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

A. KOSVYRA[1], D. FILOS[1], D. FOTOPOULOS[1], O. TSAVE[1] and I. CHOUVARDA[1]
1Aristotle University of Thessaloniki, Thessaloniki, Greece

INTRODUCTION

- Data harmonization deals with the identification and removal of any inconsistencies and/or inaccuracies that could originate from the fact that data come from multiple sites
- Key steps in the process: a) identification of data sources, data acquisition and collection, b) identification of potential inconsistencies and appropriate modifications, c) performance of quality tests 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 [1]
- Our approach focuses on the development of a Data Integration Quality Check Tool, for cancer imaging repositories that include multiple modalities and timepoints, and clinical and biological data.

AIM

This tool runs locally by the data provider to:

- Ensure 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 [2]
- Attempt to tackle the data harmonization/integration burden, when multiple data sources are involved
- Provide an integrated solution for quality check regarding data coming from multiple sources and different clinical sites

Essential for the data preparation and homogenization when multisite clinical studies are performed

ACKNOWLEDGEMENTS

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.

FUNDING

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

REFERENCES

[1] 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,

[2] 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.

DESIGN & IMPLEMENTATION

- Rule-based tool, built from scratch in Python & R programming languages.
- Implemented as a Docker image and executable file.
- Checks if the data collection requirements are followed and informs the user on corrective actions prior to data upload
- Extensible, not hard-coded check but introduced as a knowledge base (specific templates, structures, anonymization protocol).
- 5 components:

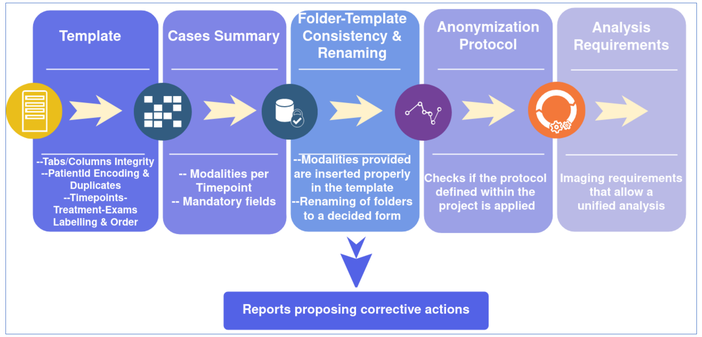

Figure 1. The DIQCT's components.

DISCUSSION - CONCLUSIONS

RESULTS

Tool Output:

- 5 different reports, one of each component described in Figure 1
- Error messages for corrective actions
- No intervention on the dataset by the tool, the errors are corrected by the user

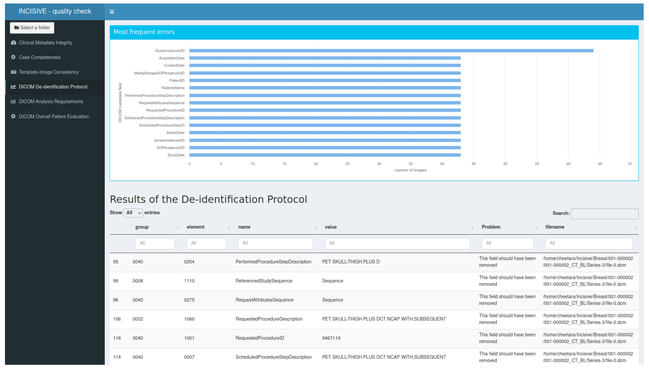

Figure 2. The web interface of the DICQT. Interactive user interface implemented using R Shiny Server. Runs the five components and depicts the results in different tabs. The current tab shows the results of the DICOM de-identification check component, with the bar plot on the top showing the most frequent findings.

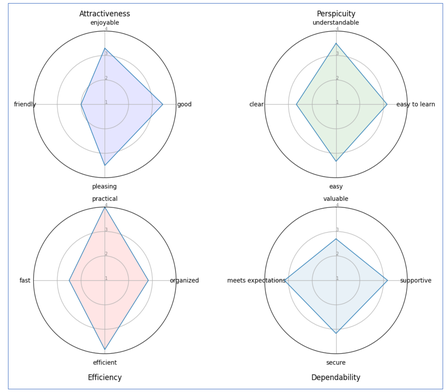

Figure 3. User experience evaluation results.

Validation:

- Internal testing by introducing errors to the mock-up datasets provided by the clinical partners
- External testing with Data Providers and optimization by including the feedback in the development
- User experience and commonly reported errors through a questionnaire circulated to all users. Some of the most reported errors were that values inserted didn't follow the allowable types/value ranges, number of images provided did not match with the information provided in the template and inconsistencies in the applied de-identification protocol.

CONTACT INFORMATION

Mail: aekosvyra@auth.gr LinkedIn: www.linkedin.com/in/alexandra-kosvyra-2599805a Twitter: @kosvyra