2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871018

Data Quality Check in Cancer Imaging Research: Deploying and Evaluating the DIQCT Tool *

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

Abstract— Data harmonization is one of the greatest challenges in cancer imaging studies, especially when it comes to multi-source data provision. Properly integrated data deriving from various sources can ensure data fairness on one side and can lead to a trusted dataset that will enhance AI engine development on the other side. Towards this direction, we are presenting a data integration quality check tool that ensures that all data uploaded to the repository are homogenized and share the same principles. The tool's aim is to report any human-induced errors and propose corrective actions. It focuses on checking the data prior to their upload to the repository in five levels: (i) clinical metadata integrity, (ii) template-imaging consistency, (iii) anonymization protocol applied, (iv) imaging analysis requirements, (v) case completeness. The tool produces reports with the corrective actions that must be followed by the user. This way the tool ensures that the data that will become available to the developers of the AI engine are homogenized, properly structured and contain all the necessary information needed for the analysis. The tool was validated in two rounds, internal and external, and at the user experience level.

Clinical Relevance— Supporting the harmonized preparation and provision of medical imaging data and related clinical data will ensure data fairness and enhance the AI development.

I. Introduction

Cancer is still considered as the leading cause of death worldwide, accounting for ~10 million deaths in 2020 [1]. According to WHO, the most common types of cancer for 2020 appear to be breast, lung, colon, and prostate based on the number of new cases diagnosed [2]. The cancer burden for most cancer types can be significantly reduced when early detection occurs, and appropriate or more personalized treatment of patients takes place. Moreover, early diagnosis increases the possibility for a better response to treatment leading to a greater probability of survival and less expensive treatment. The last decade(s), Artificial Intelligence (AI) projects a substantial contribution to the management of a plethora of medical issue, also including cancer [3]. Given that a large quantity of medical data and novel computational technologies exist, AI can be applied in multiple aspects of cancer research with emphasis on early diagnosis, patient care, prognosis, treatment response drug discovery optimization, among others. Medical decisions for cancer patient care, both for diagnosis and treatment, heavily rely on cancer imaging data in combination with clinical, histopathological, morphophysiological and other types of data for efficient patient screening. Cancer imaging is considered an umbrella term for several approaches employed in cancer research and diagnosis. This multitude of data that is used for decision making is related with both the volume and origin, and type (source, study design, infrastructure used etc.) of existing data. This can lead to human processing bottlenecks and dealing with, stands a great challenge for the development of computerized solutions.

Besides the need for data management and integration, the diversity of data that originate from multiple sources leads to a growing demand for data harmonization especially in cancer epidemiology especially when combining data sets from heterogeneous studies into one large data set. This can be accomplished with the optimization of currently used strategies and tools for data harmonization and/or the development of new ones to produce compatible and comparable datasets and enhance the high quality and utility of medical data. To date, significant work has been performed not only to construct rigorous and more universally applied strategies for data harmonization [4], [5] but also use the data model to develop data validation tools [6]. However, given that each project conducts harmonization and quality test following its internal needs and standards, there is still a great unmet need for the development of more efficient practices in data harmonization that can be applied in a wide range of studies.

The 42-month INCISIVE project (https://incisive-project.eu/) focuses on these challenges. INCISIVE aims to address the data availability challenge, towards the wide adoption of AI solutions in health imaging. INCISIVE works on the aggregation and unification of 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 data harmonization constitutes a major pillar.

*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).

Data harmonization deals with the identification and removal of any inconsistencies and/or inaccuracies that could originate from multiple sources. This process may have several distinct steps and can be a highly case-specific process. However, in general the key steps involve the a) identification of data sources, data acquisition and collection to produce datasets, b) identification of potential inconsistencies and appropriate modifications for high data quality and harmonization, c) performance of quality test(s) to ensure data validity and integrity, d) identification and selection of uniform variables for harmonization and, e) process of conversion to a common/standard format [7].

In our approach an integration quality check tool is proposed that ensures that all data uploaded to the repository are homogenized, share the same principles, and follow the harmonization rules as previously defined during the data harmonization procedure [8]. This procedure included an iterative process for defining the requirements for the data collection following specific principles [9] and standards and resulted in a set of rules that needs to be followed during the data preparation. This tool attempts to tackle the data harmonization/integration burden, when multiple data sources are involved, making the procedure of creating a pan-European federated repository more efficient. While there are several technical solutions dealing with data quality (e.g., anonymization tools etc.), this tool provides an integrated solution for quality check regarding data coming from multiple sources and different clinical sites. This tool can be considered as essential for the data preparation and homogenization when multisite clinical studies are performed.

II. Design & Implementation

A. Data Structure and Collection

Data collected within the INCISIVE studies are gathered and structured following the protocol established within the project and are aligned with the project's needs and requirements, which have been defined through the data harmonization procedure mentioned above. Data falls into two categories, clinical metadata, and imaging data. The first category is a well-defined and structured template in the '.xls' format, which the user must fill in the information extracted from the PACS system. This format was selected as it supports a semi-structured format, and it is well accepted by end users for editing. Data management within the platform is based on FHIR implementation and this type of data is handled as xml messages. This template contains information with regards to: (i) patient medical history, (ii) diagnosis, (iii) follow-up examinations in specific timepoints, (iv) histopathological examination results, (iv) laboratory examination results. Within the project, a set of mandatory fields were also specified to ensure that data provided can be used for the project's purposes. This set of data consists, among others, of information with regards to dates, cancer staging, tumor details, treatment approach and its response. The second category, imaging data, consists of all imaging examinations performed during the initial diagnosis and the follow-ups. These two categories are linked since the information of what imaging modalities are provided and at what time point is also included in the template.

Data collectors follow a specific procedure for the data collection, preparation, and uploading. This procedure includes the following steps: (i) *Data extraction*. In this step, imaging data are exported from the PACS system and organized in a 'patient folder' structure, meaning that all images from one patient lie under one folder, while no further file structure is needed. (ii) *Data de-identification*. Initially, the original patient id is converted into an alternative one, which is also used at the next step for the mapping between images and clinical metadata. This alternative patient id has a specific structure defined by the project. In addition, any personal identifier which is included in the DICOM header is removed or replaced following the DICOM Standard PS 3.15: Security and System Management Profiles that specify a set of de-identification rules. (iii) *Clinical Metadata collection*. In this step, the template is populated with clinical data. Each patient's information is inserted in the template with the identification provided by the previous step. All dates are relevant and are converted into 'months from diagnosis' with the date of diagnosis to be date zero. This approach is followed to avoid any patient identification issues. The rest of the data is inserted following rules that were decided through workshops. These rules refer to specified value ranges, proper time labeling of each insertion and internationally used medical standards. (iv) *Data quality check*. This step aims to ensure the integrity and consistency of the data, and it constitutes the target of the current paper. More details on this step are provided in the next sections. (v) *Imaging data annotation*. The imaging data are annotated following the principles designated within the project, for tumor and other malignancies annotation and labelling.

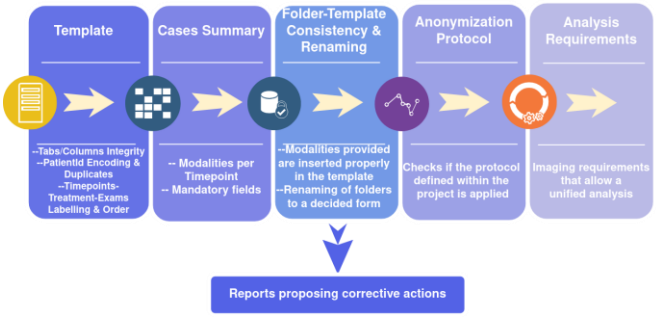

Figure 1. DIQCT workflow

B. Data Integration Quality Check Tool

This tool was developed as a rule-based quality check. Its main purpose is to check whether the data collection requirements are followed and inform the user of potential actions that must be taken to ensure the quality of the data uploaded. Moreover, the tool is extensible, the logic on checking the requirements is not hard-coded, but it is introduced from a knowledge base (specific templates, structures, anonymization protocol). The check is performed in 4 levels, clinical metadata check, images-template consistency check, pseudo-anonymization protocol check, and analysis requirements check. The tool also performs a

case completeness report. The five components of the tool are depicted in Figure 1 and described in detail below:

Clinical Metadata Integrity Check: This component checks the consistency of patient clinical metadata in the following points: (i) Template Structure – Since the templates contain several tabs and numerous fields, there is the need to maintain this structure and naming for harmonization purposes. This tool checks if the initial format is kept when the data is uploaded, (ii) Patient codification – Proper patient codification is checked in two ways, the first one is to ensure that all patient ids are following the structure specified within the project, and the second one is to avoid overlapping ids between patients, (iii) Timing Integrity -- Certain timepoints for the data collection were defined for the studies, so this tool checks the distances between these timepoints and the timing order, and (iv) Content – The data provided must follow specific semantic requirements. This tool checks if the standards and terminologies proposed for certain fields of the templates are followed.

Template-Image Consistency and Renaming: This component checks if the file structure of the images is compliant to the template and performs a proper renaming of the studies' folders so they can be stored in a unified and commonly understandable way, which is a prerequisite in INCISIVE.

Anonymization Protocol Check: The DICOM file format constitutes the standard for medical imaging. Apart from the image, this file format includes additional meta-data about patients, study, or imaging protocol. And thus, this component checks whether the anonymization protocol, already defined in INCISIVE following the respective DICOM Standard, is properly applied in the metadata included on the imaging files, ensuring on one hand that no identifiable data is used and on the other hand any valuable for the analysis information is retained.

Analysis Requirements check: For the analysis, a set of minimum requirements was defined which include, among others, the type of imaging modality that is expected for each cancer type, the pulse sequence used (in the case of MRI) and the slice thickness. The current component checks whether those requirements are met with the goal to allow the technical developers to use data of expected quality for analysis and the training of the algorithms.

Case Completeness Check: This component checks if some minimum requirements are met in each case to be considered as complete in the following points: (i) Timepoints – The tool reports for which timepoints there are available data, (ii) Image Modalities – The tool reports which imaging modalities are provided for each timepoint, and (iii) Mandatory fields – The tool reports if the mandatory fields are provided.

C. Tool Output

The tool does not intervene in the dataset in any other way except for renaming the images folders' names. Instead, it produces 5 different reports, one of each component described above. The four reports are informative and include error messages. These messages inform the user about the issues that the data collector needs to take action to revise the data structure. More specifically, the output is an excel file with 4 tabs referring to each one of the components. The output might include messages related to: (i) patient ids that were found duplicate or not properly encoded, content values that does not comply with the expected value ranges or structural issues, (ii) inconsistencies between the modalities and timepoints provided in the template and the actual folders provided, (iii) DICOM header tags that are present, while they should have been removed (iv) files that do not follow the analysis requirements. The fifth report is a report summarizing the dataset and accompanies the data in the data repository and indicates the existence or not of all mandatory fields for each patient and the modalities provided in each timepoint. An example of these reports is depicted in Figure 2.

D. Implementation

The five components of the DIQCT were developed in two programming languages, R & python and were integrated in one pipeline. This tool is related to prior to uploading work, so it runs on the client side. The tool was initially implemented in two ways: (i) as an executable file (.exe): the pipeline along with all the dependencies was built as a directly executable file (ii) as Docker Image: the pipeline along with all the dependencies was built in a docker container publicly available to all members of the consortium. This way, the pipeline can be executed in all project sites, at the local or central level. Moreover, the tool was converted into a semi-spark version to be able to run the quality checks after the data uploading on the infrastructure and to produce mass reports for the datasets provided from all partners.

E. User Interface

To improve the usability of the DIQCT, a web application was implemented using R programming language and R Studio Shiny server in particular [10], which allows the interactive execution of specific scripts through HTML pages. This application includes the five components described above and the execution of each of them is controlled by the user.

III. VALIDATION

For the validation of the tool 3 rounds of validation were performed, internal and external validation and testing and at the user experience level. In every validation round the tool was updated to cover the needs and deficiencies identified through the process. The rounds of validation performed as well as the questionnaire responses are referring to the initial version of the tool that did not include a user interface.

In the internal validation round, Testing and validation were performed by introducing errors in the template and file structure and in the anonymization process to ensure that the tool identifies every possible mistake on data preparation. This round of internal validation was performed with mock-up data provided by the project partners as example cases.

For the external validation of the tool, four partners contributed as test sites. Two partners were using the executable version and two partners the dockerized version of the tool. The tool was tested locally with real data and the feedback was used to improve the initial version of the tool to a functional level.

At the user experience level, a questionnaire was developed to gather the opinion and feedback of the users with regards to the use of the tool. This questionnaire contains four blocks of questions. The first block focuses on getting the context of the tool usage with regards to the user's background, the tool version used and the operating system. The second block of questions focuses on retrieving the most frequent errors identified by the tool and corrected by the user. The third block of questions is a user experience questionnaire [11] and focuses on retrieving information on 4 scales: Attractiveness, Perspicuity, Efficiency, Dependability. The fourth and final block aims to retrieve information on further improvement of the tool.

IV. Results

A. Web application

The web application is divided into two main areas. On the right, the user can select the folder where the data is located and will be checked for fulfilling the quality standards of the project. The execution of each component is performed by selecting the respective tab on the right, while the outcome of the check is provided in a tabular format. Finally, the user can download the results for further analysis. Figure 2 depicts the web interface, where the user selects the folder which includes the data and the results of the 5 components (right) are provided in a tabular format. In the figure, the results of the template check are shown.

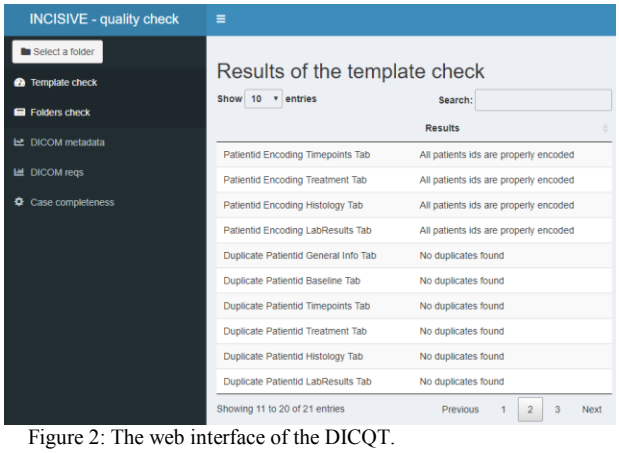

Figure 2: The web interface of the DICQT.

B. Validation

During the validation process and reported also in the specific field in the questionnaire, some common errors were identified, that most of the data collectors did during the data collection. With regards to the clinical metadata collection, one of the most frequently reported errors is that the timepoints that represent the follow-up examinations of the patients are not inside the time boundaries that were set as a requirement. However, this is understood since each patient follows a personalized treatment pathway that may differ based on his/her specific needs, and no corrective action is required. At the same time, with regards to the follow-ups, a label is required to declare the timepoint that the examination took place. It was noticed that these labels were often inserted in the wrong way. Moreover, multiple insertions for the same patient id without properly specifying the label that accompanies the insertion was a common issue during the checking. Finally, the value ranges proposed for specific fields were not followed, e.g., yes or no fields should be translated to 1 or 0 and some essential fields for the process were left blank. Regarding the DICOM file structure, the most common error was that the number of studies reported in the template did not match the number of folders inside the patient directory, which means that the user did not perform a proper matching between the template and the images provided. The most frequent reported error with regards to the anonymization was that actions that were supposed to be applied in some fields, were not. This might have derived from the fact that different protocols are applied in different data centers.

Finally, during the internal and external validation process some issues were identified and corrected by the team by updating the tool. Such issues were inherited by changes in the anonymization protocol used and, consequently, from the output of the anonymizer, e.g., changes in the patient codification, modalities such as CR, DX, and the transformation of some studies, e.g., the division of the fused PET-CT images that were not initially included in the requirements. Another issue was that the data collectors requested that they could upload studies that come from visits 'outside' the specified ones, so an alternative way for inserting these data had to be included in the process.

C. User Experience Evaluation

A total of 11 partners participated in the survey. 4 of the participants belong to technical teams while 2 of them are medical experts. 3 of them are using the dockerized version while the other 3 are using the executable version. Finally, 5 of them are using Windows operating systems, and 1 of them uses Linux. Figure 3 depicts the answers to the user experience part of the questionnaire for the 4 distinct categories. The values represented for each feature are the mean value of the votes in the range 1 to 5. In the attractiveness category, it is obvious that the tool may not be so user-friendly, but the general opinion is positive. From the

perspicuity category, it is concluded that the tool is easy to use and learn, understandable but not so clear. In terms of efficiency, we can say that the tool is efficient and practical but not so fast. Finally, the tool meets the expectations of the user, and it is considered supportive, although the users do not recognize the value of this tool.

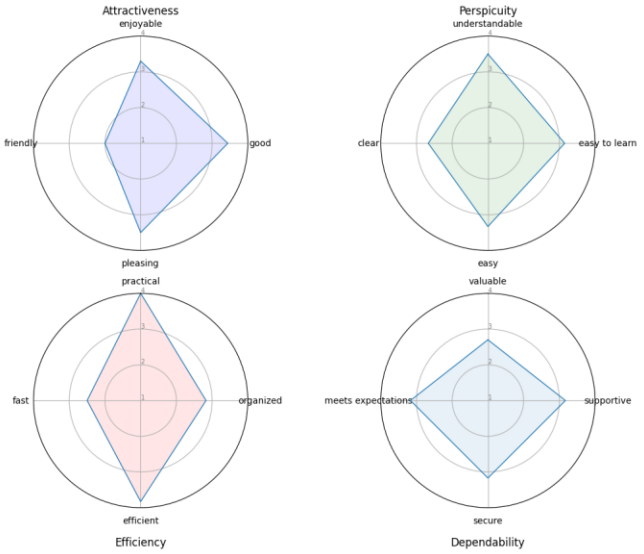

Figure 3. User experience results.

V. Conclusion

Harmonization is an essential step to improve the validity of AI modelling and the trustworthiness of multi-source data. By deploying a tool that ensures the homogeneity of data in a multicentric repository, it reduces the necessary work of data preparation prior to analysis. During this procedure we identified some common issues deriving from the time-consuming process of data collection. These issues mostly prove that following specific requirements for data preparation is not an easy procedure, since different organizations and/in different countries follow different standards for both imaging and clinical metadata storing. This tool can also be efficient in the identification of potential issues regarding the heterogeneity of the data collected from different clinical sites and thus it can later be used for the harmonization of data. Our testing proved that such issues can be identified and avoided to their genesis.

Considering the users' experience, one of the most poorly voted attributes with regards to the tool's attractiveness, was that the tool seemed unfriendly to most of the users. Taking into consideration that the questionnaire was circulated without the existence of a user interface, it is justified. So, we proceeded by adding a web interface that provides an easier way of running the tool and improves its usability. With this addition we also provide better visualization of the tool's outputs by presenting them in a more clear, human-readable and understandable way through the user interface. Finally, with regards to the tool's efficiency, users think that the execution is not too fast. However, this tool performs several actions which involve the processing of a great amount of DICOM images and clinical metadata. Summarizing, our future steps include a further evaluation of the DIQCT tool, also including the user interface, and further optimization for better efficiency.

Acknowledgment

We thank the 9 Data Providers participating in the INCISIVE project for participating in the validation and evaluation process: Aristotle University of Thessaloniki, University of Novisad, Visaris D.O.O, University of Naples Federico II, Hellenic Cancer Society, University of Rome Tor Vergata, University of Athens, Consorci Institut D'Investigacions Biomediques August Pi i Sunyer, Linac Pet-scan Onco limited.

References

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020 (https://gco.iarc.fr/today, accessed January 2022).

[2] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[3] Z. H. Chen, L. Lin, C. F. Wu, C. F. Li, R. H. Xu, and Y. Sun, "Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine," Cancer Communications, vol. 41, no. 11. John Wiley and Sons Inc, pp. 1100–1115, Nov. 01, 2021. doi: 10.1002/cac2.12215.

[4] B. Rolland et al., "Toward Rigorous Data Harmonization in Cancer Epidemiology Research: One Approach," American Journal of Epidemiology, vol. 182, no. 12, pp. 1033–1038, Jul. 2015, doi: 10.1093/aje/kwv133.

[5] R. Pomponio et al., "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," NeuroImage, vol. 208, Mar. 2020, doi: 10.1016/j.neuroimage.2019.116450.

[6] N. C. Nicholson et al., "An ontology-based approach for developing a harmonised data-validation tool for European cancer registration," Journal of Biomedical Semantics, vol. 12, no. 1, Dec. 2021, doi: 10.1186/s13326-020-00233-x.

[7] P. Avillach et al., "Harmonization process for the identification of medical events in eight European healthcare databases: The experience from the EU-ADR project," Journal of the American Medical Informatics Association, vol. 20, no. 1, pp. 184–192, 2013, doi: 10.1136/amiajnl-2012-000933.

[8] A. Kosvyra, D. Filos, D. Fotopoulos, T. Olga, and I. Chouvarda, "Towards Data Integration for AI in Cancer Research *," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Nov. 2021, pp. 2054–2057. doi: 10.1109/EMBC46164.2021.9629675.

[9] Otto. Visser et al., A proposal on cancer data quality checks one common procedure for European cancer registries.

[10] Shiny from R Studio. RStudio, PBC. URL: http://shiny.rstudio.com/ [accessed 2022-01-19]

[11] B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire.," in USAB, 2008, vol. 5298, pp. 63–76.