Journal of Digital Imaging
https://doi.org/10.1007/s10278-021-00476-9

ORIGINAL PAPER

# Research Goal-Driven Data Model and Harmonization for De-Identifying Patient Data in Radiomics

**Surajit Kundu[1] · Santam Chakraborty[2] · Jayanta Mukhopadhyay[1] · Syamantak Das[2] · Sanjoy Chatterjee[2] · Rimpa Basu Achari[2] · Indranil Mallick[2] · Partha Pratim Das[1] · Moses Arunsingh[2] · Tapesh Bhattacharyya[2] · Soumendranath Ray[2]**

Received: 24 July 2020 / Revised: 22 May 2021 / Accepted: 9 June 2021


## Abstract

There are various efforts in de-identifying patient's radiation oncology data for their uses in the advancement of research in medicine. Though the task of de-identification needs to be defined in the context of research goals and objectives, existing systems lack the flexibility of modeling data and normalization of names of attributes for accomplishing them. In this work, we describe a de-identification process of radiation and clinical oncology data, which is guided by a data model and a schema of dynamically capturing domain ontology and normalization of terminologies, defined in tune with the research goals in this area. The radiological images are obtained in DICOM format. It consists of diagnostic, radiation therapy (RT) treatment planning, RT verification, and RT response images. During the DICOM de-identification, a few crucial pieces of information are taken about the dataset. The proposed model is generic in organizing information modeling in sync with the de-identification of a patient's clinical information. The treatment and clinical data are provided in the comma-separated values (CSV) format, which follows a predefined data structure. The de-identified data is harmonized throughout the entire process. We have presented four specific case studies on four different types of cancers, namely glioblastoma multiforme, head–neck, breast, and lung. We also present experimental validation on a few patients' data in these four areas. A few aspects are taken care of during de-identification, such as preservation of longitudinal date changes (LDC), incremental de-identification, referential data integrity between the clinical and image data, de-identified data harmonization, and transformation of the data to an underlined database schema.

**Keywords** Radiomics · Radiology · Protected Health Information (PHI) · Patient Health Record (PHR) · De-identification · DICOM · Longitudinal Date Changes (LDC) · Glioblastoma · Head–neck cancer · Normalization · Harmonization

✉ Surajit Kundu
surajit.113125@gmail.com

Santam Chakraborty
drsantam@gmail.com

Jayanta Mukhopadhyay
jay@cse.iitkgp.ac.in

Syamantak Das
syamdas40@gmail.com

Sanjoy Chatterjee
chatterjee72@hotmail.com

Rimpa Basu Achari
rimpaachari@gmail.com

Indranil Mallick
imallick@gmail.com

Partha Pratim Das
ppd@cse.iitkgp.ac.in

Moses Arunsingh
85moses@gmail.com

Tapesh Bhattacharyya
tapesh27@gmail.com

Soumendranath Ray
soumen.ray@tmckolkata.com

[1] Indian Institute of Technology Kharagpur, Kharagpur, India

[2] Department of Radiation Oncology, Tata Medical Center, 14 MAR (E-W), Kolkata 700160, India

Published online: 09 July 2021

## Introduction

With the advancement of data-driven machine learning techniques, there is increasing emphasis on the banking of medical data. Digital Imaging and Communications in Medicine (DICOM) is a worldwide standard to facilitate radiological data exchange. It defines a digital imaging format, file structure, and image interchange protocol [1]. However, most patients' clinical information is stored in medical records captured in free text format. The integration of this data with radiological image data is essential for medical grade decision making using radiomics. However, appropriate de-identification of this data is necessary for the interests of patient privacy and medical ethics. In practice, radiological images and clinical data are both acquired separately during the treatment of a patient. The DICOM standards facilitate transmission, storage, retrieval, processing, and display of radiological imaging data. Such data is commonly stored in a Picture Archiving and Communication System (PACS). Furthermore, the patient health records (PHR) are managed by the Clinical Data Management System (CDMS) or Hospital Information Systems (HIS), which lack standards like DICOM. Therefore, there is no direct linkage between the PHR and radiological data.

At present, there are several robust systems, which enable complete de-identification of the DICOM datasets when they are stored in image data banks [2, 3]. Additionally, another key requirement for the research data bank is the de-identification of PHR while maintaining an associative relationship to DICOM data. A good example of such data is The Cancer Imaging Archive (TCIA) [4], where PHR is stored separately from the imaging data in spreadsheets.

Another example is the Oncospace platform [5] that has defined unified radiation oncology (RO) database schema designed for facilitating research in personalized medicine. The schema integrates treatment planning and dose data with PHR like diagnosis, pathology, treatment, and disease outcomes.

In the present system, we describe a platform for contextual modeling of data obtained from patient health records under a flexible entity relationship modeling in an image data bank (CHAVI—CompreHensive ArchiVe of Imaging in Oncology). The de-identification schema also allows data normalization in a generic form and harmonization across diverse research projects. The de-identification system retains the entity-relationship between clinical and imaging data. The de-identified dataset can facilitate complex queries' design and aid data retrieval processes in the image databank.

## Motivation

A publicly accessible image data bank for research purposes should host valid data in a consistent data structure. However, PHR in clinical practice is often recorded variably and hence needs to be structured before it can be used in a data bank. Additionally, a lexicon of data elements needs to be maintained such that consistency can be ensured. Hence, a data structure specification is required to ensure that high quality PHR can be stored and associated consistently with imaging data.

### Gap in the Area

There are many existing works on radiological image dataset de-identification, which include PHI anonymization, burned-pixel data removal, face de-identification, etc. However, most of these systems do not allow the de-identification of the clinical data and patient health records while maintaining association with de-identified DICOM data. A desirable system should be able to de-identify ongoing patient health record data while retaining the temporal association with previously de-identified data.

### Requirements

Data harmonization is an iterative process of capturing, interpreting, examining, and reconciling organization information requirements and data standardization as the mapping of the simplified data [6].

In this paper, we report the development of a standalone system, which provides the facility to de-identify both DICOM and PHR data. We propose a generalized data model for accumulating the comprehensive de-identified clinical and imaging data. DICOM imaging data can include diagnostic images (used for initial diagnosis and response assessment) and radiation therapy-related objects like RT structure sets, RT plans, RT dose, etc. PHR obtained from clinical records have patients' demographic data, disease related data (e.g., history, examination, pathology, stage), treatment, and outcomes data. As this data is available as unstructured text, a data harmonization method needs to be used before the data can be archived. This requires that a lexicon to be developed. Normalised and de-identified PHR data is temporally associated with de-identified image data and subsequently stored in the database. The entire pipeline describes the methodologies following, de-identification, data cleansing, harmonization, and ER data model for mapping de-identified data with the databank. Throughout the de-identification process, the system captures the entity relationship (ER) model of the databank and puts the de-identified data in insertion queries of the database by mentioning specific tables, attribute names, and associated values.

## Methods and Materials

The radiological images are acquired in DICOM format from the Treatment Planning System (TPS) and the PACS. The clinical data of the patient is taken in the form of comma-separated values (CSV), which follows a predefined data structure.

For demonstrating we present four typical case studies on four different disease site, following Glioblastoma multiforme, Intensifying Radiation Treatment in Advanced/Poor Prognosis Laryngeal Hypopharyngeal (LH) and Oropharyngeal Cancers (OPC) using PET-CT Based Dose Escalation Strategies (INTELHOPE), randomised phase II of immunotherapy with pembrolizumab for the prevention of lung cancer (IMPRINT), and hypofractionated radiation therapy (HYPORT-B). The same model can be applied to other studies also. INTELHOPE, IMPRINT, and HYPORT-B are short names of these studies undertaken at Tata Medical Center, Kolkata.

## Definition of Research Project

The starting point is the definition of a research project. It may be specific for a particular cancer (e.g., head–neck cancer) or include multiple cancer types. The system allows a user with appropriate privileges to create a project definition in CHAVI, which has information on the research project. Each project is assigned a unique ID that is a secondary reference for all datasets stored in the CHAVI database. At present, the system requires that all research projects included in the system should have an Institutional Review Board approval or waiver. Clinical data of patients are stored in a research database or a clinical trial management system. The data is exported from the source and then de-identified, as shown in graphical user interface (GUI) in Fig. 1. Although data from other databases can also be used, the current system has been tested exclusively on data imported from REDCap databases [7, 8]. The clinical data may be recorded prospectively or retrospectively in REDCap. A longitudinal database with repeating instruments may also be used in specific projects depending upon the research question. The current system is flexible enough to allow data from all of these different types of databases.

## Project-Driven De-Identification Schema

For each project, an external template is provided to acquire the data attribute and types. This template is currently created in a spreadsheet. It contains the attributes list of both ER model and the source data, as shown in Fig. 2. The attribute properties provide knowledge regarding the dataset and match the data elements to be exported from the medical database. A few examples are given below.

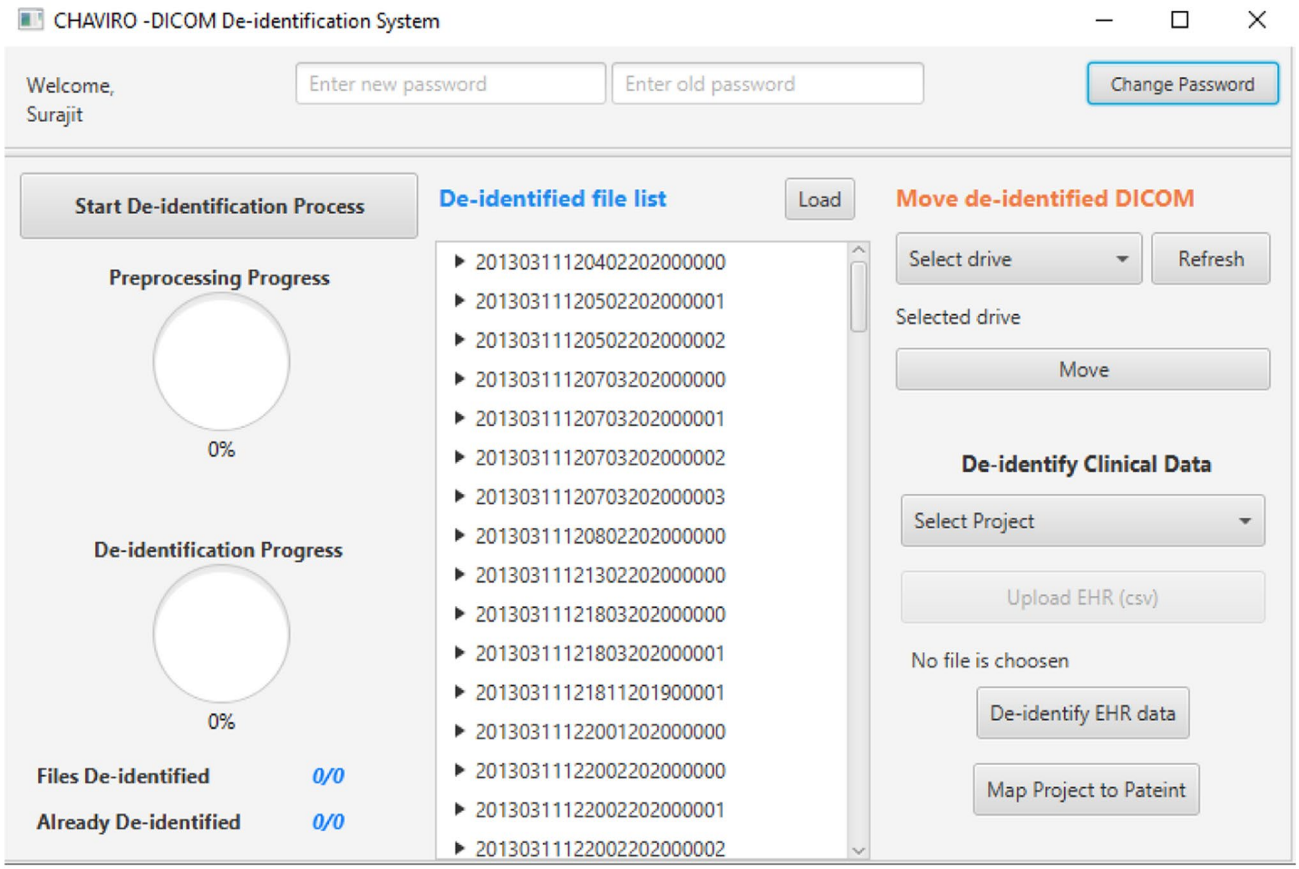


**Fig. 1** Graphical User Interface of the de-identification system

- IsIncludeInSchema (Y/N)—It specifies whether the given clinical data is to be accommodated in the schema.
- DefaultValues—The same information may be implicitly applied to all cases in certain projects but is not specified in the database. For example, if this is a brain tumor project, all patients will have the anatomic site as CNS. In this case, even if the exported data does not contain the information, the system will have the information.
- DataFieldType—It allows three types of entities to be selected following attribute (A), value (V), and multiple references (M). An attribute will have a value, which is mapped to the corresponding field in the ER model. The value itself is a piece of information, which may need de-identification. It will be associated with specific attributes of the existing ER model. Such designs are modeled to store the extensive list of the gene, protein, and other molecules. There is a particular attribute called Multiple references attribute. It encodes two pieces of information, following one corresponding attribute value, and another is attribute reference. An example is Telomerase Reverse Transcriptase (TERT) is a code name of the gene. Alongside this, it keeps the references of another attribute named mutation result.

It may occur that a part of clinical data cannot be accommodated in the existing data model. The schema contains a separate table consisting of three attributes: medical record name, value, and instance time of the data acquisition. The inharmonious data is stored in this table. The user renders this information once for a single project. It helps in the information mapping of source data with the de-identification schema, which is consistent with the ER modeling of image data bank. A patient may belong to one or more projects. In our case study, the clinical data is exported in a CSV file. Each file consists of patient-level clinical data of multiple subjects who belong to the same research protocol. However, in practice, CSV files can be generated for other databases also. The data definition in the CSV file is specified at the time project is created. Thus, updated data of multiple patients can be imported at various time points as long as the data definition is maintained. Once the de-identification is completed, the system produces a JSON file. It contains de-identified clinical data and the references of the project with each patient.

These projects deal with different cancer sites having variable endpoints and objectives. We have evaluated the process with respect to the following during this de-identification process.

- De-identification of data harmonized with entity relationship model of databank.
- Clinical data harmonization prior to storage.
- Incremental de-identification of data where longitudinal information is recorded.

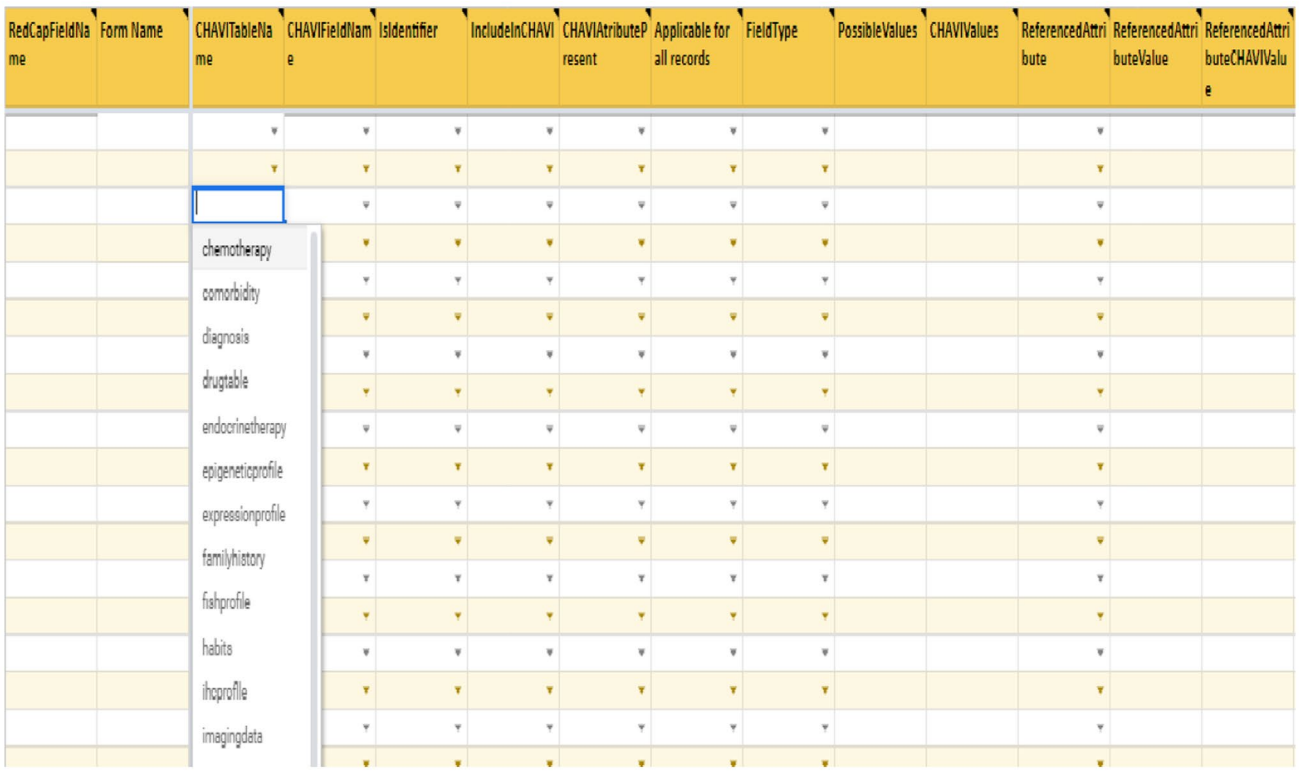


**Fig. 2** The information collecting template for data mapping

– Ensuring that temporal relationships between clinical records and DICOM data are maintained.
– Referential data integrity between clinical records and imaging data, while de-identifying multiple patients' data at a time.

## Entity-Relationship Model of the Databank

There is an underlying ER model for archiving the data after the de-identification. The system performs de-identification, harmonization, and re-transformation on both DICOM and clinical patient data. The de-identification process is defined with a project-specific data model and follows a schema for storing the clinical data. This process helps to shape the data in a uniform structure. The harmonization process helps to achieve consistent and quality data. This uniformity is maintained in the databank as well. Hence, the databank recognizes the correct de-identified data and then allows it to be uploaded in the system. Any raw data or unstructured data is rejected during the uploading process in the databank. This ensures data security, consistency, and quality. Radiological imaging data are to be uploaded incrementally or in a ZIP archive.

Clinical information is transmuted in a specific morphological order. We represented an object in the structure of JSON (JavaScript Object Notation), which is a set of unordered name/value pairs. The object is confined with curly braces. It starts with the left brace and ends with the right brace. Each name/value pairs are separated by a comma (,). A single object may contain one or more name/value pairs. In this system, the clinical data are encapsulated with the entity set, attribute, and the corresponding de-identified record. And the values are parsed from the corresponding key names. A single JSON consists of clinical data, which belongs to a particular disease site. An overview of the entire process is shown in Fig. 3. During the data transfer to the databank, the system recognizes a project specific data model and parses the data object wise. Then it performs a one-to-one data mapping in the database of the medical image databank.

## De-Identification of Radiological Images and PHI

A study on the de-identification process of both RT DICOM data and basic clinical data is published earlier [9]. The proposed de-identification system categorizes protected health information (PHI) into two classes: direct and derived identifiers. The direct identifiers are name, address, contact information, etc. These are the most serviceable and fundamental information to identify a patient. The derived identifiers (DI) are deduced from supplementary data. While these may not

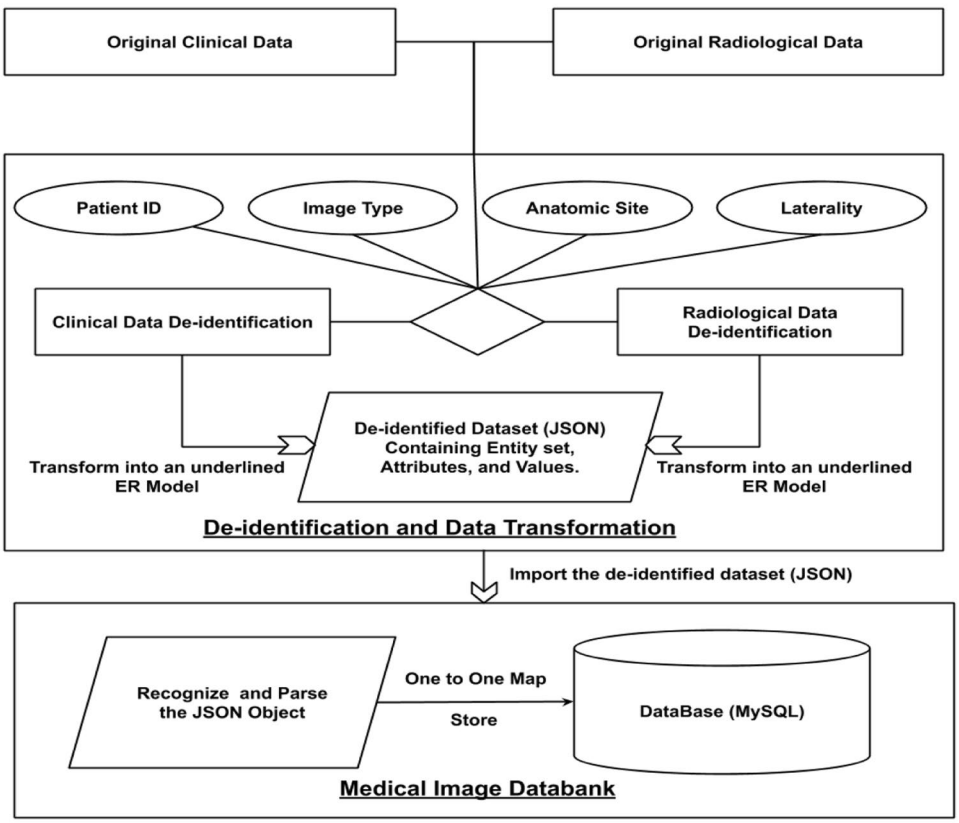


**Fig. 3** A system flowchart

hold identifiable patient information, identifying information may be deduced from this. The DIs are date, DateTime (date and time together), unique identifiers (UIDs), date of birth (DOB), etc.

The # symbol replaces all direct identifiers in patient information. DIs such as the date and DateTime values are modified such that longitudinal temporal date relationships are preserved. Patient UIDs are replaced by the system generated unique ID. The system uses a relational database management system (RDBMS) MySQL to store the original unique identifier (UID) references in a local database in an encrypted format.

## Schema of Clinical Data

A generalized schema is designed for archiving the clinical data. It maintains the hierarchy of different entities in a relational database, as shown in Fig. 4. This information hierarchy is adopted to match the way clinical information is usually collected and related to the patient. The schema is flexible enough for storing and retrieving the data based on complex queries. It is a 1:N parent–child relationship. Each record of the patient dataset is stored against a unique identifier. Concurrently, the references of the associated information are maintained. Thus, it ensures consistent data storage. In order to envisage the potential of the data model, we describe each entity and its corresponding medical data elements. The patient information like personal history, menstrual history, habits, survival is organized in a 1 to n relationship with the patient information table. Similarly, the diagnosis table has one to many relationships with patient information as a disease is a unique event for a patient. Treatment delivered for the disease and disease-related tests is linked to the diagnosis in turn by a 1 to n relationship. Hence, the treatment, treatment results, and disease investigations results can be extracted from the database.

**Project Patient Map** The "ProjectPatientMap" table holds a mapping between patient ID and project ID. The table is linked to the Project Information (PI) table through the patient ID key references. This table also contains project ID and patient entry date, which is the date on which the patient was registered at the treating center.

**Patient Information** The "Patient Information" (PI) contains a set of data elements related to patient profile based details. This table stores age, gender, center code in which the patient originates, de-identified date of registration at the center, patient date of birth, and Performance Status of the patient. These data elements at this hierarchical level relate directly to the patient and do not vary across diagnoses. Most of these elements are captured in history during a patient encounter.

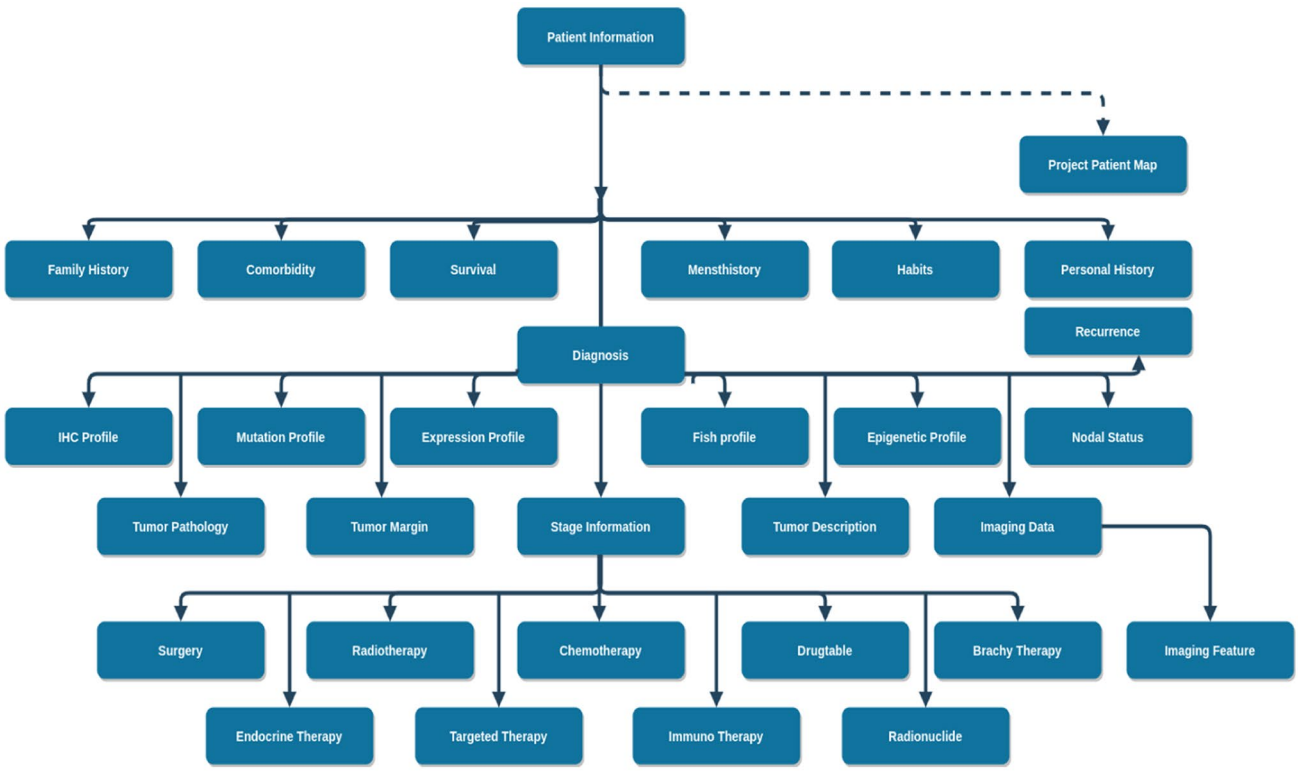


**Fig. 4** Hierarchical database model of clinical data

**Family History** This part of the schema manages the information regarding the family history of malignancies. The family history dataset has one-to-one relation to PI. In addition, it contains patient-specific detail such as the relationship of the affected family member with malignancy, age of the family member at the diagnosis of his/her cancer, the patient's survival status (alive or dead), and the type of cancer the family member had. Typically, this would include the site as well as the pathology if available.

**Comorbidity** The comorbidity grading and classification are variable. The relational table can accommodate multiple comorbidities to be coded separately. The attribute set is as follows: the date on which comorbidity was diagnosed, the duration for which comorbidity was present, the severity of the comorbidity, comorbidity type status, and name of the comorbidity.

**Mensthistory** The menstrual history attribute sets are following, age at menarche, attained menopause status, age at which patient become postmenopausal, duration of each menstrual cycle in days, duration of time each menstrual period lasts in days, menstrual cycle, date of last menstrual period, pain with menstrual periods records.

**Habits** The substance abuse-related habits the patient may have or had in the past. The system de-identifies and acquires the relevant information like substance type, substance quantity, substance unit, substance duration in the year, last use of the substance, and current usage of the substance.

**Personal History** This part of the schema stores information on the patient's personal history, such as marital status and immunization status.

**Diagnosis** Diagnosis is a key element that comprises disease related information, e.g., pathology, stage, etc. There is a 1 to n relationship between patient information and diagnosis. Thus, a single patient can have multiple disease diagnoses, but each diagnosis can only belong to a single patient. Thus a patient with bilateral breast cancer will have information obtained for left and right breast cancers captured as a separate diagnosis. It allows differential treatments for the two diagnoses. The diagnosis contains the following component of medical records, recurrence—Indicate if the diagnosis is recurrent cancer, laterality of the disease, anatomic sites, subsite of the disease, pathology type of the disease, and the date when the pathology was obtained first.

**Stage Information** The staging of the patient is usually done after the diagnosis. Every individual stage information keeps the corresponding diagnosis references after the de-identification. The system allows the user to choose the staging system type and stores record on T, N, and M stages if the AJCC staging system is being followed.

**Survival** The patients' survival status is included in the clinical information. It tracks the current state (dead or alive) of survival. Alongside this, queries can be executed to obtain information on when the patient died after the diagnosis of a disease. The survival object incorporates the following information in the de-identified JSON, status of the patient of the last follow-up, date of death of a patient, overall survival from the patient's admission date, and the date of the last follow-up. This date can be increased as the length of follow-up increases for a patient.

**Tumor Pathology** The system enables the de-identification of different aspects of the tumor pathology such as histological type, grade of the tumor, necrosis status, existence of angioinvasion, lymphatic invasion status, appearance of perineural invasion, tumor deposits fact, and treatment effect status.

**Tumor Description** The gross description of the tumor is noted in the surgical pathology. The tumor description consists of the maximum size of the tumor in millimeters (mm), tumor site, focality (unifocal, multifocal, or multicentric), description of the tumor extent, and tumor perforation of the viscus.

**Tumor Margin** Margin status often holds prognostic significance when tumors are resected. Therefore the system allows the granular description of the margin status, which includes the following properties margin name, involvement, and distance in mm. The database is also capable of storing multiple margins for the same tumor pathology specimen.

**Nodal Status** The nodal status is typically obtained from information after a nodal dissection or sampling procedure. The attributes set are pathology specimen, node level, number of positive nodes, number of nodes with isolated tumor cells found in pathology, number of nodes with micrometastases, extranodal extension status.

**Recurrence** This table holds the details of all recurrences and tumor response data for the patient's disease. Note that the recurrence table is kept necessarily separate from the survival table as a disease can have one or many recurrences without affecting the patient's survival. The attributes that are available include the recurrence type (recurrent disease, stable disease, or progressive disease), recurrence location, recurrence date, duration of time which the patient was recurrence-free, response assessment, response location, response to treatment, time at which intervention is assessed in days, and date on which response is assessed.

**Genetic Information** The genetic makeup of the tumor is acquired in harmonized form. The system maintains a catalog of the gene, protein, and other molecules. The new substances can be added through the user interface of the system. The results of the genomics tests performed can be recorded for an arbitrary number of genes and proteins. The gene from the clinical data is mapped to the existing catalog. All relevant objects are individually linked to the diagnosis dataset, allowing different genes and protein abnormalities to be recorded. The patient data related to MutationProfile, FishProfile, ExpressionProfile, IHCProfile, and EpigeneticProfile are stored after the de-identification.

**Imaging Data and Features** It is mostly semantic information, as the quantitative imaging data will be extracted for each imaging dataset. The image data keeps its type and acquisition date. The attribute set of "ImageFeature" is the timing of the imaging in relation to the disease, site abnormality, contrast enhancement presence, status of edema in the image, disease status during the time of imagery, and the type of disease progression noted on imaging.

**Treatment information** The treatments are performed for a particular diagnosis, the associated information transitively dependent on that diagnosis. Alongside this, it keeps the references of the corresponding stage information so that the treatment delivered for different stages of the same diagnosis may be captured. The treatment information of the patients is stored in several objects of the data model, which is listed below.

- Surgery—Consists of surgery date, surgery details, sentinel node biopsy status, side surgery status, nodal dissection status, and reconstruction type.
- Radiotherapy—Data fields are radiotherapy status, radiotherapy intent, radiation dose in cGy, number of fractions, radiotherapy start date, radiotherapy complete date, radiated volume, side of radiation, concurrent chemotherapy status, and radiotherapy setting it which it is delivered.
- Chemotherapy—Contains chemotherapy status, intent of chemotherapy treatment, chemotherapy setting, regimen, number of chemotherapy cycles are received, chemotherapy start date, and complete date.
- Brachy therapy—List of attributes are following brachytherapy status, brachytherapy setting, total brachytherapy dose in cGy unit, number of fractions, radiated volumes, radiated side, re-irradiation status, name of equipment, brachytherapy applicator, brachytherapy dose point, brachytherapy dose volume, brachytherapy start date, and complete date.
- Endocrine therapy—Consists of endocrinetherapy status, treatment intent, setting, regimen, endocrine therapy regimen start date, and end date.
- Targeted therapy—Data fields are current status, targeted therapy treatment intent, setting, regimen, number of targeted therapy cycles, start date, and complete date.
- Immuno therapy—Contains immunotherapy status, immunotherapy treatment intent, setting, regimen, number of immunotherapy cycles, immunotherapy regimen start date, and complete date.
- Radionuclide therapy—List of attributes are following status, radionuclide isotope therapy type, radionuclide dose in Bequrel, number of radionuclide therapy, start date, and end date.

Additionally, a "drugtable" is provided so that granular information on each drug delivered to the patient can be accommodated as required.

## Clinical Data Harmonization

Every medical institute collects the clinical data in a specific format through a CDMS. Therefore, there is always a hazard that each medical specialist may interpret each data element differently. For example, a glioblastoma multiforme patient's anatomic site could be written as the central nervous system, CNS, or brain. Similarly, a study date can be obtained in a different date format. Likewise, the attribute name can vary in several CDMS, such as the side of the tumor can be discerned through laterality, disease side, or other phrases. Therefore, the attribute list is mapped to the data model for every individual project data to overcome these issues.

At the first step toward the clinical data de-identification, the system creates a hash-map containing data fields of the actual clinical data and encapsulated attribute and entity set of the underline database schema, as shown in Fig. 6. The $a_1, a_2, a_3 ... a_n$ is the list of data elements in exported original data, $L_1, L_2, L_3 ... L_n$ is the composition of attributes and entity set. For every project, there is one-to-one attribute mapping between the source and the de-identified data field. For example, in Fig. 5, $a_1$ is mapped to $A_1$ ($a_1 \rightarrow A_1$). It creates a knowledge base to discern each record. The attribute list is mapped in a relational table. This is exerted for moving inaccurate, broken, and erroneous data from the original treatment dataset. The project configuration file comprises a set of structural JSON objects containing the expected medical record and the harmonized value. Once an attribute matches, the system parses the corresponding clinical data and converts it to the normalized form. There are general use cases like the gender of a patient who is male, female, or transgender. However, this same data can be collected inclusive of {male, female, transgender}, {M, F, T}, {0, 1, 2}, and many other forms. This type of data is contradictory to the database. In this circumstance, this file is utilized for transforming multi-source data into one cohesive data set.

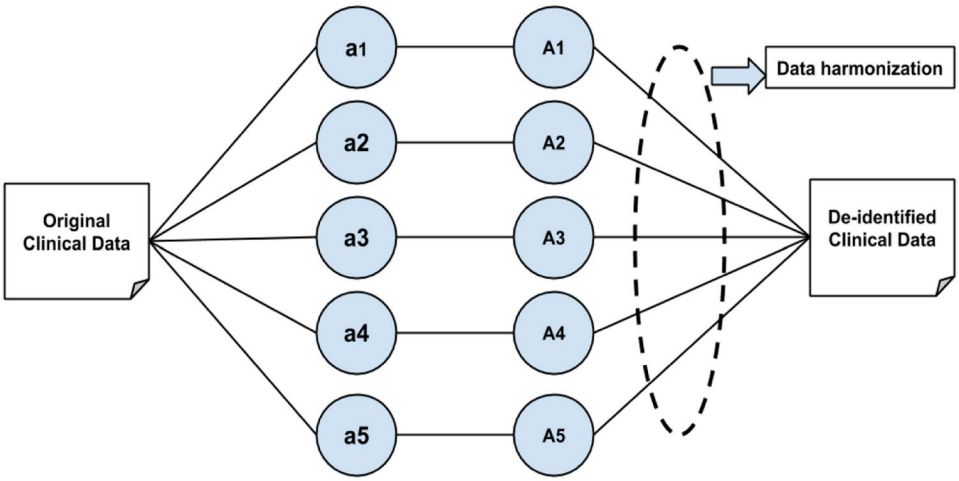


**Fig. 5** Overview of data harmonization

A configuration file is perpetuated for regulating data harmonization and validation. It includes project information, attribute set of medical records, actual data, and harmonized values. A typical example of the configuration file with dummy data shows in Listing 1. It has two objects along with the project id. The first one ("default") is a JSON array, contains the default value, such as the anatomic site for a specific project. This array may have N number of JSON objects. An inner object "system_attribute_name" —attribute name in the data model that keeps the default value. The object holds an array having two indexes. The first index contains the default value, and the second is the table name where the value will be stored. The "valuesmap" object consists of the necessary attribute list, which corresponding value needs to be harmonized. The "input_data_attribute" is the data field name in the actual clinical data. It holds another inner JSON object. "system_attribute" is the key of the object, which contains two arrays of the same length. One array contains the expected value from the input data, and another holds actual values that are acquired in the de-identified file. It is mapped one to one like $array1[0] \xrightarrow{\text{Mapped}} array2[0]$.

```
{
    "projectid":"Project ID",
    "default":[
        {"system attribute name":["default value","system table name"]},
        {"system attribute name1":["default value1","system table name"]}
    ],
    "valuesmap":{
        "input data attribute":{
            "system attribute":[
                ["Expected value from original data","2","3"],
                ["Actual data stored in system","Right","Bilateral"]
            ]
        }
    }
}
```

The same approach is followed for retaining longitudinal temporal information as reported in a previous publication [9]. The proposed de-identification system preserves the longitudinal date changes (LDC) in both projects-driven clinical and image data. The same is applied for incremental de-identification as well.

## Clinical Data De-Identification

The complete process of the clinical data de-identification is shown in Fig. 6. At first, a CSV file containing the clinical data from a single project is imported to the de-identification system. Next, the user has to select the associated project name from the list. Once the de-identification starts, the system creates a hash map of the original and de-identified

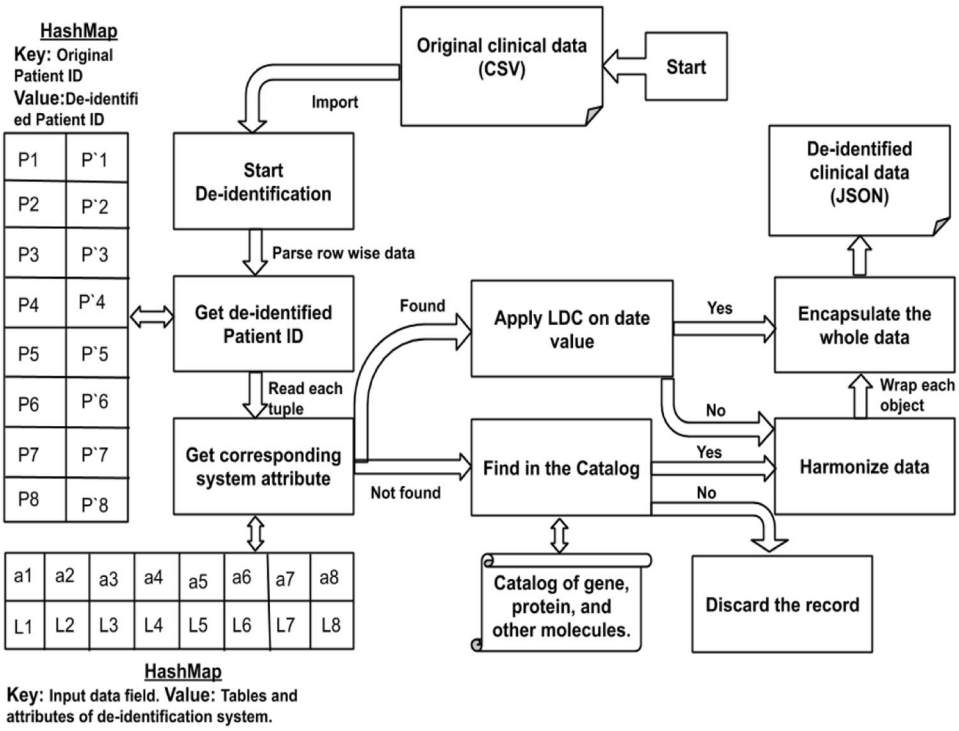


**Fig. 6** Workflow of clinical data de-identification

patient ID. Then the system goes through each row of the CSV against its actual medical record number. A hash table of the attribute set is created with the entity set of the generalized data model. During the de-identification, the system parses the value of each mapped attribute and applies LDC on all dates. If a corresponding medical record name can not be found in the hash list, the system searches it from the catalog. If the record exists in the catalog, It is discerned as gene, protein, or other molecules. The nonexistence records are excluded from the de-identified dataset. Concurrently, a log file will be generated, including the list of all discarded data fields for emending. Once the data parsing is complete, the system applies the harmonization process to every single clinical data. Then all distinct tuples are merged in different relative objects. And all the objects are encapsulated in a single JSON. It contains de-identified clinical data with the references of the project.

## Case Studies

We have considered at least 30 patients' data from every project. As shown in Table 1, a case study has been performed on a total of 150 patients' complete treatment data. It consists of radiological images (diagnostic, treatment therapy plan, treatment verification images) and clinical data. The image data is compressed in the ZIP format and imported in the de-identification system. Total 543 radiological studies are de-identified, which includes different modalities such as CT, PET, CBCT, MRI, RTSTRUCT, RTPLAN, and RTDOSE.

### Glioblastoma Multiforme Treatment Data Acquisition

A total of 30 patients with glioblastoma multiforme data are taken. The process starts with radiological images, followed by the clinical data of the patient. The DICOM images, including diagnostic images, RT treatment planning annotated DICOM, plan parameters, RT verification images such as CBCT, and RT response images, are exported from TPS. A CSV file containing extracted clinical data for each patient is obtained from RedCap. Each record contains the actual hospital UID corresponding to the clinical records of that patient. The treatment information comprises patient demographic details, imaging data and features, gene mutation, patient followup detail, surgery, tumor pathology, chemotherapy, radiation, and re-irradiation data. The demographic details are incorporated with the PI while replacing the identifiers like the patient's hospital UID with de-identified UID throughout the de-identification process. Pathology information is accommodated in diagnosis, tumor pathology, tumor margin, and tumor description. Imaging correlated data are mapped to imaging data and imaging feature modules. The followup records of the patient are amassed in survival and recurrence. Chemotherapy and radiation data can directly be shifted to the chemotherapy and radiotherapy correspondingly in the data model. Re-irradiation details are collected

**Table 1** The glossary of the dataset

| Total Patient = 150 \|\| Total study = 543 | | | | |
|---|---|---|---|---|
| Site | Glioblastoma = 30 | IntelHope = 30 | HYPORT-B = 30 | Lung = 60 |
| Radiological study | 99 | 82 | 62 | 300 |
| Diagnostic Images | MRI = 15 | PET = 30 | PET = 22 | NA |
| | | | CT = 17 | |
| RT Therapy planning | CT = 30 | CT = 30 | PET = 23 | CT = 60 |
| RT Therapy verification | CBCT = 22 | CBCT = 30 | | CBCT = 240 |
| RT Therapy Response | MRI = 24 | NA | | NA |
| Clinical data | Patient profile information (age, gender, registration date, performance status), diagnosis, stages, recurrence, imaging features, expression profile, treatment details, and survival | Patient profile information (age, gender, registration date, performance status), diagnosis, stages, recurrence, imaging features, expression profile, comorbidities, treatment details, and smoking/drinking habits | Patient profile information smoking-drinking habits, diagnosis, stages, imaging features, stage information, comorbidities, survival, recurrence, and tumor pathology —grade | Patient profile information (age, gender, registration date, performance status), diagnosis, stages, recurrence, imaging features, and survival |

in the radiotherapy module as a new instance. There are some semantic descriptions of images like edema volume, t2 enhancement, sub-ventricular zone (SVZ) involvement, etc. These entities are not defined distinctly in the existing data model. In this situation, the data are collected in the form of name and value pairs. Then it is kept in a relational database table of the current schema.

The values against clinical data elements are harmonized. The anatomic site is CNS across all the patients of this project. Likewise, the diagnosis site is the brain, and the radiotherapy setting is adjuvant. Pathology information in the source data is following 1, 2, 3, 4, 5, and 6. After the harmonization, the data is replaced with the consecutive denomination "Anaplastic astrocytoma, IDH-mutant", "Anaplastic astrocytoma, IDH-wildtype", "Anaplastic astrocytoma, NOS", "Glioblastoma, IDH-wildtype", "Glioblastoma, IDH-mutant", "Glioblastoma, NOS". As an example, the normalized values of image timing are "Baseline", "Post primary surgery", "Pre-chemotherapy", "Interim during adjuvant chemotherapy", "End of treatment", "Followup", and "Post re-do surgery".

After the completion of de-identification, the de-identified images are stored in a folder tagged with patient ID and study instance UID. Then the patient's clinical data is de-identified and encapsulated in a JSON. It contains a de-identified patient id along with the reformed clinical data and the corresponding project identity. The system de-identifies only delineated data fields, and the remaining attributes list appears in a warning dialog box. The LDC is preserved on both de-identified images and patient health data. In some cases, patients underwent multiple treatments due to a second primary or recurrence of the tumor. In such circumstances, collective repeat instances occur for the recapitulated diagnosis, stage information, and treatment details. The clinical data of a patient is shown in Table 7. The data field of the input data source and attributes of the data model represents a one to one medical record mapping. The "Is Modified" column shows the participation of a single record in the de-identification process. It has three indications following Yes, No, Not Included. If the system applies harmonization, LDC, or system-generated value replacement on the original clinical data, it shows status "Yes". "No" indication means the original record is acquired without any modification. There is a certainty that all attributes of the input data source may not be mapped in the system. The unmapped data does not participate in the de-identification process. It is specified "Not Included" in the column of the table.

It may also occur that the schema has $N$ number of attributes in the corresponding table. But data sources provide K ($K \subseteq N$) number of records. So in this situation, the system keeps $K$ number of records while $(N - K)$ remains NULL.

As shown in Table 2, the statistics for the glioblastoma multiforme data de-identification are displayed in tabular

**Table 2** The statistics for the de-identification status of Glioblastoma multiforme data

| | Records de-identified | Records not de-identified |
|---|---|---|
| Data fields need to be de-identified | 35 | 0 |
| Data fields not to be de-identified | 0 | 14 |

**Table 3** The statistics for the de-identification status of INTELHOPE data

| Total = 39 | Records de-identified | Records not de-identified |
|---|---|---|
| Data fields need to be de-identified | 28 | 0 |
| Data fields not to be de-identified | 0 | 11 |

form. A total of 35 medical records should be de-identified, which is successfully de-identified. Similarly, 14 data fields are not de-identified, which does not need de-identification.

## INTELHOPE Data Acquisition

INTELHOPE is a study on head and neck cancer patients. We have taken ten patients' data for this project. A total number of 90 studies are de-identified where the anatomic site is chosen as head–neck, followed by the laterality of the disease. The radiological studies comprise ten diagnostic PET, 30 RT planning CT, and 22 treatment verification CBCT. The patients are planned for radiation therapy in either TomoTherapy®or Eclipse treatment planning system (Varian Medical Systems, Palo Alto, USA) [10]. First, the RT structure set, RT plan, and RT dose files are exported from the TPS. Then the RT treatment planning images are de-identified.

The clinical data are extracted from the CDMS following, smoking and drinking habits details, comorbidities (hypertension, diabetes, ischemic heart disease, chronic renal disease, altered hepatic function, chronic obstructive pulmonary disease), histology, staging, imaging data, treatment, recurrence, and survival. The data are imported to the de-identification system as a CSV file. After the de-identification, the system produces one JSON file containing the de-identified clinical data. It also generates a log file that includes all the attribute names, which are not de-identified in this process. Finally, the de-identified clinical data are mapped to the data model, and the actual mr_number is replaced with the de-identified patient ID. Smoking and drinking addiction details relate to the "Habits" in the schema. The comorbidities of the patient are kept in the "Comorbidity" module. The histological subtype, date of histopathological, and tumor location are incorporated in "Diagnosis". PETscan and planning ct dates are associated with the "Imaging data". Chemotherapy cycles and agents are mapped to the chemotherapy module. Total dose, number of fractions, concurrent chemotherapy, RT course correspond to the radiotherapy. The randomization group, local recurrence, regional recurrence, distant metastases details are wrapped in "Recurrence," and the date of the last assessment is saved in the survival status. The clinical records of an INTELHOPE patient are shown in Table 6. The patient's medical data field mapping, de-identified values, and associated modules are displayed in tabular form. It may transpire that the de-identified value is an empty string. For example, the patient has hypertension or not; A flag is set in "comorbidities___1" to get the status of that. Similarly, each delineated number with the "comorbidities___" specifies a comorbidity type. As shown in Table 6 Rows 4-9, the patient has hypertension and diabetes. Alongside this, the patient does not have Ischemic heart disease, chronic renal disease, Altered hepatic function, and Chronic obstructive pulmonary disease. The de-identified values of the comorbidity remain empty if it is not present in the patient. In such cases, the databank can identify these data and keep them out of the uploading process. As shown in Table 3, It shows the statistics on de-identified and not de-identified attributes. A total of 28 medical records needs to be de-identified, which is successfully de-identified. Similarly, 11 data fields are not de-identified as it is expected to keep the same.

## HYPORT-B Treatment Data Acquisition

In hypofractionated radiation therapy (HYPORT-B) protocol, ten breast cancer patients' data are taken. Those are having aggressive tumor pathology. A total of 30 studies are associated with the treatment process. First, the radiological data are de-identified, followed by the clinical data. Every radiological study contains a JSON file after the de-identification. It includes the anatomic site "Breast", indicates the determined laterality of the tumor and image type. The clinical data includes patient information (age, gender, registration date, performance status), smoking/drinking habits, diagnosis, stages, imaging features, stage information, comorbidities, survival, recurrence, and tumor pathology—grade. The treatment data consist of radiotherapy and chemotherapy. Some recurrent features are used for all the breast cancer patients, such as the intent of the treatment is palliative, recurrence type as a progressive disease, pathology being Invasive mammary carcinoma, adjuvant type chemotherapy is given.

## Lung Treatment Data Acquisition

Randomized Phase II of Immunotherapy With Pembrolizumab for the Prevention of Lung Cancer (IMPRINT-Lung) trial is used to treat the patients. As shown in Table 1, 60

patients' data consists of 300 studies is de-identified with the associated clinical data. The LDC is applied to the date and DateTime values for both datasets. The radiological images contain CT and CBCT modalities. The CT images are used for diagnostic and treatment planning. The CBCT images are taken for the therapy verification.

## Experimental Validation

The de-identified results are validated for the individual research project data. The clinical records de-identification is manually scrutinized by comparing it with the actual data—a group of the radiation oncologist and physicist involved in this process. The decisive points of the validation were the following,

- Feasibility of the data model for archiving in a databank.
- Harmonization effects to ensure data quality and consistency.

### Feasibility of the Data Model for Archiving in a Databank

After the completion of the de-identification process, the dataset is transformed into a specific structure. The de-identified data is mapped to the associated attribute and table name, which is defined in the database schema. All the medical records are then uploaded to the databank. Thus it is ensured that any data entered in the data bank should be screened by the de-identification system. Manual or external data entry is not allowed to ensure error-free and redundant data entry. We test the upload process with the existing data. The databank first recognizes the project details and obtains the dataset from the JSON. Then it extracts each object as a form of a database table. And it parses the inner object contains key/value pairs. The keys are attribute set, and values are the corresponding de-identified record. In the process, each data is captured and fit in the ER model of the databank. We have shown in Listing 2 the structure of data after de-identification.

```
{
    "projectid":"Project ID",
    "tables":[
            {
                "dcmpatientid":"De-identified Patient ID",
                "TableName":[
                    {"AttributeName1":"values1"},
                    {"AttributeName2":"Values2"}
                ],
                "objectid":"Number of instances"
            },
            {
                "dcmpatientid":"20130311122002202000002",
                "diagnosis":[
                    {"pathology":"non small cell carcinoma"},
                    {"anatomicsite":"Lung"},
                    {"diagnosis_date":"2017-08-23"},
                    {"diagnosis_site":"lung"},
                    {"laterality":"right"}
                ],
                "objectid":"1"
            }
    ],

}
```

**Technology Details of the JSON Dataset** The de-identified dataset is assembled in a single object into the JSON. A single object is a collection of name/value pairs, which are coated with curly braces. The JSON array is an ordered list of values and is covered by square braces. The complete de-identified data is a set of umpteen JSON objects and arrays. The outer object contains the two elements following project ID and table information. The "tables" is a JSON array that contains the entity list as set objects. Each object has three components: de-identified

**Table 4** Result set for query example 1

| Patient ID | Last Followup Date | Status | Date of Death |
|---|---|---|---|
| TTML019201303111220052021000002 | 2019-08-09 | Dead | 2020-02-06 |
| TTML019201303111220052021000003 | 2019-06-08 | Alive | |
| TTML019201303111220052021000020 | 2020-02-06 | Alive | |
| TTML019201303111220052021000032 | 2020-01-11 | Alive | |

patient ID ("dcmpatientid"), table name, and study occurrence. The "diagnosis" is the table name of the database schema. It contains the list of attributes and associated values in an object. This inner object is typically a key/value pair. Key is the column name of the table, and value is the corresponding attribute value. At last, "objectid" indicates the number of occurrences of the study for the particular patient.

We have shown in Fig. 3 the key attributes for connecting the clinical records with the radiological data. Every treatment module has a unique identifier that is created based on the "objectid" and the position in the hierarchy.

## Harmonization Effects to Ensure Data Quality and Consistency

The radiation oncologist and physicist exported the actual clinical data from the CDMS. Then de-identified JSON dataset was reviewed and verified manually against the original dataset. We observed that LDC was consistently applied, while any date format was changed in the YYYY-MM-DD form. A few clinical data were encoded correctly with code values. In this situation, the code values are replaced with the harmonized string. Project-specific data were also included in the de-identified dataset.

## Complex Query Processing and Benefits

The de-identified records are relational data. The desirable data can be extracted from given complex queries. A few typical examples are presented below.

- **Query example 1:** Find the survival status of more than 37-year-old female having invasive mammary carcinoma (ca) diagnosed of ca breast visceral crisis (cN2M1) human epidermal growth factor receptor (HER) 2 (+ve) started on anti HER 2 therapy.
- **SQL**

```
select * from Survival where PatientID in
    (select PatientID from PatientInformation
        where age > 37 and gender = 'F' and
        PatientID in
         (select PatientID from Diagnosis where
             pathology = 'Invasive mammary
             carcinoma' and anatomicsite = 'Breast'
              and diagnosisID in
              (select stage.diagnosisID from
                  StageInformation stage inner join
                  IHCProfile ihc on stage.
                  diagnosisID = ihc.diagnosisID
                  where stage.nstage = 'N1' and
                  stage.mstage = 'M1' and ihc.
                  protein_tested = '3' and ihc.
                  ihc_result = 'HER2')
        )
    );
```

**Table 5** Result set for query example 2

| Patient ID | DICOM Study UID | Image Type | Anatomic Site | Modality | DICOMReference |
|---|---|---|---|---|---|
| TMCKL019201303111227012020000004 | 2013031112270120200000 4.0 | RT Treatment Planning | Brain | CT | /home/surajit/ iMediX/tomcat7 /webapps/ CHAVIRO-DATABANK/ DICOM-DATA/ TMCKL0192013031112270120 2000004/ 2013031112270120200000 4.0 |

- **Utilization:** This kind of query is used for finding the survival details of real world patient (See Table 4).
- **Result:**
- **Query example 2:** Find the radiological studies of the alive patients who are suffering from Glioblastoma, IDH-wild type and the timing of the imaging is End of treatment with presence of edema having gross total resection surgery (See Table 5).
- **SQL**

```
select * from RadiologicalStudies where patientID
    in
    (select Diagnosis.patientID from Diagnosis
        inner join Survival on Diagnosis.patientID
         = Survival.patientID where pathology = '
        Glioblastoma, IDH-wildtype' and status = '
        Alive' and diagnosisID in
         (select diagnosisID from ImagingData
            inner join ImagingFeature on
            ImagingData.imageID = ImagingFeature.
            imageID where imaging_timing = 'End of
             Treatment' and edema = '1' and
            diagnosisID in
             (select diagnosisID from Surgery,
                StageInformation where
                surgery_details = 'Gross Total
                Resection' and Surgery.stageID =
                StageInformation.stageID)
         )
    );
```

- **Result:**
- **Utilization:** This case is worth in finding the radiological images from the clinical records.

## Discussion And Conclusion

There are several studies on patient health records and clinical data de-identification. An effective de-identification system, named "deid.pl" [11] is developed in 2009. It discusses different strategies for de-identifying the patient health records and clinical data. A development process of biomarker [12] is reported for quantitative imaging research. In this biomarker, the clinical data of head and neck cancer patients is manually extracted and stored in a PostgreSQL relational database. The same patient's PET/CT data are taken in DICOM format and demonstrate the capability of DICOM standard. Also, it represents the interrelationships between imaging and clinical information. A study on archiving image and clinical data of head and neck squamous cell carcinoma patients [13] was carried out in 2018. The patients are treated with curative intent RT. The clinical data comprises patient's demographics information, stages, risk factors, recurrence, grade, and survival data. The Posda tool [14] from TCIA is used for the preservation of the hierarchy structure. Another state-of-the-art describes non-small cell lung cancer dataset acquisition [15]. The radiological images are CT and PET-CT in this study. These data are de-identified using the CTP anonymization tool [16]. The patient information other than imaging data consists of the mutation profile, expression profile, imaging features, survival outcomes, smoking status, stage information, histopathological grade, and profile information such as age, sex, weight, and ethnicity.

The research-driven data model is developed, facilitating de-identification and segmenting diverse disciplines of cancer patients' data. This dataset can be exploited in inflexible and accurate modeling of information related to several research projects. The Oncospace is used for efficient data retrieval and curation. It has a very good impact on statistical analysis and other decision support on real-life clinical data. The proposed system harmonized the dataset to maintain data quality and consistency, which increases the usability of the data for further research. However, we have similarities like both methods are using a hierarchical ER model. In our exposition, the schema includes an extensive list of studies and treatment details. The system is compatible to accommodate the data that does not belong to the existing data model. We made the database schema generic to store various disease site data. The data can also be classified and segmented site-wise. Both the radiological and clinical data are involved in the de-identification process, which perpetuates

the interrelationships among them. An open-source application, the Posda tool uses a normalized permissive database schema that differs from traditional DICOM databases. It is capable of automatic integrity checking on a bulk basis DICOM data [17]. It collects imaging data, clinical data, pathology data, particular sites, cancer types, and treatment or imaging modalities. The clinical dataset is kept in a file and stored in the database. Our proposed schema represents a robust design. Each radiomics study can easily be classified as every entity is connected in the relational model. The classification can be done at each level like disease site wise, patient wise, study wise.

We recommend this system for radiation oncology patients' data collection. An experimental trial of de-identification and validation is performed on a few studies. The clinical data and imaging data both maintain referential integrity. The dataset is capture based on an underlined ER model in JSON format while the clinical data is normalized. This will helps in complex query processing and desirable data searching.

## A Appendix I

### The Input and Output Clinical Dataset Uses for De-Identification Process

**Table 6** The clinical dataset of a patient under INTELHOPE study

| Medical Terms (Semantic) | Input Attributes | Is Modified | Mapped Attributes | De-identified Values | Associated Module |
|---|---|---|---|---|---|
| Patient Hospital ID | mrn | Yes | systempatientid | 2013031803202000001 | Patient Information |
| Date of Registration | date_of_registration | Yes | date_registered_center | 2017-07-21 | Patient Information |
| Patient age | age | No | age | 61 | Patient Information |
| Hypertension | co_morbidities___1 | Yes | comorbidity_type | Hypertension | Comorbidity |
| Diabetes | co_morbidities___2 | Yes | comorbidity_type | Diabetes | Comorbidity |
| Ischemic heart disease | co_morbidities___3 | Yes | comorbidity_type | | Comorbidity |
| Chronic renal disease | co_morbidities___4 | Yes | comorbidity_type | | Comorbidity |
| Altered hepatic function | co_morbidities___5 | Yes | comorbidity_type | | Comorbidity |
| Chronic Obstructive Pulmonary Disease | co_morbidities___6 | Yes | comorbidity_type | | Comorbidity |
| WHO Performance Status | performance_status | No | performance_status | 1 | Patient Information |
| Habit of taking alcohol | alcohol | Yes | substance_used | | Habits |
| Habit of smoking | smoking | Yes | substance_used | Smoking | Habits |
| Histological Subtype | hpe_subtype | Yes | pathology | squamous cell carcinoma | Diagnosis |
| Date of histopathological diagnosis | hpe_date | Yes | diagnosis_date | 2016-07-20 | Diagnosis |
| Location of tumour | tumour_location | No | diagnsosis_site | 2 | Diagnosis |
| HPV status | hpvstatus | Yes | protein_tested | p16 | IHC Profile |
| | | | ihc_result | Positive | |
| Randomization Group | rand_grp | No | randomization_group | 1 | Name/value Pair |
| Tumour stage | t_stage | Yes | t_stage | 3 | Stage information |
| Nodal stage | n_stage | Yes | n_stage | 2C | Stage Information |
| Planning PET Scan date | petscan_date | Yes | imagedate | 2017-08-06 | Imaging data |
| Planning CT scan date | planning_ct_date | Yes | imagedate | 2017-08-06 | Imaging data |
| Planned Total Dose(Gy) of RT | total_dose | No | radiotherapy_dose | 66 | Radiotherapy |
| Total Number of Fractions of radiotherapy | total_fractions | No | radiotherapy_fractions | 30 | Radiotherapy |
| Radiotherapy Start Date | rt_start_date | Yes | startdate | 2017-08-19 | Radiotherapy |
| Radiotherapy End Date | rt_end_date | Yes | enddate | 2017-09-28 | Radiotherapy |
| Concurrent Chemotherapy | conc_chemo | No | concurrentchemotherapy | 1 | Radiotherapy |
| Number of cycles of chemotherapy given | chemo_cycles | No | Chemotherapy cycles | 6 | Chemotherapy |
| Chemotherapy agent used | chemo_agent | No | Chemotherapy regimen | CISPLATIN | Chemotherapy |
| Boost target volume | btv | No | boost_volume | 55.67 | Name/value Pair |
| Planning target volume | ptv1 | No | planning_target_volume | 188.03 | Name/value Pair |

**Table 6** (continued)

| Medical Terms (Semantic) | Input Attributes | Is Modified | Mapped Attributes | De-identified Values | Associated Module |
|---|---|---|---|---|---|
| 3 months followup date of assessment | mon3_date | Yes | last_followup_date | 01-01-2018 | Survival |
| Local Recurrence in 3 months followup | mon3_lr | Yes | responselocation | | Recurrence |
| Date of local recurrence in 3 months followup | mon3_date_lr | Yes | dateresponseassess | | Recurrence |
| Regional Recurrence in 3 months followup | mon3_rr | Yes | responselocation | | Recurrence |
| Date of regional recurrence in 3 months followup | mon3_date_rr | Yes | dateresponseassess | | Recurrence |
| Distant Metastases in 3 months followup | mon3_dm | Yes | responselocation | | Recurrence |
| Date of distant metastases in 3 months followup | mon3_date_dm | Yes | dateresponseassess | | Recurrence |
| Death known in 3 months followup | mon3_death | Yes | status | | Survival |
| Date of death in 3 months followup | mon3_date_death | Yes | dateofdeath | | Survival |

**Table 7** The clinical dataset of a patient under Glioblastoma multiforme study

| Medical Terms (Semantic) | Input Attributes | Is Modified | Mapped Attributes | De-identified Values | Associated Module |
|---|---|---|---|---|---|
| Patient Hospital ID | mr_number | Yes | systempatientid | 2013032611201900005 | Patient Information |
| | redcap_repeat_instance | No | objectid | 1…N | All |
| Date of first registration | date_first_registration | Yes | date_registered_center | 2016-11-12 | Patient Information |
| Patient Name | name | Not Included | X | X | X |
| Patient Age | age | No | age | 30 | Patient Information |
| Patient Gender | gender | No | gender | FEMALE | Patient Information |
| NPS/ECOG_baseline | performance_status | No | performance_status | 1 | Patient Information |
| MRI_Date | mri_date | Yes | imagedate | 2016-11-15 | Imaging Data |
| Location | location | Yes | site_abnormality | Frontal | Imaging Feature |
| MRI_Timing | mri_timing | Yes | imaging_timing | Post primary surgery | Imaging Feature |
| T1W Contrast enhancement | enhancement | Yes | enhancement | Yes | Imaging Feature |
| Perilesional edema | edema | Yes | edema | 1 | Imaging Feature |
| T2W Changes | enhancement_t2 | Yes | enhancement_t2 | 1 | Name/value pair |
| Subventricular zone involvement | svz | Yes | svz | 1 | Name/value pair |
| Crossing Midline | midline | Yes | midline | 0 | Name/value pair |
| Corpus Callosum Involvement | callosum | Yes | callosum | 0 | Name/value pair |
| Necrotic Core | necrotic_core | No | necrotic_core | 0 | Name/value pair |
| SVZ SIDE | side_svz | Yes | side_svz | 1 | Name/value pair |
| T1C volume in cc | t1c_volume | No | t1c_volume | 36.3 | Name/value pair |
| T2FL PBZ volume in cc | edema_vol | No | edema_vol | 78.4 | Name/value pair |
| T2 volume in cc | t2contrast_vol | No | t2contrast_vol | 81.4 | Name/value pair |
| MRI Disease Status | mri_disease_status | Yes | disease_status_img | Subtotal resection | Imaging Feature |
| MRI Disease Details | mri_disease_progression_details | No | disease_progression_type | 5 | Imaging Feature |
| Remarks MRI | remarks_mri | Not Included | X | X | X |
| Imaging Impression | imaging_impression | Not Included | X | X | X |
| Date of Surgery | date_of_surgery | Yes | surgerydate | 2016-10-17 | Surgery |

**Table 7** (continued)

| Medical Terms (Semantic) | Input Attributes | Is Modified | Mapped Attributes | De-identified Values | Associated Module |
|---|---|---|---|---|---|
| Extent of resection | extent_of_resection | No | extent_of_resection | 2 | Surgery |
| Surgery Remarks | surgery_remarks | Not Included | X | X | X |
| Pathology Date | pathology_date | Yes | pathology_date | 2016-11-27 | Diagnosis |
| Diffuse astrocytic and oligodendroglial tumours | diffuse_astrocytic_and_oligodendroglial_tumours | Yes | pathology | Glioblastoma, IDH-mutant | Diagnosis |
| MGMT methylation status | mgmt | Yes | gene_tested | MGMT | Expression Profile |
| | | | expression_result | Not Known | |
| TERT mutation | tert | Yes | gene_tested | TERT | Expression Profile |
| | | | expression_result | Not Known | |
| ATRX | atrx | Yes | gene_tested | ATRX | Expression Profile |
| | | | expression_result | Not Known | |
| TP53 | p53 | Yes | gene_tested | TP53 | Expression Profile |
| | | | expression_result | Wildtype | |
| Remarks | remarks | Not Included | X | X | X |
| Paraffin block available | wax_block | Not Included | X | X | X |
| Radiation therapy | rt_used_or_not | Yes | radiotherapy_given | Yes | Radiotherapy |
| Radiotherapy Dose | rt_dose | No | radiotherapy_dose | 1 | Radiotherapy |
| Radiotherapy Start Date | rt_start_date | Yes | radiotherapy_startdate | 2016-12-03 | Radiotherapy |
| Radiotherapy End Date | rt_end_date | Yes | radiotherapy_enddate | 2017-01-15 | Radiotherapy |
| Corticosteroid use at any point during RT | steroids_during_rt | Yes | concurrentmedication | Yes | Radiotherapy |
| RT_Concurrent TMZ @ 75mg/m2 | conc_tmz | Yes | concurrentchemotherapy | TMZ | Radiotherapy |
| Radiotherapy Remarks | rt_remarks | Not Included | X | X | X |
| Reirradiation Therapy | re_rt_used_or_not | Yes | is_reiradiation | Yes | Radiotherapy |
| Re-Radiotherapy Dose | re_rt_dose | Not Included | X | X | X |
| Re-Radiotherapy Start Date | re_rt_start_date | Yes | radiotherapy_dose | 2019-07-14 | Radiotherapy |
| Re-Radiotherapy End Date | re_rt_end_date | Yes | radiotherapy_enddate | 2019-09-13 | Radiotherapy |
| Corticosteroid use at any point during RE_RT | steroids_during_re_rt | Yes | concurrentmedication | Yes | Radiotherapy |
| RE_RT_Concurrent TMZ @ 75mg/m2 | conc_tmz_re_rt | Yes | concurrentchemotherapy | TMZ | Radiotherapy |
| Re-radiotherapy Remarks | re_rt_remarks | Not Included | X | X | X |
| Primary chemotherapy protocol | primary_chemo_protocol_used | No | chemotherapyregimen | 1 | Chemotherapy |
| Number of cycles | cycles | No | chemotherapycycles | 6 | Chemotherapy |
| Salvage Chemotherapy Protocol | salvage_chemotherapy_proto | Not Included | X | X | X |
| Number of cycles | cycle | Not Included | X | X | X |
| Date of last followup | date_registration | Yes | last_followup_date | 2018-03-29 | Survival |
| Survival Status at the last followup | status_lastfu | Yes | status | Dead | Survival |
| Date of death | date_of_death | Yes | dateofdeath | 2018-03-29 | Survival |
| Disease status at last followup | disease_last_followup | Yes | recurrencetype | Progressive Disease | Recurrence |
| Date of disease progression | progression_date | Yes | recurrencedate | 2017-11-11 | Recurrence |
| PFS Initial Treatment | pfs_initial_treatment | No | durationrecurrencefree | 403260 | Recurrence |

**Acknowledgements** This project is funded under National Digital Library of India (NDLI) sponsored by Ministry of Human Resource Development (MHRD), Govt. of India.

**Funding** This study has been funded by the Ministry of Human Resource Development IN (IIT/SRIC/CS/NDM/2018-19/096). None of the authors have conflicts of interest to declare. The CHAVI protocol is approved by the institutional review board at the Tata Medical Center Kolkata and consent waiver for taken for storing data from retrospective studies. The reference no is EC/GOVT/24/IRB23 on August 31, 2018. After the inception of the biobank, patients have given written informed consent for storing their images and clinical data in the biobank prospectively.

## References

1. W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle, "Understanding and using dicom, the data interchange standard for biomedical imaging," *Journal of the American Medical Informatics Association*, 4(3):199–212, 1997.
2. K. Aryanto, M. Oudkerk, and P. van Ooijen, "Free dicom de-identification tools in clinical research: functioning and safety of patient privacy," *European radiology*, 25(12):3685–3695, 2015.
3. P. Vcelak, M. Kryl, M. Kratochvil, and J. Kleckova, "Identification and classification of dicom files with burned-in text content," *International journal of medical informatics*, 126:128–137, 2019.
4. F. Prior, K. Smith, A. Sharma, J. Kirby, L. Tarbox, K. Clark, W. Bennett, T. Nolan, and J. Freymann, "The public cancer radiology imaging collections of the cancer imaging archive," *Scientific data*, 4:170124, 2017.
5. M. R. Bowers, T. R. McNutt, J. W. Wong, M. H. Phillips, K. R. Hendrickson, P. Kwok, W. Song, and T. L. DeWeese, "Oncospace consortium: A shared radiation oncology database system designed for personalized medicine and research," *International Journal of Radiation Oncology Biology Physics*, 93(3):E385, 2015.
6. U. UNNExT, UNESCAP, "Data harmonization and modelling guide for single windows environment," 2012.
7. P. A. Harris, R. Taylor, B. L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, et al., "The redcap consortium: Building an international community of software platform partners," *Journal of biomedical informatics*, 95:103208, 2019.
8. P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research electronic data capture (redcap)–a metadata-driven methodology and workflow process for providing translational research informatics support," *Journal of biomedical informatics*, 42(2):377–381, 2009.
9. S. Kundu, S. Chakraborty, S. Chatterjee, S. Das, R. B. Achari, J. Mukhopadhyay, and P. P. Das, "De-identification of radiomics data retaining longitudinal temporal information" *Journal of Medical Systems*, 2020.
10. M. W. Kan, L. H. Leung, and K. Peter, "The use of biologically related model (eclipse) for the intensity-modulated radiation therapy planning of nasopharyngeal carcinomas," *PloS One*, 9(11):e112229, 2014.
11. F. P. Morrison, S. Sengupta, and G. Hripcsak, "Using a pipeline to improve de-identification performance," In *AMIA Annual Symposium Proceedings*, volume 2009, page 447. American Medical Informatics Association, 2009.
12. A. Fedorov, D. Clunie, E. Ulrich, C. Bauer, A. Wahle, B. Brown, M. Onken, J. Riesmeier, S. Pieper, R. Kikinis, et al., "Dicom for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured pet/ct analysis results in head and neck cancer research," *PeerJ*, 4:e2057, 2016.
13. A. J. Grossberg, A. S. Mohamed, H. Elhalawani, W. C. Bennett, K. E. Smith, T. S. Nolan, B. Williams, S. Chamchod, J. Heukelom, M. E. Kantor, et al., "Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy," *Scientific data*, 5:180173, 2018.
14. W. Bennett, J. Matthews, and W. Bosch, "Su-gg-t-262: Open-source tool for assessing variability in dicom data," *Medical Physics*, 37(6Part19):3245, 2010.
15. S. Bakr, O. Gevaert, S. Echegaray, K. Ayers, M. Zhou, M. Shafiq, H. Zheng, J. A. Benson, W. Zhang, A. N. Leung, et al., "A radiogenomic dataset of non-small cell lung cancer," *Scientific data*, 5(1):1–9, 2018.
16. O. Brook, "Radiological society of north america, inc. ctp-the rsna clinical trial processor,"
17. W. Bennett, K. Smith, Q. Jarosz, T. Nolan, and W. Bosch, "Reengineering workflow for curation of dicom datasets," *Journal of digital imaging*, 31(6):783–791, 2018.