vol. 22, no. 6, pp. 1212–1219, 2015.

[23] M. Bogowicz *et al.*, "Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer," *Sci. Rep.*, vol. 10, p. 4542, Mar. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-61297-4

[24] A. Jochems *et al.*, "Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital—A real life proof of concept," *Radiotherapy Oncol.*, vol. 121, no. 3, pp. 459–467, 2016, doi: 10.1016/j.radonc.2016.10.002.

[25] T. M. Deist *et al.*, "Distributed learning on 20000+ lung cancer patients—The personal health train," *Radiotherapy Oncol.*, vol. 144, pp. 189–200, Mar. 2020, doi: 10.1016/j.radonc.2019.11.019.

[26] F. Dankers, "Prediction modeling and distributed learning for radiotherapy outcomes in lung cancer patients," Ph.D. dissertation, Radboud Inst. Health Sci., Radboud Univ., Nijmegen, The Netherlands, 2019.

[27] Z. Shi, A. Traverso, J. van Soest, A. Dekker, and L. Wee, "Technical note: Ontology-guided radiomics analysis workflow (O-RAW)," *Med. Phys.*, vol. 46, no. 12, pp. 5677–5684, Dec. 2019.

[28] *Plastimatch V1.9.0 Software*. Accessed: Apr. 30, 2021. [Online]. Available: https://sourceforge.net/projects/plastimatch/

[29] *Pyradiomics Package V2.2.0*. Accessed: Apr. 30, 2021. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/

[30] J. J. M. van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, p. e104, 2017.

[31] *Radiation Oncology Ontology—NCBO BioPortal*. Accessed: Apr. 30, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/ROO

[32] A. Traverso, J. van Soest, L. Wee, and A. Dekker, "The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic Web and ontology techniques," *Med. Phys.*, vol. 45, no. 10, pp. e854–e862, Oct. 2018.

[33] *National Cancer Institute Thesaurus(NCIT)*. Accessed: Apr. 30, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/NCIT

[34] *Semantic DICOM Ontology*. Accessed: Apr. 30, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/SEDI

[35] *Healthcare Vocab*. Accessed: Apr. 30, 2021. [Online]. Available: http://purl.org/healthcarevocab/v1

[36] *Radiomics Ontology NCBOBioPortal*. Accessed: Apr. 30, 2021. [Online]. Available: https://bioportal.bioontology.org/ontologies/RO

[37] *The Image Biomarker Standardisation Initiative*. Accessed: Apr. 30, 2021. [Online]. Available: https://ibsi.readthedocs.io/en/latest/index.html

[38] *Radiological Society of North America. Quantitative Imaging Biomarkers Alliance (QIBA). rsna.org*. Accessed: Apr. 30, 2021. [Online]. Available: https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance/

[39] *RDFLIB 5.0.0*. Accessed: Apr. 30, 2021. [Online]. Available: https://rdflib.readthedocs.io/en/stable/

[40] *Apache Jena—Apache Jena Fuseki*. Accessed: Apr. 30, 2021. [Online]. Available: https://jena.apache.org/documentation/fuseki2/

[41] *Apache Jena—SOH—SPARQL Over HTTP*. Accessed: Apr. 30, 2021. [Online]. Available: https://jena.apache.org/documentation/fuseki2/soh.html

[42] *TDB*. Accessed: Apr. 30, 2021. [Online]. Available: https://jena.apache.org/documentation/tdb/

[43] *ARQ-Extending Query Execution*. Accessed: Apr. 30, 2021. [Online]. Available: https://jena.apache.org/documentation/query/arq-query-eval.html

[44] E. Pfaehler, A. Zwanenburg, J. R. de Jong, and R. Boellaard, "RaCaT: An open source and easy to use radiomics calculator tool," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0212223.

[45] L. E. Court, X. Fave, D. Mackin, J. Lee, J. Yang, and L. Zhang, "Computational resources for radiomics," *Transl. Cancer Res.*, vol. 5, no. 4, pp. 340–348, 2016.

[46] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review current status and future potential," *Comput. Med. Imag. Graph.*, vol. 31, pp. 198–211, Jun. 2007.

[47] W. Rogers, B. Ryack, and G. Moeller, "Computer-aided medical diagnosis: Literature review," *Int. J. Biomed. Comput.*, vol. 10, no. 4, pp. 267–289, 1979.

[48] Z. Dlamini, F. Z. Francies, R. Hull, and R. Marima, "Artificial intelligence (AI) and big data in cancer and precision oncology," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 2300–2311, 2020, doi: 10.1016/j.csbj.2020.08.019.

[49] I. S. Boon, T. P. T. A. Yong, and C. S. Boon, "Assessing the role of artificial intelligence (AI) in clinical oncology: Utility of machine learning in radiotherapy target volume delineation," *Medicines*, vol. 5, no. 4, p. 131, Dec. 2018.

[50] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, Art. no. baaa010, Jan. 2020.

[51] F. Shaikh *et al.*, "Translational radiomics: Defining the strategy pipeline and considerations for application—Part 2: From clinical implementation to enterprise," *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 543–549, Mar. 2018.

[52] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, Mar. 2016, Art. no. 160018, doi: 10.1038/sdata.2016.18.

[53] R. Bache, S. Miles, and A. Taweel, "An adaptable architecture for patient cohort identification from diverse data sources," *J. Amer. Med. Inf. Assoc.*, vol. 20, no. e2, pp. e327–e333, 2013, doi: 10.1136/amiajnl-2013-001858.

[54] A. J. McMurry *et al.*, "SHRINE: Enabling nationally scalable multi-site disease studies," *PLoS ONE*, vol. 8, no. 3, 2013, Art. no. e55811.

[55] S. N. Murphy *et al.*, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," *J. Amer. Med. Inf. Assoc.*, vol. 17, no. 2, pp. 124–130, 2010.

[56] S. Mate, F. Köpcke, and D. Toddenroth, "Ontology-based data integration between clinical and research systems," *PLoS ONE*, vol. 10, no. 3, 2015, Art. no. e0122172, doi: 10.1371/journal.pone.0116656.