2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Towards Data Integration for AI in Cancer Research *

Alexandra Kosvyra, Dimitrios Filos, Dimitrios Fotopoulos, Tsave Olga and Ioanna Chouvarda, *Member, IEEE*

Abstract— **Cancer research is increasing relying on data-driven methods and Artificial Intelligence (AI), to increase accuracy and efficiency in decision making. Such methods can solve a variety of clinically relevant problems in cancer diagnosis and treatment, provided that an adequate data availability is ensured. The generation of multicentric data repositories poses a series of integration and harmonization challenges. This work discusses the strategy, solutions and further issues identified along this procedure within the EU project INCISIVE that aims to generate an interoperable pan-European federated repository of medical images and an AI-based toolbox for medical imaging in cancer diagnosis and treatment.**

Clinical Relevance— **Supporting the integration of medical imaging data and related clinical data into large interoperable repositories will enable the development, and validation, and wider adoption of AI-based methods in cancer diagnosis, prediction, treatment and follow-up.**

I. Introduction

Cancer remains one of the main causes of morbidity and mortality worldwide with a rising prevalence in the developed countries [1]. Medical decisions for cancer-related patient care both diagnosis and treatment, heavily rely on cancer imaging data. In addition, clinical, histopathological and other types of data complement screening of cancer patients. This wealth and multitude of data for decision making creates human processing bottlenecks, while this is also an opportunity where computerized problem solving can contribute to efficiency and accuracy.

AI, and Machine Learning (ML) as a main branch of AI, are emerging as key constituents in healthcare and medicine. Recent advancements in AI and ML have paved the way for the analysis of big datasets in a cost- and time-effective manner [2]. Cancer offers a unique context for medical decisions given not only its multidimensional and heterogeneous forms with evolution of disease but also the need to consider the individual condition of patients, their ability to receive treatment, and their responses to treatment. Deep learning has been widely used with massive imaging data fir diagnosis [3]. ML on the other hand can be used to developed predictive models that combine information from different sources, including clinical data and radiomics in order to improve decision making and thus lead to better patient management. To this end, AI and ML can be effectively applied to imaging data to recognize unique and complex features to facilitate automated assessments. [4].

Lately, the increasing consortia-based research has led to a better understanding of cancer etiology, diagnosis, treatment, prognosis etc. shedding light into lifestyle, clinical, and genetic determinants of such pathologies and their outcomes. The employment of existing data sets (e.g. imaging) in combination with newly-produced/obtained types of data significantly enhance the management of salient health burdens in a more cost- and time-efficient way. However, in order to take the full advantage of the AI/ML technology in cancer research, interconnectivity, interoperability and harmonization challenges need to be addressed.

The 42-month INCISIVE project (https://incisive-project.eu/) aims to address these challenges. In particular, INCISIVE aims to address the data availability challenge, towards the wide adoption of AI solutions in health imaging. INCISIVE aims to aggregate and unify the fragmented cancer imaging datasets across European healthcare systems and institutions, characterized by a multiplicity of data sources, to enable the integration and full exploitation of current initiatives and isolated databases and to reach a critical mass of gathered data. Together with the generation of an AI-toolbox, the end goal of the INCISIVE is the implementation of a pan-European repository of health images following a federated approach. In this respect the data harmonization constitutes a major pillar.

Data Integration is the procedure of combining multi-source data in a single view. Data integration refers to the semantic integration of big data in order to be easily accessible, usable and updated. Prior to integration, a harmonization procedure must be applied, that results in a common data schema.

Data harmonization, the process of evaluation and management of compatibility of data acquired from heterogeneous sources, arises as an important task in an effort to enhance both retrospective and prospective integrative analyses. Data harmonization is often a difficult process, as it is often heavily based on experience and tacit knowledge and consensus by multiple data providers [5]. As phrased in [6], *"reproducibility of any one group's processes is questionable"*, posing challenges related to available data and quality control processes, which will allow the unified analysis of the acquired data.

A major goal for the INCISIVE harmonization procedure is to follow the FAIR principles [7]. This implies the provision of a well-described, searchable, uniquely identifiable and standardized imaging repository, accompanied by image metadata, which is supported by the data model proposed in this work. This data model will constitute the basis for facilitating the integration of multiple data sources.

* This work has received funding from the EU's H2020 RIA programme under grant agreement No 952179.

A. Kosvyra, D Filos, D Fotopoulos, O Tsave and I Chouvarda are with the School of Medicine, Aristotle University of Thessaloniki, GR (phone: 30-2310999247; fax: 30-2310999263; e-mail: {aekosvyra, dimfilos, difoto, tsaveolga, ioannach}@auth.gr).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

This paper aims to present the methodology followed on integrating multi-source data into a common model and is a multiple steps procedure. This process ensures that different types of data from multiple sites can be linked, shared and reused in INCISIVE towards the harmonization of data collected from 9 data providers participating in the project and existing data from open databases.

II. Data Integration – Harmonisation Methodology

The activities performed to achieve the data harmonization follow the best practices [8]. For imaging data in DICOM, this procedure mainly focused on the headers part.

A. Types of INCISIVE Data

The data collected during the lifetime of the project come from 5 different countries and 9 Data providers, while 4 cancer types are considered namely, lung, breast, colorectal and prostate. These data are divided in two categories, (a) clinical and biological data, and (b) imaging data.

The first category, provided in structured text form, includes demographic and medical history data, histological and blood markers, treatment and tumor details, as well as the imaging acquisition protocol. The second category includes body scans in different modalities (MRI, CT, PET, US, MMG) DICOM format and histopathological images in *png* or *tiff* format. Imaging data in DICOM include besides the image, also the metadata related, among other, to the screening protocol and patient information.

These data will be collected in distinct timepoints during the patients' treatment: (1) Diagnosis, (2) after first treatment (surgery or therapy), (3) 1st Follow-Up (between 4 to 7 months after diagnosis), (4) 2nd Follow-Up (between 9 to 12 months after diagnosis). An overview of this approach is presented in Figure 1.

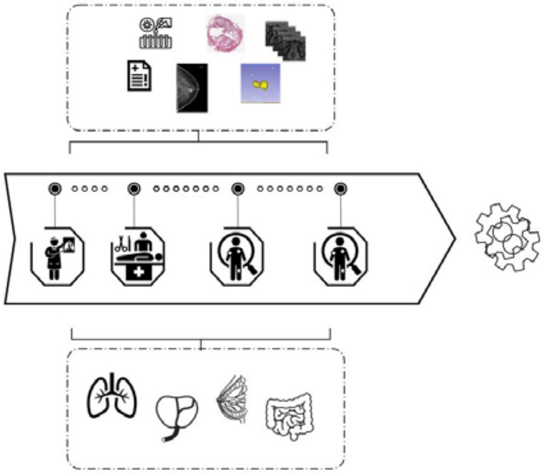

Figure 1 An overview of the INCISIVE data integration process

In addition to the data collected inside the project, an extensive review of the available open databases was performed in order to investigate the type of data included, the de-identification and integration approaches followed, and, finally, integrate these datasets with the project's data. The Cancer Imaging Archive (TCIA) [9] is an open-source archive that hosts oncology image data, annotation details as well as clinical information details, and genomic, proteomics, etc. These data collections are de-identified and organized according to image modality, target organ of the disease (e.g. lung cancer) or research focus. Following this categorization, we searched the archive and organized our findings.

B. Data integration strategy

As mentioned, the data are collected from different sites, spread in five countries and multiple sites, and thus various diversity issues arise. The clinical pathways followed, and the standards used in every site are different. To that end, the integration procedure applied focuses on 3 levels:

1. *Structural / Functional*. A structure that will be able to embed all the information derived from each site has to be defined. This structure will represent the data model that will be used as the basis for the data collection. The challenge in this level was to identify the timepoints that the data will be collected and the subcategories deriving from the information collected. For each one of the subcategories, the fields of data included had to be homogenized to meet the different perspectives of the standards followed in each case.
2. *Privacy*. All privacy issues that may arise during the collection of data have to be identified from the beginning. In clinical data category, such issues arise when it comes to dates of events and patient identification. In imaging data, the DICOM header may contain dates and names, while in some cases such data are burned-in the images.
3. *Semantic*. Another challenge identified was the harmonization of clinical data provided, in terms of terminology used for medication, tumor characterization, grading and staging. Since multiple standards can be used for these categories, a standard that will unify the differentiated data had to be defined and followed by all sites.

C. Definition of an iterative procedure for non-imaging data integration

In order to define the protocol and overcome the homogenization challenges in functional, semantics and privacy levels, an iterative procedure took place following the steps:

i. *Identification*: Proposition of a template per cancer type based on bibliography and medical experience,

ii. *Review*: The templates were circulated through the Data providers, reviewed and discussed,

iii. *Merge:* A consensus of each template was extracted and discussed in a meeting to resolve homogenization issues,

iv. *Redefine:* The data providers were asked to provide a sample case. The cases were reviewed for integrity and privacy issues

v. *Standardize:* Standardization of the fields content and adopted terminologies based on medical standards as ICD-11, ATC,

vi. *Review and Refine:* The templates were circulated again for verification.

The resulting templates constitute the data collection protocol for the non-imaging data.

D. Imaging data integration procedure

The imaging data from all the DRs were analyzed in order to investigate the harmonization issues that may arise. The first step for the harmonization consists the analysis of the anonymization methodology applied. In this respect, the metadata of all DICOM files were processed and a list of all the attributes that are defined in][10] regarding the confidentiality profile, for each DP, was created. A similar approach was followed for the identification for the open DBs. Towards this goal all the databased, from the Cancer Imaging Archive, addressing one of the cancer types of interest were identified and a list of the DICOM metadata fields related to anonymization was created. All those lists were compared to identify the differences on the anonymization procedure applied according to the imaging modality and the data provider. The methodology followed in each cases was compared with the DICOM anonymization standards.

As a next step, additional attributes related to the image, such as Field of view, slice thickness etc. will also analyzed for harmonization purposes.

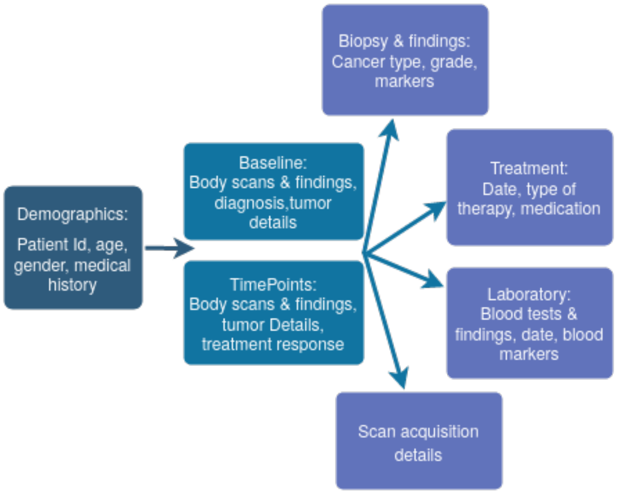
Biopsy & findings: Cancer type, grade, markers
Baseline: Body scans & findings, diagnosis,tumor details
Treatment: Date, type of therapy, medication
Demographics: Patient Id, age, gender, medical history
TimePoints: Body scans & findings, tumor Details, treatment response
Laboratory: Blood tests & findings, date, blood markers
Scan acquisition details

Figure 2. Cancer Imaging Data Model .

III. Results

A. The templates and data model

The templates resulting from the iterative procedure constitute the data model that will be followed for the data collection. Figure 2 depicts the overall structure of this model. The data are classified in 3 levels: (i) Demographics, which contains personal information of the patient and medical history, (ii) Timepoints, which contains information about the tumor characteristics resulting from scan examinations, the progression and the status of the patient in various timepoints, Actual scan examinations are included in the timepoints and linked with the findings, (iii) information about histopathology findings, treatment and blood tests connected to each timepoint.

Of note, the metadata in each timepoint are split into mandatory and optional ones. As mandatory we consider the patient, tumor/cancer and treatment profiles as well as an imaging examination and as optional the blood and histological examinations, medical history, metastases and other signs. This mandatory information was selected as they provide the patient profile, demographics and diagnosis, which are the necessary information needed for the data analysis together with imaging scans. For example, a minimum set of information for a breast cancer patient could be: "Woman, 50 years old, with Invasive Ductal Carcinoma grade 3, BIRADS classification: 6, with a tumor of 13 cm maximum diameter, in UIQ of left breast, identified by Mammography.

B. The privacy issues

Regarding the timepoints and events, all dates were transformed into months from diagnosis. With regards to the patient identification, the data were pseudo-anonymized. A random unique id was assigned to each patient in order to connect the different types of data and make them findable.

As regards the anonymisation of DICOM images, DICOM part 15 [10] lists the metadata attribute fields that must be modified while different confidentially profiles may be applied on each of them. This approach does not allow the identification of the patient while all the information needed for the analysis is retained or protected.

Apart from the DICOM images, other types will also be available, such as histopathology images. Those images are obtained using common photographic devices adapted on a microscope and thus no personal information is included. The name of the file is also modified based on a hashing algorithm to avoid any personal information to appear.

C. The standardization and quality issues

Based on the analysis of the available imaging data, it was observed that diverse anonymization methods applied by each data provider. The application of the same de-identification method will ensure the successful integration of data. In this respect the DICOM standards is adopted which described in detail the action be taken for each attribute which could be used for the identification of the patient.

In terms of terminology unification, specific standards were proposed to achieve the data homogeneity: (1) ATC codes for medications, (2) ICD-11 for cancer classification and tumor characterization, (3) consensus decisions based on prior experience, as for example the type of tumor classification (Stage, TNM, T-plus). Using these standards for the terminology in data collection, the model proposed ensures interoperability.

Given the diversity of the information that may appear in each of the data sources, an integration quality check tool is needed. Quality control tests are foreseen to ensure data quality, overcome dissimilar formats, availability with respect to adequate and uniform sampling, and to assess whether the harmonized data fall within expected value ranges. This tool is envisioned as a rule-based engine, following the constraints defined within the INCISIVE domain, in order to identify whether data follow the data harmonization requirements defined (e.g. follow the same anonymization protocol, include all the mandatory fields, and encodings, etc.) as well as the integrity and consistency of them. This approach is regarded as complementary to the data export and anonymization approach followed by each data provider.

D. Are open data integratable and usable?

After a review of the available data in TCIA, 27 collections were identified, specifically: 12 for breast cancer, 23 for lung cancer, 11 for prostate cancer and 5 for colon cancer. These collections include DICOM supported and/or histopathology image data and other types of data associated with the images of the disease. We categorized these findings based on the image modality used in each dataset, what other non-image data are available - clinical or annotations (which includes every type of data that characterizes the cancer images), cancer type and the clinical timepoints of the image data. An overview of the result is presented in Table 1. The purpose of this review is to clarify if it is possible to integrate the data of the cancer archive in the final repository that will host the data that will be collected during the span of the project. This task needs further investigation, but initial results showed there are datasets that can be integrated and utilized for the training of AI models, based on the data fields they share with the project's data templates. For example, a few datasets have multiple clinical timepoints for the medical images, which is valuable information for modeling and predicting the trajectory of the disease. Another relevant example of a usable dataset is one which contains information about the patient's treatment response.

TABLE I. OVERVIEW OF THE AVAILABLE DATA IN OPEN DATASETS

	Breast	**Lung**	**Prostate**	**Colon**
Databases	12	23	11	5
MR	7		10	
CT	3	17	2	2
PT	2	7	2	
US			1	
MG	5			
other		2	1	1
Histopathology	4	5	3	2
Clinical	12	8	2	3
Omics	2	7	1	2
Annotations	7	9	6	1
Timepoints	8	7	6	1

IV. CONCLUSION

Quantitative analysis in medical imaging, via the calculation of radiomics features, is an emerging field of research and particularly in cancer. AI in cancer research promises to improve the automated quantification of radiographic images [Simon]. Predictive models have been implemented for the discrimination of benign and malignant lesions, prognosis, response to treatment and grading. In order to achieve robust models and reduce bias, the need to be trained with large qualities of diverse data, and thus large multicentric data repositories are needed, to comprise the basis of AI models.

The integration of data coming from different clinical sites can facilitate cancer research, in terms of data volumes and data bias, and leverage the adoption of AI in clinical routine towards improving workflows' efficiency and effectiveness. This work discusses the ongoing effort for data integration, and challenges in doing so, and the strategy to establish a smooth integration procedure for a European cancer data repository. Current efforts focus on the imaging data privacy quality issues and the non-imaging data structural, privacy and semantic issues. A data model, including mandatory and optional data is proposed, along with an anonymization strategy. These are currently being incorporated in a quality tool that is expected to help in the integration process.

Further harmonization approaches are necessary in order to improve the validity and clinical importance of the multicentric AI/radiomics analysis, including imaging protocols and vendors, and histopathology procedures. Annotation of images requires a level of standardisation to increase their employability in an analytics pipeline. Harmonisation of radiomics features (not data) consitutes an additional approach [11] as in ComBat system. Next steps in the integration procedure will consider incorporation of these approaches.

REFERENCES

[1] M. Ghoncheh, and H. Salehiniya, "Inequality in the Incidence and Mortality of All Cancers in the World," *Iran. J. Public Health,* vol 45 no. 12, pp. 1675-1677, 2016.

[2] T. Davenport, and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthc J,* vol 6 no. 2, pp. 94-98, 2019.

[3] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.J. Bae, and N. Kim, "Deep Learning in Medical Imaging," *Neurospine*, vol 16, no. 4,pp. 657–668, 2019.

[4] W.L. Bi, A. Hosny, M.B. Schabath, M.L. Giger, N.J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I.F. Dunn, R.H. Mak, R.M. Tamimi, C.M. Tempany, C. Swanton, U. Hoffmann, L.H. Schwartz, R.J. Gillies, R.Y. Huang, and H.J.W.L. Aerts, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA Cancer J Clin.,* vol 69 no. 2, pp. 127-157, March 2019.

[5] S. Nierkens, A.C. Lankester, R.M. Egeler, P. Bader, F. Locatelli, M.A. Pulsipher, C.M. Bollard, J.J. Boelens, Westhafen Intercontinental Group. "Challenges in the harmonization of immune monitoring studies and trial design for cell-based therapies in the context of hematopoietic cell transplantation for pediatric cancer patients, " *Cytotherapy*, vol 17 no. 12, pp. 1667-1664, December 2015.

[6] B. Rolland, S. Reid, D. Stelling, G. Warnick, M. Thornquist,Z. Feng, and J.D. Potter JD, "Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach," *Am. J. Epidemiol*. Vol 182 no. 12, pp.1033-1038, 2015.

[7] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data*,vol 3 no. 1, pp. 1-9, 2016.

[8] B.M. Schmidt, C.J. Colvin, A. Hohlfeld, and N. Leon , "Definitions, components and processes of data harmonisation in healthcare: a scoping review," *BMC Med Inform. Decis.Mak,* vol 20 no. 222, 2020.

[9] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *J. Digit. Imaging*, vol 26, no. 6, pp. 1045-1057, December, 2013.

[10] http://dicom.nema.org/medical/dicom/current/output/html/part15.html

[11] R. Da-ano, I. Masson, F. Lucia,. et al., "Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies,". *Sci Rep*, vol 10, no. 10248, 2020.