Ref. Ares(2023)4555195 - 30/06/2023

A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology

# Deliverable D4.3:
# Quality control of image and non-image cancer data

| | |
|---|---|
| Reference | D4.3_ EuCanImage_institute_UAMS |
| Lead Beneficiary | University of Arkansas for Medical Sciences (UAMS) |
| Author(s) | Prior, Fred (UAMS)<br>Rutherford, Michael (UAMS)<br>Bona, Jonathan (UAMS)<br>Maciej Bobowicz (GUMED)<br>Rygusik, Marlena (GUMED)<br>Teresa Garcia Lezana (CRG)<br>Aldar Cabrelles Munoz (CRG)<br>Santiago Andres Frid (FCRB)<br>Martijn Starmans (EMC)<br>Ivan Bocharov (EMC)<br>Alexander Harms (EMC) |
| Dissemination level | Public |
| Type | Report |
| Official Delivery Date | June 30, 2023 |
| Date of validation of the WP leader | |
| Date of validation by the Project Coordinator | |
| Project Coordinator Signature | |

EuCanImage is funded by the European Union's H2020 Framework
Under Grant Agreement No 952103

## Version log

| Issue Date | Version | Involved | Comments |
|---|---|---|---|
| 26/06/2023 | V1.0 | Prior, Fred | Draft version for review |
| 28/06/2023 | V2.0 | Prior, Fred | Final Draft |

## Executive Summary

Deliverable 4.3 aims to define for EuCanImage appropriate tools and procedures for quality control (QC) of imaging data and associated clinical (non-imaging) data to meet Objective 4.3 "Expand POSDA tools and curation procedures to curate labelled data and non-imaging data." As the program developed it became clear that multiple approaches were required, for non-image data, but a common set of components could be shared across these approaches and integrated into POSDA. It was also concluded that automated image quality assessment tools would be more efficiently implemented as containerized plug-ins to the XNAT-based Euro-BioImaging repository. With slight changes to the plug-in interface, the same tools are deployable as extensions to POSDA.

The document is structured in three main parts: (1) **Quality Control of Non-image Data**. Because of the wide variance in the capabilities of clinical sites to provide non-image data, three data collection strategies were undertaken. At Hospital Clinic de Barcelona, data was easily accessible and tools already existed to map this data to the EuCanImage data model so only a shared set of rules for QC was required. At GUMED, data export from the hospital EHR was possible and tools were created to ingest the exports, present the data for structured selection, QC, and export of the data in the format needed for the final submission. The default option was manual chart review with data capture via a REDCap[1] eCRF followed by QC using the common QC rules and a built-in QC tool. (2) **Quality Control of Image Data**. A containerized tool that retrieves acquisition protocols from each image series in the EuCanImage collection in the Euro-BioImaging repository, creates a sample of 10% of the images in each series with a minimum of 10 samples or the entire series if smaller and calculates the PIQUE[2] metric for each. The average PIQUE score and series acquisition protocol are stored in XNAT as a descriptor of that series. (3) **Final Considerations**. Data collection and QC is ongoing at all clinical sites and the tools described here continue to be enhanced as new situations arise.

## Table of Contents

## Acronyms

| Name | Abbreviation |
|---|---|
| Perl Open Source DICOM Archive | POSDA |
| Extensible Neuroimaging Archive Toolkit | XNAT |
| The Cancer Imaging Archive | TCIA |
| Digital Imaging and Communications in Medicine (ISO 12052:2017) | DICOM® |
| Collective Minds Radiology | CMRAD |
| European Union | EU |
| Research Electronic Data Capture | REDCap |
| Quality Control | QC |
| Perception based Image Quality Evaluator | PIQUE (or PIQE) |

# 1 Introduction

This report will contribute to the achievement of the final goal of EuCanImage, the creation of a GDPR compliant integrated platform for large-scale cancer imaging, and AI solutions. Our initial task was to determine the relative capabilities of our clinical sites to extract non-image data relevant for each EuCanImage use case and working with Work Packages 2 and 3. Each site had different capabilities requiring the development of multiple tailored solutions to capture non-image data and perform quality checks. Image data quality evaluation happens naturally as part of the annotation process.  Additional, automated quality metrics are computed within the XNAT framework of the Euro-BioImaging repository and captured in XNAT to provide guidance to future users of the image data.

Quality Assurance refers to data collection procedures which must be designed to ensure data is collected in conformance to the requirements of the project. Quality Control refers to those direct steps taken to assess and improve data quality once the data has been collected. Both techniques are employed. For simplicity, in the text that follows, we will use QC as the general term to encompass both assurance and control.

# 2 Quality Control of Non-Image Data

Information quality may be defined as data that is fit for purpose[3], e.g., the data are sufficient for a specified research need although there are many other dimensions of data and information quality[4]. Data quality issues can be introduced at many points in the data collection and management cycle, e.g., during data collection and integration, storage and management, analysis, and publishing and sharing. Our approach is to define a standard set of tools and procedures for data QC and integrate these into the pipeline for data collection and storage. We focus on automation but retain a human in the loop approach at the data controller to facilitate error correction.

## 2.1 Methods of Data Collection

Due to the various capabilities for data extraction at each site, multiple methods of data collection were deployed. Some sites already housed structured elements, facilitating easy extraction, while others had no extraction capabilities at all, requiring manual chart review. We pursued multiple efforts to accommodate the various levels of extraction.

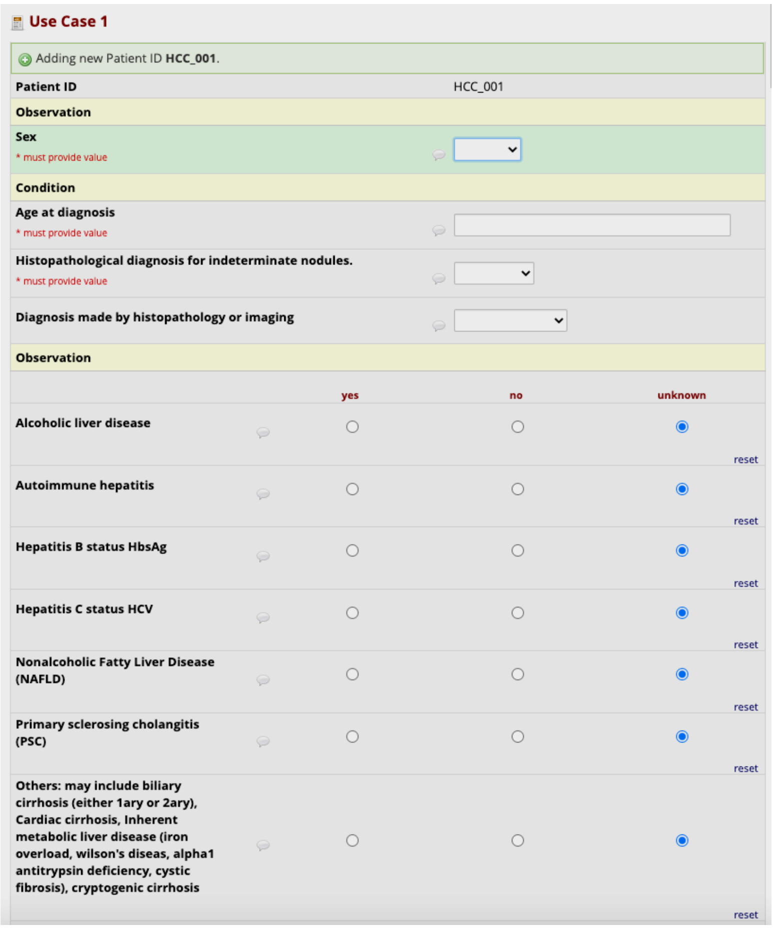
At the heart of the data ingestion effort is the REDCap database. We use REDCap as the intermediary data store where all other data collection methods funnel their data. Those with structured data simply reformat their data to match the format for ingestion by REDCap, while others with no capabilities manually enter their data into the REDCap eCRF. We also pursued a customized solution for presenting and extracting data that falls in the middle of the scale for a site which was able to export JSON files.

### 2.1.1 Basic method: REDCap eCRF

With the completion of the final data model, CRG created of a REDCap[1] form for ingesting non-image data. This consisted of a detailed form for each use case with QA/QC built in using the REDCap data quality module. REDCap is a secure web application for building custom eCRFs and manage online data capture mainly for clinical research studies. Its server is hosted by the EGA@CRG. REDCap has a user-friendly interface to design the data entry forms that allows field validation, custom logic patterns, calculated fields, data import/export options, data quality control and role-based user access. It also has a set of APIs for integration with other platforms. We created five different eCRFs (example in figure 1) with appropriate custom QC rules to cover the needs of the different use cases.

Data entry from hospitals into REDCap can be performed in two different pathways: 1) by direct filling of REDCap form on the web or 2) by filling specific csv templates with predetermined codes (specified in the data dictionary). As a result of this process, for each use case, all the related clinical data coming from the different institutions will merge in a single harmonized database (figure 2). Once all datasets are collected, an ETL process will transform the database, exported as CSV file, to a FHIR compliant JSON file. Given that data is captured in a structured way under QC procedures, we expect that the potential number of errors will be low.

REDCap has a data quality module that allows the execution of QC rules to check for discrepancies in the data. The QC analysis can be performed in real-time in case of direct filling of the data or after submission in case of csv import. Quality control rules include two levels: a) pre-defined data rules, standard rules pre-established by the app b) custom rules, designed to fulfill the specific needs of the project.



Use Case 1

Adding new Patient ID **HCC_001**.

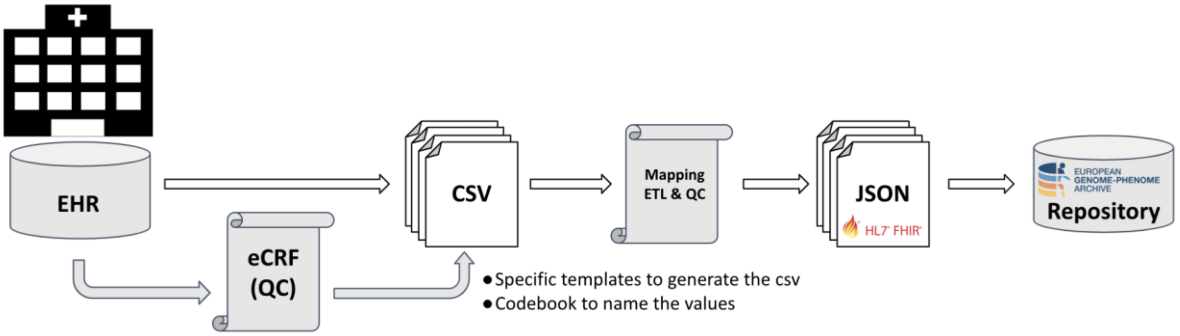
| | |
|---|---|
| **Patient ID** | HCC_001 |
| **Observation** | |
| **Sex**<br>* must provide value | |
| **Condition** | |
| **Age at diagnosis**<br>* must provide value | |
| **Histopathological diagnosis for indeterminate nodules.**<br>* must provide value | |
| **Diagnosis made by histopathology or imaging** | |

| **Observation** | yes | no | unknown |
|---|---|---|---|
| **Alcoholic liver disease** | ○ | ○ | ● |
| **Autoimmune hepatitis** | ○ | ○ | ● |
| **Hepatitis B status HbsAg** | ○ | ○ | ● |
| **Hepatitis C status HCV** | ○ | ○ | ● |
| **Nonalcoholic Fatty Liver Disease (NAFLD)** | ○ | ○ | ● |
| **Primary sclerosing cholangitis (PSC)** | ○ | ○ | ● |
| **Others: may include biliary cirrhosis (either 1ary or 2ary), Cardiac cirrhosis, Inherent metabolic liver disease (iron overload, wilson's diseas, alpha1 antitrypsin deficiency, cystic fibrosis), cryptogenic cirrhosis** | ○ | ○ | ● |

Figure 1: Screenshot from REDCap eCRF

Pre-established data quality rules include:

- Blank values: for mandatory (required) fields
- Field validation errors: for incorrect data types
- Outliers: for numerical fields
- Hidden fields that contain values: refers to any fields on a survey or data entry form that are not being displayed because branching logic is hiding them, which assumes that the field's value should be blank/null.
- Multiple choice fields with invalid values
- Incorrect values for calculated fields
- Fields containing "missing data codes"



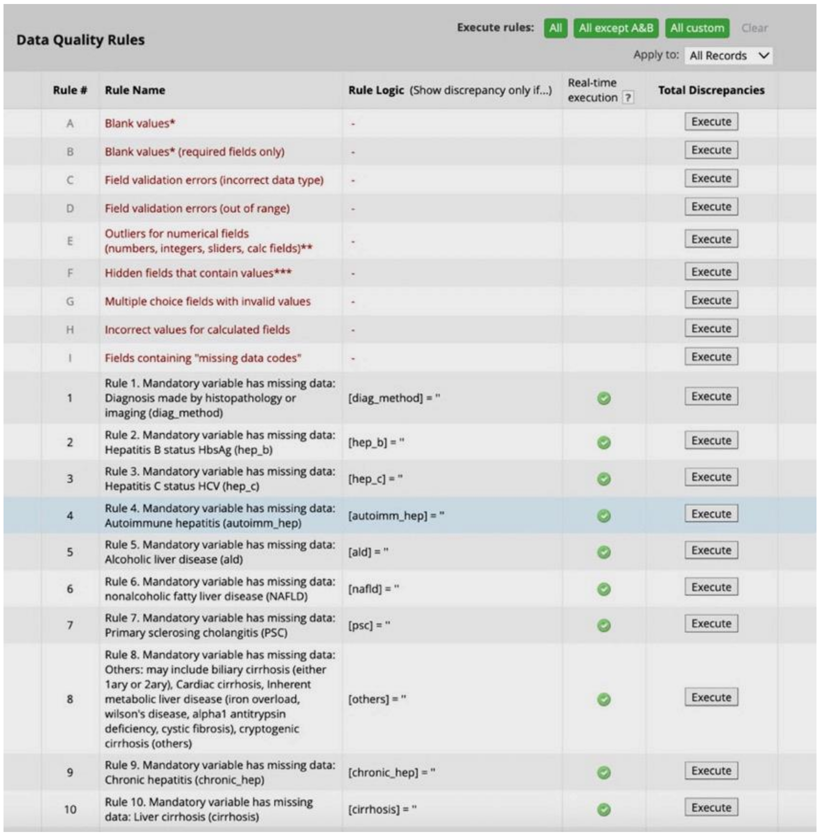
- Specific templates to generate the csv
- Codebook to name the values

Figure 2. Current data flow of the non-imaging data from clinical centers to EGA. Relevant data is captured at hospitals using REDCap, by direct data entry or by uploading a csv. The REDCap output is a csv file that will undergo a mapping ETL and QC process to obtain a FHIR-compliant json file that can be submitted at the EGA repository.

Beyond the pre-established rules we have created, depending on the use case, from 10 to 30 custom QC rules mainly focus on covering the clinical correctness of the data and creating the logic to adequately provide the three levels of clinical information (minimal, mandatory and recommended). These QC rules are also aligned with the need of collecting all minimal fields by all clinical centers to obtain unbiased data (figure 3).

### 2.1.2 Semi-automated data collection tools at GUMED

At the GUMED clinical site, we are piloting the use of a novel clinical data collection tool, a clinical data bus, developed by GUMED in collaboration with the University Clinical Center, that ingests data extracted from their EMR system. The purpose of this tool is to facilitate collection of clinical data from electronic medical records for the EuCanImage project, and to support semantic integration of data from disparate sources (clinical sites submitting data) by aligning collected data with the common data model for the project, while providing real-time quality assurance checks on these data as they are collected.



| Rule # | Rule Name | Rule Logic (Show discrepancy only if...) | Real-time execution ? | Total Discrepancies |
|---|---|---|---|---|
| A | Blank values* | - | | Execute |
| B | Blank values* (required fields only) | - | | Execute |
| C | Field validation errors (incorrect data type) | - | | Execute |
| D | Field validation errors (out of range) | - | | Execute |
| E | Outliers for numerical fields (numbers, integers, sliders, calc fields)** | - | | Execute |
| F | Hidden fields that contain values*** | - | | Execute |
| G | Multiple choice fields with invalid values | - | | Execute |
| H | Incorrect values for calculated fields | - | | Execute |
| I | Fields containing "missing data codes" | - | | Execute |
| 1 | Rule 1. Mandatory variable has missing data: Diagnosis made by histopathology or imaging (diag_method) | [diag_method] = '' | ✓ | Execute |
| 2 | Rule 2. Mandatory variable has missing data: Hepatitis B status HbsAg (hep_b) | [hep_b] = '' | ✓ | Execute |
| 3 | Rule 3. Mandatory variable has missing data: Hepatitis C status HCV (hep_c) | [hep_c] = '' | ✓ | Execute |
| 4 | Rule 4. Mandatory variable has missing data: Autoimmune hepatitis (autoimm_hep) | [autoimm_hep] = '' | ✓ | Execute |
| 5 | Rule 5. Mandatory variable has missing data: Alcoholic liver disease (ald) | [ald] = '' | ✓ | Execute |
| 6 | Rule 6. Mandatory variable has missing data: nonalcoholic fatty liver disease (NAFLD) | [nafld] = '' | ✓ | Execute |
| 7 | Rule 7. Mandatory variable has missing data: Primary sclerosing cholangitis (PSC) | [psc] = '' | ✓ | Execute |
| 8 | Rule 8. Mandatory variable has missing data: Others: may include biliary cirrhosis (either 1ary or 2ary), Cardiac cirrhosis, Inherent metabolic liver disease (iron overload, wilson's disease, alpha1 antitrypsin deficiency, cystic fibrosis), cryptogenic cirrhosis (others) | [others] = '' | ✓ | Execute |
| 9 | Rule 9. Mandatory variable has missing data: Chronic hepatitis (chronic_hep) | [chronic_hep] = '' | ✓ | Execute |
| 10 | Rule 10. Mandatory variable has missing data: Liver cirrhosis (cirrhosis) | [cirrhosis] = '' | ✓ | Execute |

Figure 3. Screenshot from Quality control module in REDCap

The capabilities at GUMED to extract their EMR data to JSON files, preserving the original data structure of the hospital system, enabled the creation of a semi-automated tool for traversing the multiple use cases, importing the JSON files and displaying the needed sections as the form is being filled out. Patient data are anonymized on site in the process of extracting them to JSON files, which are then loaded into the clinical data collection tool. The tool extracts key elements, allowing the user to add or modify as needed, and exports the anonymized data in the pre-defined format for upload into the REDCap repository. (Figure 4).

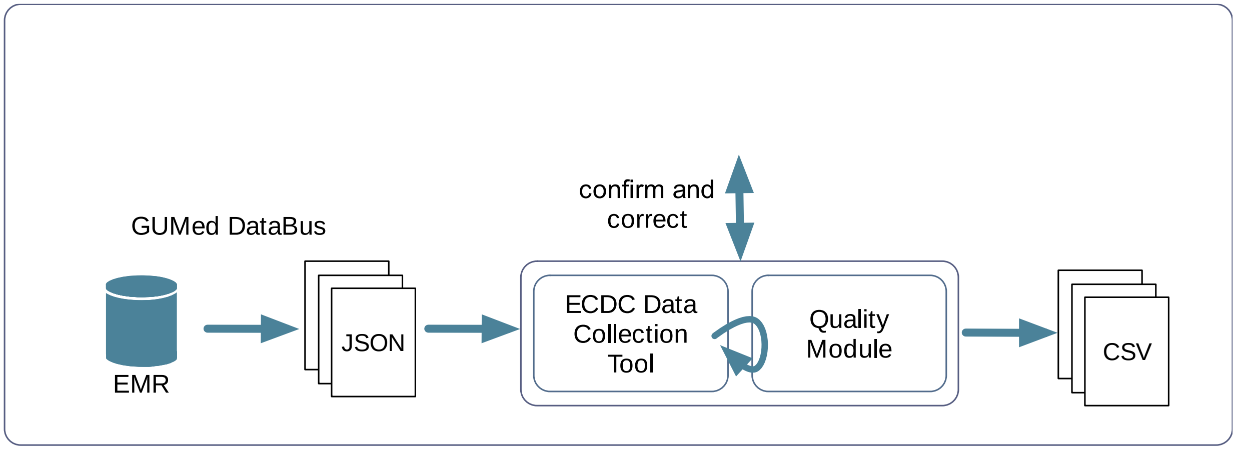


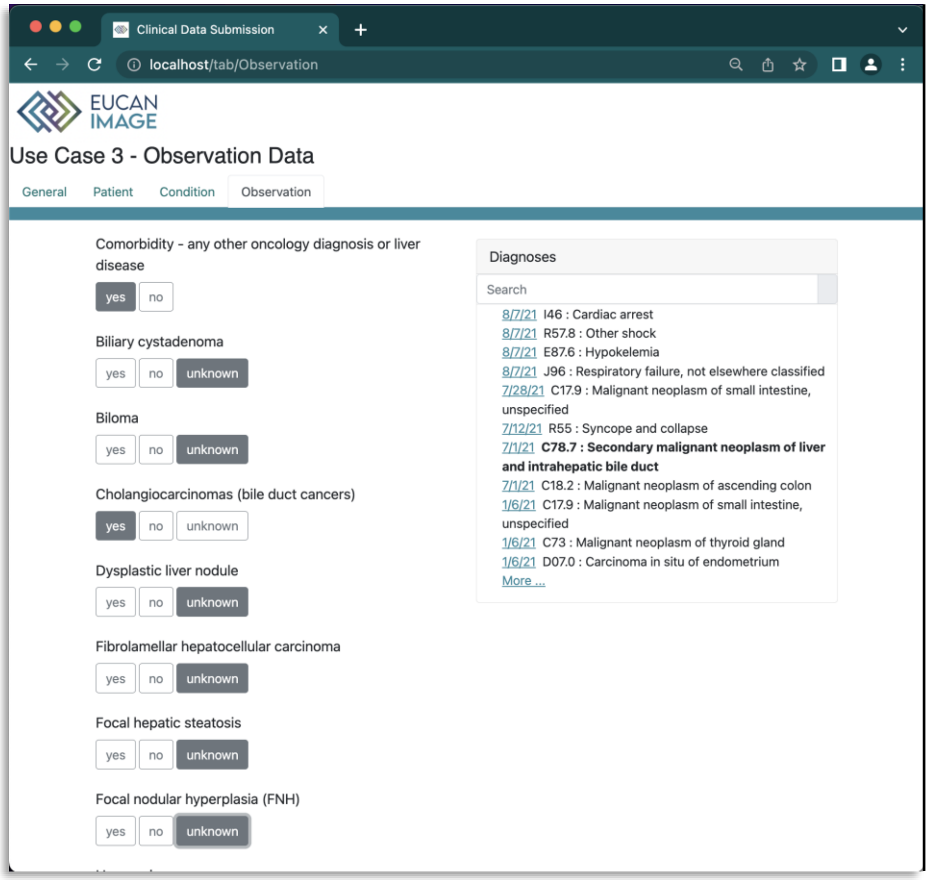
Figure 4. workflow for the ECDC tool. EMR data extracted as JSON files are loaded into the tool. The user interacts with the tool's interface to confirm and augment the collected data. Data quality checks are run automatically. Finished data can be exported to a CSV file aligned with the EuCanImage data model, and submitted to the REDCap repository.

Some data elements required for EuCanImage use cases are directly extracted from the clinical data files, e.g. patient sex. Other elements need to be calculated/inferred based on some rules, e.g. age at diagnosis may need to be calculated from the patient's age and the date of the patient's cancer diagnosis. Other elements can be found only in free text fields. An example is tumor staging values (TNM), found in pathology reports. In cases like this, the data collection tool assists the user in locating, highlighting, and extracting the appropriate element from the text. (Figure 5)

The data collection tool is built using Python, the Flask web framework, and MongoDB database system, distributed as a portable Docker container. By distributing the tool as a Docker container, which can be deployed on any machine that has the Docker tool installed, we support 100% local collection of the data, which is securely stored within the container on the user's computer, on-site. The forms for each use case are dynamically generated from configuration files which are created based on the REDCap codebook ensuring this tool presents the same values and exports resulting records in the form needed to import into the REDCap database.

### 2.1.3 Automated procedures at Hospital Clinic de Barcelona

The Hospital Clinic de Barcelona has managed what many other institutions strive for, a semantically rich dataset based on inputs and confirmations by clinicians at input rather than attempting to code and translate retrospectively. This information is stored in an ontology-based clinical repository called OntoCR. Using SPARQL queries, data is extracted into csv files, and the cohorts are selected according to the mandatory variables for each use case. Using the data dictionary and data quality rule spreadsheets generated from the REDCap codebook, the structured data is then translated and structured into the CSV template with the format needed to import data directly into the REDCap database.



Use Case 3 - Observation Data

General | Patient | Condition | Observation

Comorbidity - any other oncology diagnosis or liver disease
[✓ yes] no

Biliary cystadenoma
yes no [✓ unknown]

Biloma
yes no [✓ unknown]

Cholangiocarcinomas (bile duct cancers)
[✓ yes] no unknown

Dysplastic liver nodule
yes no [✓ unknown]

Fibrolamellar hepatocellular carcinoma
yes no [✓ unknown]

Focal hepatic steatosis
yes no [✓ unknown]

Focal nodular hyperplasia (FNH)
yes no [✓ unknown]

Diagnoses

Search

8/7/21 I46 : Cardiac arrest
8/7/21 R57.8 : Other shock
8/7/21 E87.6 : Hypokelemia
8/7/21 J96 : Respiratory failure, not elsewhere classified
7/28/21 C17.9 : Malignant neoplasm of small intestine, unspecified
7/12/21 R55 : Syncope and collapse
7/1/21 **C78.7 : Secondary malignant neoplasm of liver and intrahepatic bile duct**
7/1/21 C18.2 : Malignant neoplasm of ascending colon
1/6/21 C17.9 : Malignant neoplasm of small intestine, unspecified
1/6/21 C73 : Malignant neoplasm of thyroid gland
1/6/21 D07.0 : Carcinoma in situ of endometrium
More ...

*Figure 5. ECDC data collection tool. At right are example patient diagnoses loaded from an EMR JSON file. At left are EuCanImage use case 3 data elements. Where possible, the tool automatically populates these elements based on EMR data, and allows the user to change or add selected answers.*

## 2.2 Common QC Metrics

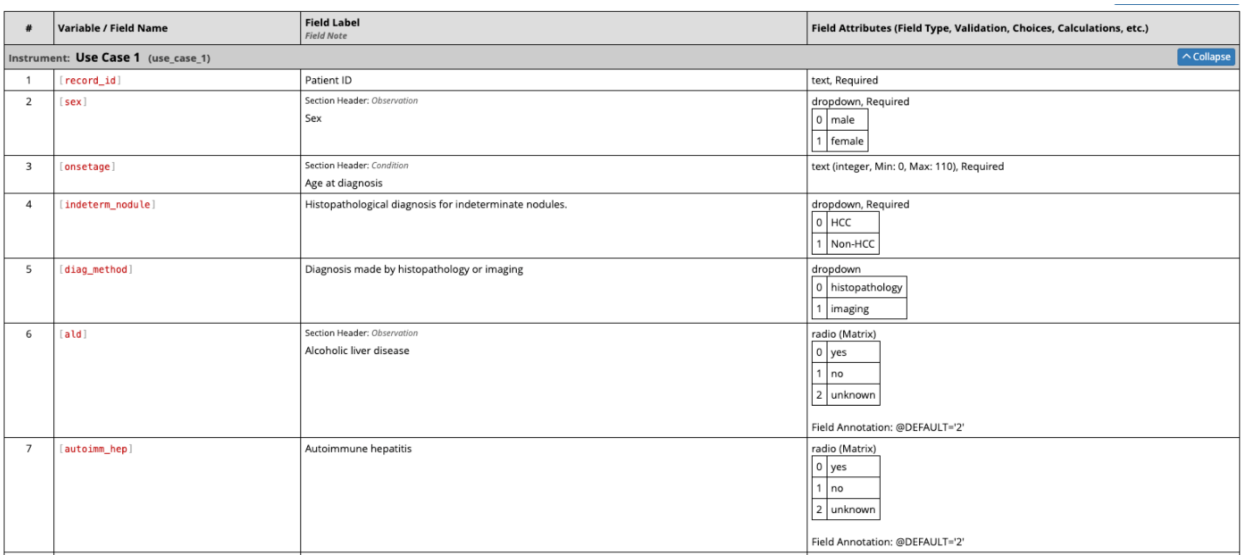
Data quality can be evaluated over many different dimentions[4]. We have focused on three critical dimensions for our evaluation: Completeness, Conformance and Plausibility. Specifically, the assessments applied address the following questions:

- **Completeness**
  - Are required values present?
- **Conformance**
  - Do values conform to syntactic and semantic requirements?
  - Do values conform to relational or structural constraints?
  - Do computed values match calculations on values from which they are derived?
- **Plausibility**

- Are objects such as entities and observations unique when they should be?
- Do observed data values, distributions, or densities agree with common knowledge or gold standards?
- Do time-varying variables change as expected based on known temporal properties?

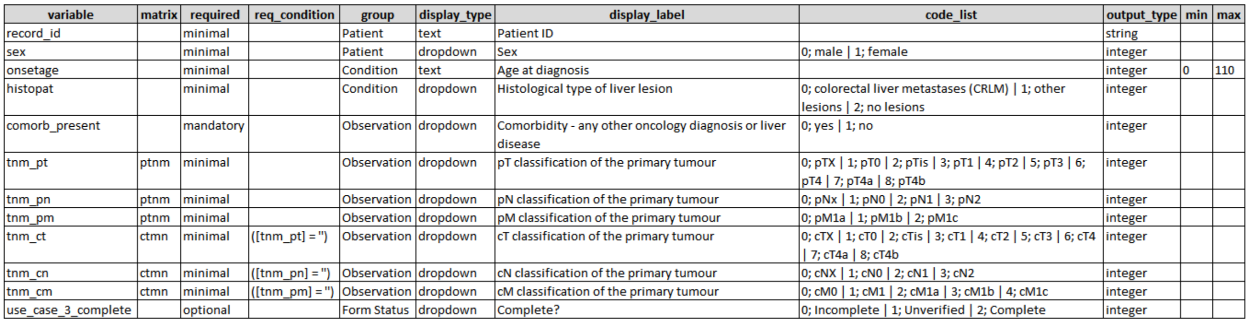
### 2.2.1 Data Dictionary

At the heart of any data project is the data dictionary. The codebook (Figure 6) generated from the REDCap application is utilized to generate a computationally ingestible data dictionary which is used to define the specific quality assessments for each use case.



| # | Variable / Field Name | Field Label<br>*Field Note* | Field Attributes (Field Type, Validation, Choices, Calculations, etc.) |
|---|---|---|---|
| Instrument: **Use Case 1** (use_case_1) | | | ^ Collapse |
| 1 | [record_id] | Patient ID | text, Required |
| 2 | [sex] | Section Header: *Observation*<br>Sex | dropdown, Required<br>0 male<br>1 female |
| 3 | [onsetage] | Section Header: *Condition*<br>Age at diagnosis | text (integer, Min: 0, Max: 110), Required |
| 4 | [indeterm_nodule] | Histopathological diagnosis for indeterminate nodules. | dropdown, Required<br>0 HCC<br>1 Non-HCC |
| 5 | [diag_method] | Diagnosis made by histopathology or imaging | dropdown<br>0 histopathology<br>1 imaging |
| 6 | [ald] | Section Header: *Observation*<br>Alcoholic liver disease | radio (Matrix)<br>0 yes<br>1 no<br>2 unknown<br>Field Annotation: @DEFAULT='2' |
| 7 | [autoimm_hep] | Autoimmune hepatitis | radio (Matrix)<br>0 yes<br>1 no<br>2 unknown<br>Field Annotation: @DEFAULT='2' |

*Figure 6.* REDCap Codebook example

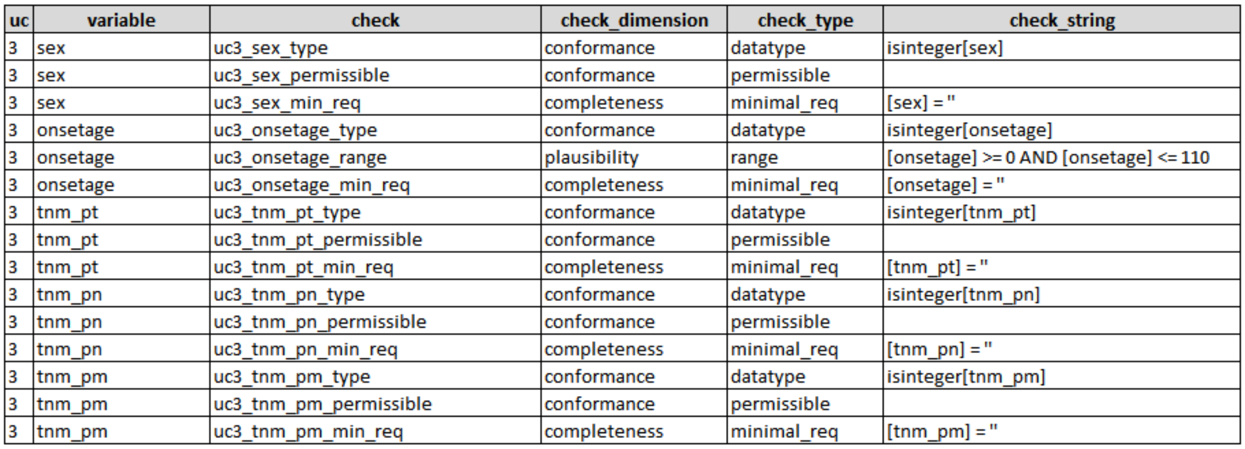


| variable | matrix | required | req_condition | group | display_type | display_label | code_list | output_type | min | max |
|---|---|---|---|---|---|---|---|---|---|---|
| record_id | | minimal | | Patient | text | Patient ID | | string | | |
| sex | | minimal | | Patient | dropdown | Sex | 0; male \| 1; female | integer | | |
| onsetage | | minimal | | Condition | text | Age at diagnosis | | integer | 0 | 110 |
| histopat | | minimal | | Condition | dropdown | Histological type of liver lesion | 0; colorectal liver metastases (CRLM) \| 1; other lesions \| 2; no lesions | integer | | |
| comorb_present | | mandatory | | Observation | dropdown | Comorbidity - any other oncology diagnosis or liver disease | 0; yes \| 1; no | integer | | |
| tnm_pt | ptnm | minimal | | Observation | dropdown | pT classification of the primary tumour | 0; pTX \| 1; pT0 \| 2; pTis \| 3; pT1 \| 4; pT2 \| 5; pT3 \| 6; pT4 \| 7; pT4a \| 8; pT4b | integer | | |
| tnm_pn | ptnm | minimal | | Observation | dropdown | pN classification of the primary tumour | 0; pNx \| 1; pN0 \| 2; pN1 \| 3; pN2 | integer | | |
| tnm_pm | ptnm | minimal | | Observation | dropdown | pM classification of the primary tumour | 0; pM1a \| 1; pM1b \| 2; pM1c | integer | | |
| tnm_ct | ctmn | minimal | ([tnm_pt] = '') | Observation | dropdown | cT classification of the primary tumour | 0; cTX \| 1; cT0 \| 2; cTis \| 3; cT1 \| 4; cT2 \| 5; cT3 \| 6; cT4 \| 7; cT4a \| 8; cT4b | integer | | |
| tnm_cn | ctmn | minimal | ([tnm_pn] = '') | Observation | dropdown | cN classification of the primary tumour | 0; cNX \| 1; cN0 \| 2; cN1 \| 3; cN2 | integer | | |
| tnm_cm | ctmn | minimal | ([tnm_pm] = '') | Observation | dropdown | cM classification of the primary tumour | 0; cM0 \| 1; cM1 \| 2; cM1a \| 3; cM1b \| 4; cM1c | integer | | |
| use_case_3_complete | | optional | | Form Status | dropdown | Complete? | 0; Incomplete \| 1; Unverified \| 2; Complete | integer | | |

Figure 7. Data Dictionary Subset example

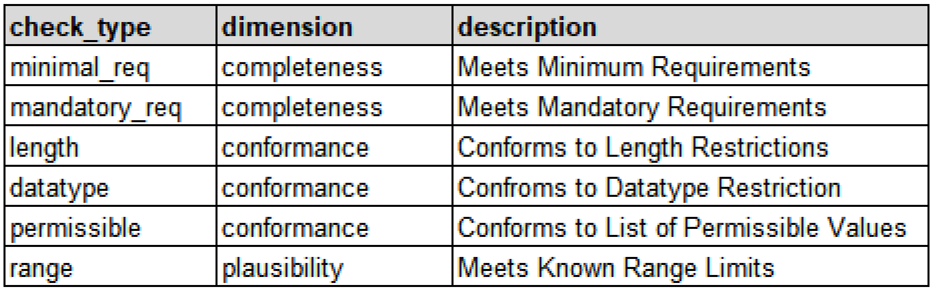
### 2.2.2 Quality Assessment

Specific quality assessments are generated for each quality dimension for each variable in the data dictionary (Figure 7). The assessments are organized by quality dimension and check type (Figure 8). These check types are organized into multiple assessment organizing principles listed in Figure 9.



| uc | variable | check | check_dimension | check_type | check_string |
|---|---|---|---|---|---|
| 3 | sex | uc3_sex_type | conformance | datatype | isinteger[sex] |
| 3 | sex | uc3_sex_permissible | conformance | permissible | |
| 3 | sex | uc3_sex_min_req | completeness | minimal_req | [sex] = '' |
| 3 | onsetage | uc3_onsetage_type | conformance | datatype | isinteger[onsetage] |
| 3 | onsetage | uc3_onsetage_range | plausibility | range | [onsetage] >= 0 AND [onsetage] <= 110 |
| 3 | onsetage | uc3_onsetage_min_req | completeness | minimal_req | [onsetage] = '' |
| 3 | tnm_pt | uc3_tnm_pt_type | conformance | datatype | isinteger[tnm_pt] |
| 3 | tnm_pt | uc3_tnm_pt_permissible | conformance | permissible | |
| 3 | tnm_pt | uc3_tnm_pt_min_req | completeness | minimal_req | [tnm_pt] = '' |
| 3 | tnm_pn | uc3_tnm_pn_type | conformance | datatype | isinteger[tnm_pn] |
| 3 | tnm_pn | uc3_tnm_pn_permissible | conformance | permissible | |
| 3 | tnm_pn | uc3_tnm_pn_min_req | completeness | minimal_req | [tnm_pn] = '' |
| 3 | tnm_pm | uc3_tnm_pm_type | conformance | datatype | isinteger[tnm_pm] |
| 3 | tnm_pm | uc3_tnm_pm_permissible | conformance | permissible | |
| 3 | tnm_pm | uc3_tnm_pm_min_req | completeness | minimal_req | [tnm_pm] = '' |

*Figure 8.* Data quality assessments generated from the data dictionary

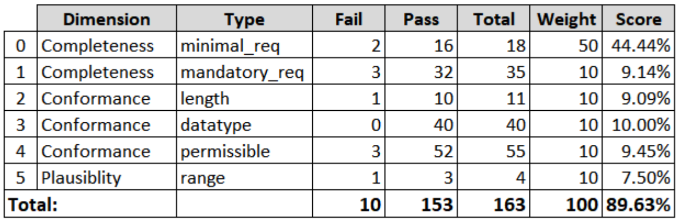


| check_type | dimension | description |
|---|---|---|
| minimal_req | completeness | Meets Minimum Requirements |
| mandatory_req | completeness | Meets Mandatory Requirements |
| length | conformance | Conforms to Length Restrictions |
| datatype | conformance | Confroms to Datatype Restriction |
| permissible | conformance | Conforms to List of Permissible Values |
| range | plausibility | Meets Known Range Limits |

Figure 9. Assessment organizing principles

### 2.2.3 Quality Scoring

The scores are based on the dimension and various assessment types. The final scores can be grouped by use cases, patients, or sites. Weights can be added at any point, as needed to increase, or decrease the importance of various rules. Below, in Figure 10 is an example of what a site's scoring for a particular use case may look like.



| | Dimension | Type | Fail | Pass | Total | Weight | Score |
|---|---|---|---|---|---|---|---|
| 0 | Completeness | minimal_req | 2 | 16 | 18 | 50 | 44.44% |
| 1 | Completeness | mandatory_req | 3 | 32 | 35 | 10 | 9.14% |
| 2 | Conformance | length | 1 | 10 | 11 | 10 | 9.09% |
| 3 | Conformance | datatype | 0 | 40 | 40 | 10 | 10.00% |
| 4 | Conformance | permissible | 3 | 52 | 55 | 10 | 9.45% |
| 5 | Plausiblity | range | 1 | 3 | 4 | 10 | 7.50% |
| **Total:** | | | **10** | **153** | **163** | **100** | **89.63%** |

Figure 10. Scoring Report

## 3 Quality Control of Image Data

Image quality assessment is conducted in two phases. During the annotation process there is a visual assessment by radiologists.  Post annotation, an automated process collects the data acquisition protocol for each image series and calculates a perceptual quality metric, PIQUE[2], on a representative sample of data from each image series.

## 3.1 Expert assessment during annotation

The first phase of quality assessment is performed by radiologists during the annotation process (see Deliverable D4.2). If images in a given study are of too poor quality to permit annotation, the radiologist will reject the entire study. These data will be removed from the annotation and data storage pipeline and the submitting data controller will be requested to provide a replacement study. This is the ONLY point at which studies will be rejected and replaced.

## 3.2 Automated assessment

Once image annotation is complete, the images and annotations are uploaded to the XNAT-based Euro-BioImaging repository. An automated process is executed by XNAT to perform image quality assessment which includes collection and documentation of the acquisition protocol for each study and the calculation of an image quality metric that is used to guide the AI team and other future users of EuCanImage data.

A containerized XNAT plug-in was built to provide means to access and index data (images) from XNAT and store results (acquisition protocol and quality metric) to be returned to XNAT for publication as added information attached to each imaging study.

### 3.2.1 Image Acquisition Protocols

EuCanImage data is collected retrospectively and represents standard of care imaging at each of the partner clinical sites. Thus, image acquisition protocols cannot be controlled as they might be in a clinical trial. It is well known that variation in modality (scanner type and manufacturer) and acquisition protocol introduce variance and potential bias into machine learning training and testing data. To permit an analysis of this bias, the protocol used to acquire each image series is automatically extracted from the data stored in XNAT and a protocol descriptor returned to XNAT to serve as a label for that series. Table 1 identifies the DICOM attributes that comprise the acquisition protocols for all EuCanImage use cases.

| Modality | DICOM Tag | DICOM Name |
|---|---|---|
| **ALL** | (0020,000D) | Study Instance UID |
| **ALL** | (0020,000E) | Series Instance UID |
| **ALL** | (0008,0070) | Manufacturer |
| **ALL** | (0008,0060) | Modality |
| **ALL** | (0008,0008) | Image Type |
| **ALL** | (0018,0050) | Slice Thickness |
| **ALL** | (0018,0088) | Spacing Between Slices |
| **ALL** | (0020,0032) | Image Position (Patient) |
| **ALL** | (0020,0037) | Image Orientation (Patient) |
| **ALL** | (0028,0030) | Pixel Spacing |
| **MR** | (0018,0010) | Contrast/Bolus Agent |
| **MR** | (0018,0020) | Scanning Sequence |
| **MR** | (0018,0021) | Sequence Variant |
| **MR** | (0018,0022) | Scan Options |
| **MR** | (0018,0023) | MR Acquisition Type |
| **MR** | (0018,0024) | Sequence Name |
| **MR** | (0018,0080) | Repetition Time |
| **MR** | (0018,0081) | Echo Time |
| **MR** | (0018,0087) | Magnetic Field Strength |
| **MR** | (0018,0091) | Echo Train Length |
| **MR** | (0018,1314) | Flip Angle |
| **MR** | (0018,9341) | Contrast/Bolus Usage Sequence |

| | | |
|---|---|---|
| **MR** | (0018,9344) | Contrast/Bolus Agent Phase |
| **CT / MG** | (0018,0060) | KVP |
| **CT / MG** | (0018,1190) | Focal Spots |
| **CT / MG** | (0018,1110) | Distance Source to Detector |
| **CT / MG** | (0018,1111) | Distance Source to Patient |
| **CT / MG** | (0018,1150) | Exposure Time |
| **CT / MG** | (0018,1151) | X-Ray Tube Current |
| **CT / MG** | (0018,1152) | Exposure |
| **CT / MG** | (0018,1160) | Filter Type |
| **CT** | (0018,1120) | Gantry/Detector Tilt |
| **CT** | (0018,1130) | Table Height |
| **CT** | (0018,1210) | Convolution Kernel |
| **CT** | (0018,9311) | Spiral Pitch Factor |
| **MG** | (0018,11A0) | Body Part Thickness |
| **MG** | (0018,11A2) | Compression Force |
| **MG** | (0018,5101) | View Position |
| **MG** | (0020,0062) | Image Laterality |

Table 1. Each DICOM image series can be characterized by a set of DICOM attributes that define the essential acquisition parameters. The acquisition protocol for each EuCanImage use case is constructed by selecting the appropriate attributes from this table for the imaging modality or modalities used in that use case.

### 3.2.2 PIQUE metric

Although it is common practice to use general quality metrics such at root means square error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), these metrics require the selection of a reference image. A reference free metric was chosen instead, to permit automation and simplify the quality assessment process. PIQUE (also referred to as PIQE) estimates block-wise distortion and measures the local variance of perceptibly distorted blocks to compute a quality score. PIQUE scores fall in the range [0,100]. The PIQUE score is inversely correlated to the perceptual quality of an image. A low score value indicates high perceptual quality and high score value indicates low perceptual quality.

PIQE is a MATLAB$^{TM}$ (Mathworks$^{TM}$ Natick, MA) function that we implemented in Python based on the algorithm proposed by Venkatanath ,et al.[2] Our version was validated against the MATLAB$^{TM}$ function using data from The Cancer Imaging Archive. This was done so our implementation could be released open source.

The PIQUE algorithm computes the Mean Subtracted Contrast Normalized (MSCN) coefficient for each pixel in the input image[2]. The image is then divided into nonoverlapping blocks and high spatially active blocks are identified based on the variance of the MSCN coefficients. In each active block distortion is evaluated using the MSCN coefficients and threshold criteria used to score the blocks as distorted with blocking artifacts, with Gaussian noise, or undistorted. The PIQUE score is computed as the mean of scores in the distorted blocks. Based on the literature a relative quality scale has been defined by Mathworks$^{TM}$ and illustrated in Table 2.

| Relative Quality | PIQUE Score Range |
|---|---|
| **Excellent** | [0, 20] |
| **Good** | [21, 35] |
| **Fair** | [36, 50] |
| **Poor** | [51, 80] |
| **Bad** | [81, 100] |

Table 2. Relative image quality based on PIQUE score

PIQE/PIQUE has been successfully used as a no-reference quality metric for a variety of medical imaging modalities and synthetic data[5, 6]. Figure 11 illustrates PIQUE scores for TCIA images in their original as acquired form and with varying degrees of added distortion to illustrate how PIQUE scores capture the distortions.

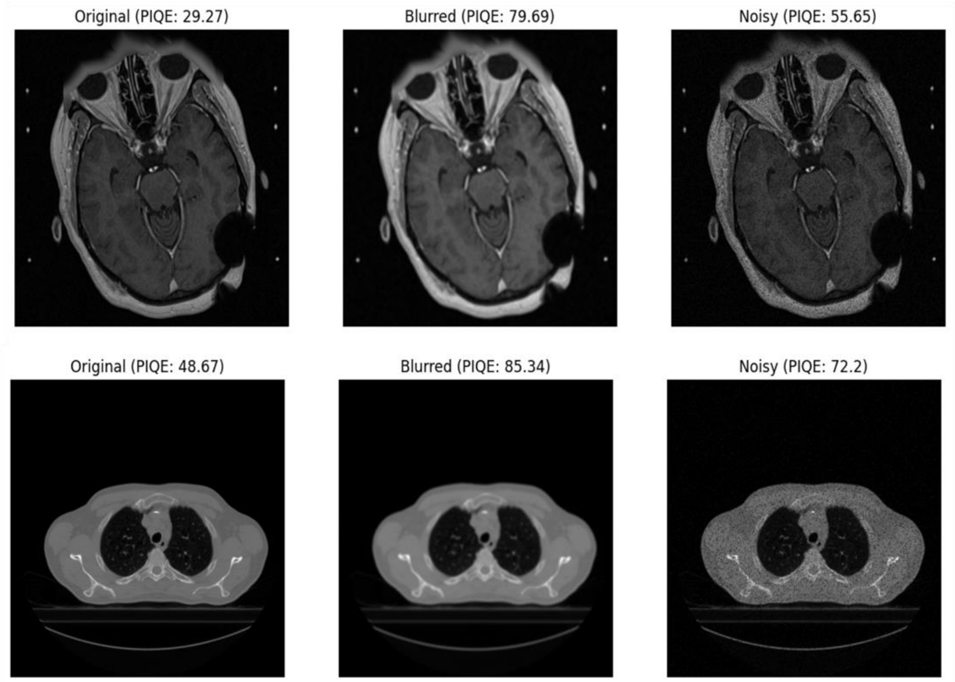


Figure 10. Examples of PIQE (PIQUE) socres for original TCIA images and distorted versions of the original.

## 4 Final Considerations

Data collection and QC is ongoing at all clinical sites. The tools described here continue to be enhanced as new situations arise. GUMED tools are being incorporated into POSDA as are the non-image quality assessment processes. We are considering adding the automated QC tools from POSDA in to the image quality assessment pipeline as future enhancement.

The Python implementation of PIQUE and the container used to deploy image quality assessments in the XNAT framework will be released open source on Github.

## 5 References

1. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, JG C. Research electronic data capture (REDCap)--ametadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377-81.
2. Venkatanath N, Praneeth D, Bh MC, Channappayya SS, Medasani SS, editors. Blind image quality evaluation using perception based features. 2015 twenty first national conference on communications (NCC); 2015: IEEE.
3. Talburt JR. Entity resolution and information quality: Elsevier; 2011.
4. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, Wu Y, Prosperi M, George Jr TJ, Harle CA. Assessing the practice of data quality evaluation in a national clinical data

research network through a systematic scoping review in the era of real-world data. Journal of the American Medical Informatics Association. 2020;27(12):1999-2010.

5. Higashiyama S, Katayama Y, Yoshida A, Inoue N, Yamanaga T, Kawahata H, Ichida T, Miki Y, Kawabe J. Usefulness of a No-Reference Metric for Evaluation of Images in Nuclear Medicine-A Comparative Study with Visual Assessment2021.
6. Karthik K, Kamath S. Deep neural models for automated multi-task diagnostic scan management—quality enhancement, view classification and report generation. Biomedical Physics & Engineering Express. 2021;8(1):015011.