# Integrating Multimodal Radiation Therapy Data into i2b2

Eric Zapletal[1] Jean-Emmanuel Bibault[2,3] Philippe Giraud[2] Anita Burgun[1,3]

[1] Department of Medical Informatics, Biostatistics, and Public Health, Georges Pompidou European Hospital, Assistance Publique-Hôpitaux de Paris, Paris Descartes Faculty of Medicine, Paris, France
[2] Department of Radiation Oncology, Georges Pompidou European Hospital, Assistance Publique-Hôpitaux de Paris, Paris Descartes Faculty of Medicine, Paris, France
[3] INSERM UMR 1138 Eq22, Cordeliers Research Centre, Paris Descartes University, Paris, France

Address for correspondence Eric Zapletal, PhD, Department of Medical Informatics, Biostatistics, and Public Health, Georges Pompidou European Hospital, Assistance Publique-Hôpitaux de Paris, Paris Descartes Faculty of Medicine, Paris, France (e-mail: eric.zapletal@aphp.fr).

Appl Clin Inform 2018;9:377–390.

## Abstract

**Background** Clinical data warehouses are now widely used to foster clinical and translational research and the Informatics for Integrating Biology and the Bedside (i2b2) platform has become a de facto standard for storing clinical data in many projects. However, to design predictive models and assist in personalized treatment planning in cancer or radiation oncology, all available patient data need to be integrated into i2b2, including radiation therapy data that are currently not addressed in many existing i2b2 sites.

**Objective** To use radiation therapy data in projects related to rectal cancer patients, we assessed the feasibility of integrating radiation oncology data into the i2b2 platform.

**Methods** The Georges Pompidou European Hospital, a hospital from the Assistance Publique – Hôpitaux de Paris group, has developed an i2b2-based clinical data warehouse of various structured and unstructured clinical data for research since 2008. To store and reuse various radiation therapy data—dose details, activities scheduling, and dose-volume histogram (DVH) curves—in this repository, we first extracted raw data by using some reverse engineering techniques and a vendor's application programming interface. Then, we implemented a hybrid storage approach by combining the standard i2b2 "Entity-Attribute-Value" storage mechanism with a "JavaScript Object Notation (JSON) document-based" storage mechanism without modifying the i2b2 core tables. Validation was performed using (1) the Business Objects framework for replicating vendor's application screens showing dose details and activities scheduling data and (2) the R software for displaying the DVH curves.

**Results** We developed a pipeline to integrate the radiation therapy data into the Georges Pompidou European Hospital i2b2 instance and evaluated it on a cohort of 262 patients. We were able to use the radiation therapy data on a preliminary use case by fetching the DVH curve data from the clinical data warehouse and displaying them in a R chart.

**Conclusion** By adding radiation therapy data into the clinical data warehouse, we were able to analyze radiation therapy response in cancer patients and we have leveraged the i2b2 platform to store radiation therapy data, including detailed information such as the DVH to create new ontology-based modules that provides research investigators with a wider spectrum of clinical data.

**Keywords**
- data warehouse
- radiation therapy
- software validation

**received**
December 12, 2017
**accepted after revision**
April 7, 2018

DOI https://doi.org/10.1055/s-0038-1651497.
ISSN 1869-0327.

## Background and Significance

Clinical data warehouses (CDWs) have proven their efficiency for fostering translational research, and the research opportunities opened by secondary use of such clinical data has been demonstrated, e.g., in Vanderbilt[1] or Harvard.[2] Thanks to its early adoption of an electronic healthcare record (EHR) system in 2000,[3] The Georges Pompidou European Hospital (Hôpital Européen Georges Pompidou – HEGP) started a repository of structured and unstructured clinical data for care and research with the Informatics for Integrating Biology and the Bedside (i2b2) platform in 2008.[4] Since then, numerous data sources have been integrated into the HEGP CDW.[5]

In the early years of the HEGP CDW project, the priority was given to the integration of data that were both present in a standard format and frequently needed, such as demographic data, biology results, and medical activity codes (diagnosis, procedures, Diagnosis Related Group [DRG], etc.). Then, new data sources were added to support projects with different needs (clinical reports, EHR structured forms, etc.). More specifically, data related to cancer patients are of special interest for the CAncer Research and PErsonalized Medicine (CARPEM) program.[6] In 2012, the French National Cancer Institute (INCa) granted eight SIRICs (Site de Recherche Intégré sur le Cancer in French, or Integrated Cancer Research Site) labels in France. SIRICs' goals are to provide new operational resources to oncology research, to optimize and accelerate the production of knowledge, and to favor knowledge dissemination and application in patient care. CARPEM is one of these eight SIRICs, with focus on digestive, endocrine, head and neck, hematological, lung, ovarian, and renal tumors. More generally, the multidimensional characterization of cancer patients is the first step to achieve precision medicine and make decision based on far more complex diagnostic and prognostic categories than are currently in use. The multivariate descriptors of cancer patients will allow better understanding of the disease and develop new decision support tools derived from data to assist in everyday patient care.[7,8] With the objective of designing predictive models and assisting in personalized treatment planning, all available patient data need to be integrated and explored. However, the variables come from multiple fields such as genomics, imaging, biology, surgery, medical oncology, and radiation oncology.[9] To support cross-disciplinary research objectives leading to the development of personalized medicine, HEGP integrated in the last years chemotherapy data and is currently developing a program dedicated to personalized radiation oncology therapy.

The HEGP CDW is based on i2b2, an open source standard system developed by Harvard Medical School,[10] which has been adopted by more than 130 academic hospitals around the world.[11] However, while the core infrastructure of i2b2 is widely shared and improved by the community, Extraction/Transform/Load (ETL) modules are still mostly developed by each hospital to load data from their local information systems into i2b2, an approach that was used at HEGP. Some data sources used to populate the HEGP CDW were hosted in applications provided by private software vendors and others were hosted in applications provided by the Assistance Publique – Hôpitaux de Paris (AP-HP) institution to which HEGP belongs to. For the latter ones, documentations and technical support were easily available, but for the first ones, poor or no technical support was available. Therefore, a significant amount of time was spent to analyze the source data storage model to export the required observations.

Data sources have been imported in the CDW thanks to the i2b2 generic data storage model.

This generic storage model enables fast and efficient queries through an easy-to-use Web graphical interface dedicated to clinicians (the i2b2 Web client). However, this simplicity has a cost: every data source must be heavily transformed to fit the i2b2 data model. For some complex data sources, this transformation must be carefully analyzed because it has an impact on the way the data are latter accessed.

## Objective

Hospital data related to cancer treatment is mainly contained in the core electronic clinical record, but not limited to it. The radiation therapy data are usually produced and stored in dedicated systems and contains information in different formats. Such complex data needs to be integrated with other types of individual data in CDWs. However, the integration of radiation therapy data into i2b2 is an issue currently not addressed in referenced i2b2 sites.[12] In this article, we present the method developed to integrate radiation therapy data in the i2b2 CDW and its implementation at the HEGP.

## Methods

To integrate the radiation therapy data into the HEGP CDW, the actions listed below (also shown in ►Fig. 1) were performed.

### Selection of the Items to Integrate in the CDW

There are currently several treatment planning and record-and-verify systems for radiation oncology. Among these, VARIAN (ARIA and Eclipse) is used at HEGP. The selection of the items to integrate was established with a radiation oncology expert by analyzing the screens of VARIAN ARIA and Eclipse applications where data of interest were displayed (see Step "A" in ►Fig. 1). Four domains were retained:

1. "Dose details" domain
2. "Activities scheduling" domain
3. "Dose-volume histogram (DVH) curves" domain
4. "Couch correction" domain

These features, digitally recorded during treatment planning and delivery, are necessary to predict treatment outcomes (both efficacy and toxicity) in any predictive model.[8,13]

The structures and treatment plans associated to the DVH curves were selected but the images were not selected for the

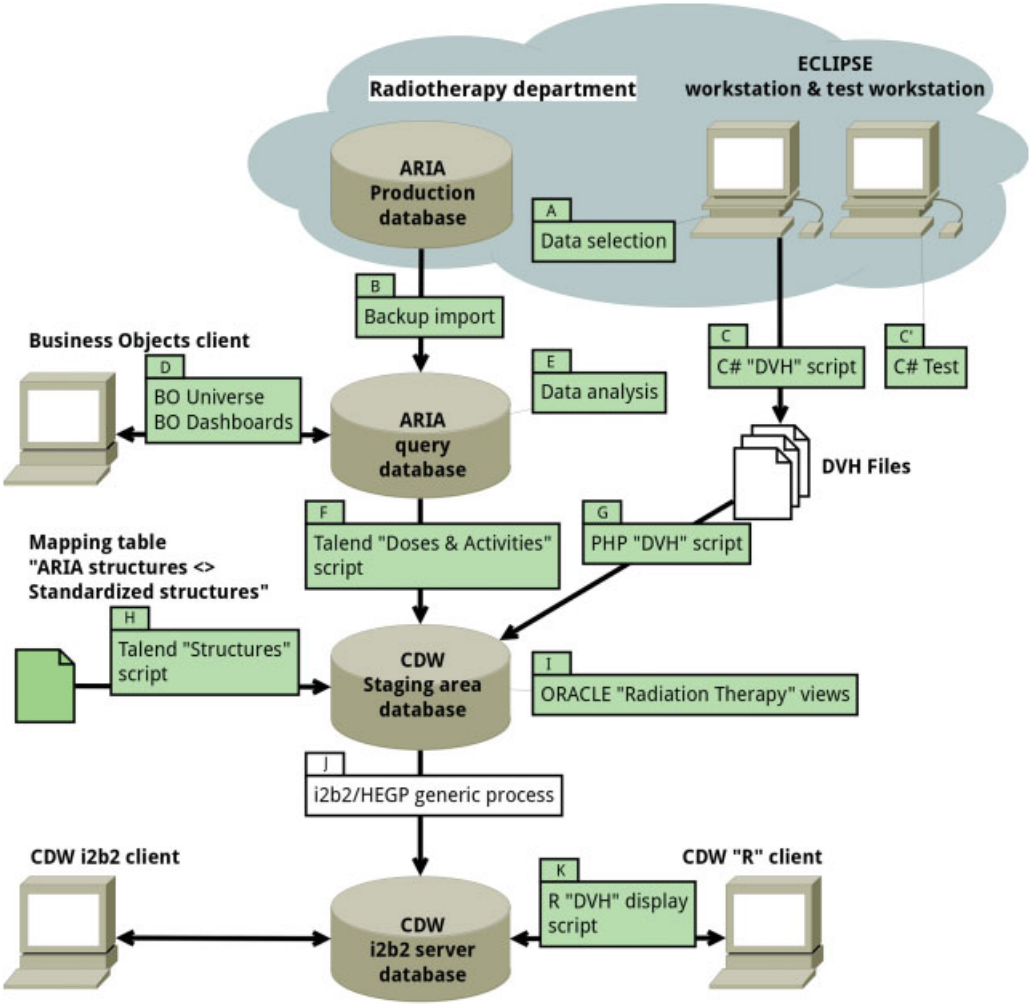


**Fig. 1** The integration pipeline of the radiation therapy data into the Hôpital Européen Georges Pompidou (HEGP) Informatics for Integrating Biology and the Bedside (i2b2) clinical data warehouse.

integration in the i2b2 repository since they are already available via the hospital picture archiving and communication system (PACS) architecture and not yet supported in the standard i2b2 data repository storage model, although some projects, such as the mi2b2 project,[14] allow fetching images from a remote PACS.

The pilot study has been limited to a patient cohort of 262 individuals. This pilot study was approved by the Institutional Review Board (IRB) and ethics committee CPP Ile-de-France II (IRB Committee # 00001072, study reference # CDW_2015_0024).

## Analysis of the Source System

A "backup" of the VARIAN/ARIA production database was installed on a specific computer to facilitate its analysis (see step "B" in ►Fig. 1) and a HTML documentation of the source storage model was generated by the SchemaSpy tool[15] to facilitate the analysis of the source system (see step "E" in ►Fig. 1).

The VARIAN/Eclipse software also offers an application programming interface (ESAPI): this Microsoft .NET package gives access to treatment data such as plans, images, doses, structures, and DVHs and it is actually available through a Web site[16] hosted by VARIAN.

## Radiation Therapy Data Extraction into the CDW Staging Area

From the analysis of the source storage model, two SQL procedures were developed for the "Dose details" and "Activities scheduling" domains. These two SQL procedures have been tested on the VARIAN/ARIA backup database.

For the "DVH curves" domain, a C# template script using the ESAPI package was retrieved from the VARIAN Web site and modified to suit the project needs to extract the DVH curves (see step "C" in ►Fig. 1). This C# export script has been tested and validated on a dedicated VARIAN/Eclipse test workstation including the whole VARIAN software suite but running a dedicated patient database not linked to the daily care process (see step "C'" in ►Fig. 1).

Together with the primary selected items, we decided to include as many as possible "related data" (annotated by a " + " symbol in the following enumerations):

For the "Doses details" domain, we extracted:

- Prescribed dose
- Received dose
  + Course ID
  + Plan Setup ID
  + Reference Point ID
  + Total Dose Limit
  + Daily Dose Limit
  + Session Dose Limit
  + Dose Delta

For the "Activities scheduling" domain, we extracted:

- Date/Time of the activity
- Duration of the activity
- Text comment entered by the physician during the activity

For the "DVH curves" domain, we extracted:

- DVH Curve vector:
  + Course ID
  + Anatomic structure ID
  + Volume of the affected anatomic structure
  + Coverage
  + Minimum dose
  + Maximal dose
  + Mean dose
  + Median dose
  + Standard deviation dose
  + Sampling coverage

To ensure integrity and quality prior to the integration into i2b2, these data were exported in the staging area of the HEGP CDW with Talend Open Studio scripts[17] and additional PHP scripts.

### Talend Open Studio Scripts

The two SQL procedures developed for the "Dose details" and "Activities scheduling" domains were integrated into the Talend Open Studio scripts to export the data of these domains from the radiation therapy backup database into the CDW staging area (see step "F" in ►Fig. 1).

### PHP Scripts

PHP scripts were used to import the DVH files (exported with the C# script) into the CDW staging area. This feature was not directly implemented in the C# script to minimize the dependencies of the global ETL workflow toward proprietary technologies such as the .NET framework (see step "G" in ►Fig. 1).

## Validation of the Data Imported in the Staging Area

The validation of the imported data has been a key issue in the whole integration process, requiring specific developments not directly used by the ETL modules. The data sets from the "Dose details" and "Activities scheduling" domains were validated by creating two Business Objects (BO)[18] dashboards replicating the screens of the vendor's application (see step "D" in ►Fig. 1). The rationale of this method is based on the fact that BO dashboards are built on an intermediate layer called a BO "Universe." The design of a BO "Universe" requires a manual extraction of the relevant relationships between objects in the source data (see ►Fig. 2). Therefore, the BO dashboards may be seen as a proof of a correct understanding of the source data model: if the BO dashboards display the same content as the Eclipse application, then it means that the relationships between objects in the source data model have been correctly interpreted. Furthermore, the tool used to create the BO dashboards is also a SQL generator and the artifacts generated by this tool have been used to validate the two SQL procedures designed to extract the data. Then, the comparison of the Eclipse application screenshots made during data selection (►Fig. 3) and the BO dashboards (►Fig. 4) on some selected patients allowed validating the inner structure of the data source model inferred from the analysis step. The data set from the "DVH curve" domain was validated by developing a R[19] script displaying the curve data extracted from the CDW (see step "K" in ►Fig. 1).

These validations steps were mainly manual processes as only a few patient data were used in the screen comparisons with the BO dashboards and the R script, but considering the complexity of these two data sets it was not possible to perform an automatic validation during this preliminary study.

## Integration of the Radiation Therapy Data into i2b2

### The Generic i2b2 Data Storage Model

The generic i2b2 data storage model is designed around a central facts table (OBSERVATION_FACT) that stores all the observations in an Entity-Attribute-Value (EAV) model and five additional dimensional tables are used to precisely qualify the observations.[20] In the i2b2 OBSERVATION_FACT table, observations contents are encoded through a small number of columns (the other columns are links to the dimensional tables, secondary qualifiers, or technical timestamps):

- Valtype_cd: stores the type of the observation content (string, number, or text)
- Tval_char: stores the value of a string-based observation
- Nval_num & Units_cd: store the value and the unit of a number-based observation
- Observation_blob: stores the value of a text-based observation

### Radiation Therapy i2b2 Concepts

*"Dose Details" and "Activities Scheduling" Domains*

Every observation in i2b2 is indexed by a set of concepts that are used by clinicians to build their queries. For the "Dose details" and "Activities scheduling" domains, three new concepts were created:

- "RTX:PRESCRIBEDDOSE"
- "RTX:ACTUALDOSE"
- "RTX:ACTIVITY"

*"DVH Curves" Domain*

There are several domain-specific information systems for radiation oncology: Elekta (MOSAIQ), VARIAN (ARIA),

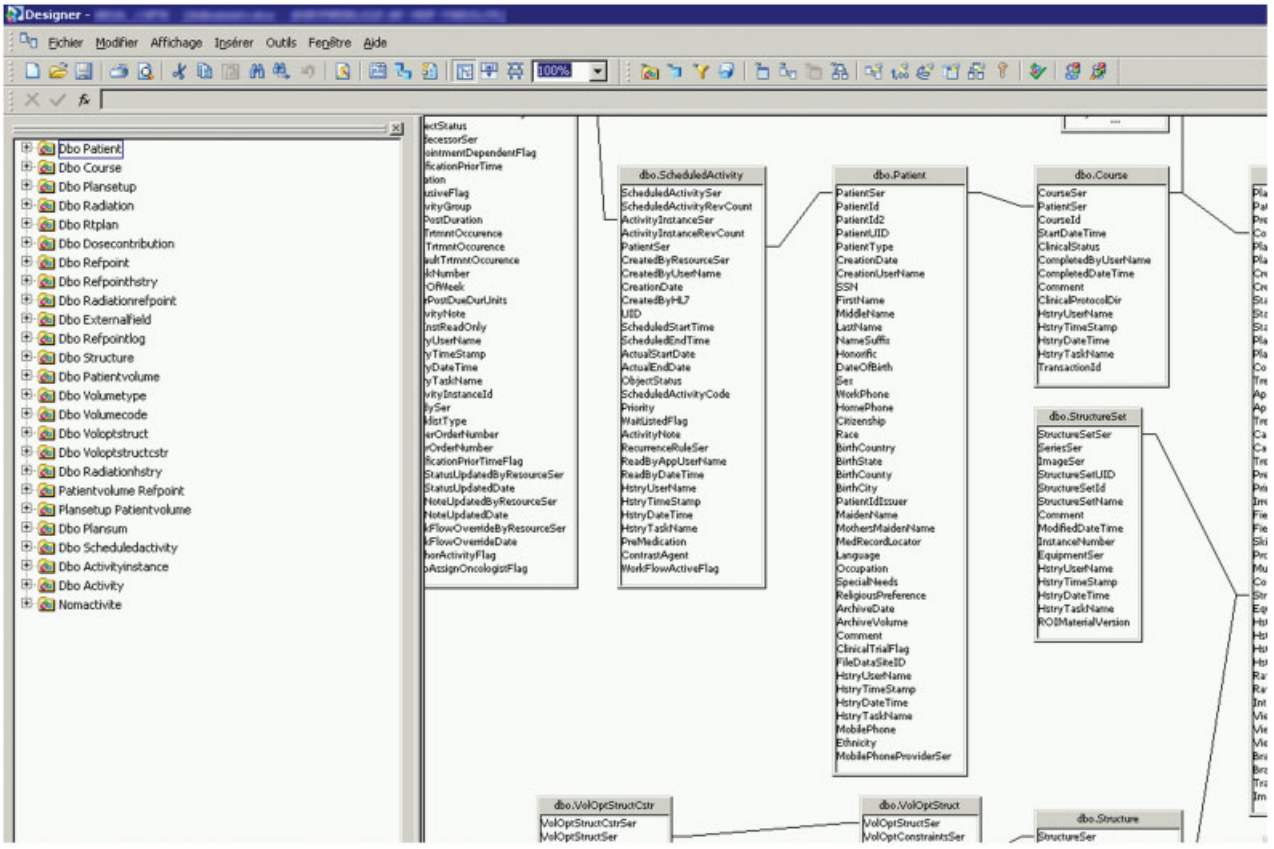


**Fig. 2** The Business Objects Universe created to validate the integration of radiation therapy data into the Hôpital Européen Georges Pompidou (HEGP) Informatics for Integrating Biology and the Bedside (i2b2) clinical data warehouse (CDW).

Accuray (Multiplan and Tomotherapy Data Management System), and BrainLab (iPlan). Each of these treatment planning and record-and-verify systems has its own structures labeling, which is not consistent across platforms, making it difficult to extract and analyze dosimetric data. For uniform data integration, we needed to create a classification and mapping that leads to an accurate ontology. A solution to this issue is to use an ontology, a set of common concepts that can be used, independently of the software, to represent medical knowledge, and in our case, anatomical and target volumes. There are currently around 440 biomedical ontologies. The most commonly used include Systematized Nomenclature of Medicine (SNOMED),[21] the National Cancer Institute (NCI) Thesaurus,[22] Common Terminology Criteria for Adverse Events (CTC AE),[23] and the Unified Medical Language System (UMLS) Metathesaurus.[24] These ontologies do not include specific radiation oncology terms, which led to the creation of the Radiation Oncology Ontology (ROO),[25] that reused other ontologies and added RO terms such as region of interest (ROI), target volumes (gross tumor volume [GTV], clinical target volume [CTV], planning target volume [PTV]), and DVHs. However, the ROO does not provide enough anatomical or target volume concepts for an easy use of routine practice data. For example, lymph nodes levels, that are essential for the planning of nodal CTV in radiotherapy, are not included.[26,27] Moreover, in the radiation therapy software, the names of the anatomical structure associated to the DVH are manually entered, leading to heterogeneity issues in labels. To address that issue and to enable semantic integration and standard representation of anatomy tailored for radiation therapy, we created a new ontology dedicated to radiation oncology structures: the Radiation Oncology Structure (ROS) ontology.[28] We then mapped all the original terms entered by the users to the concepts of the ROS ontology. The mapping table has been integrated into the staging area with a Talend Open Studio script (see step "H" in ►Fig. 1). However, the original name of the anatomical structure is stored as the main value of the i2b2 observation in the TVAL_CHAR column. As for the doses detail, the curve data and the other contextual data are stored as a semi-structured text field in the JavaScript Object Notation (JSON) format.

## Storing Radiation Therapy Structured Data as i2b2 Observations

Three levels of aggregation were actually used to model and store DVH data, as shown in ►Fig. 5:

1. A DVH: aggregation of contextual data (volume, coverage, minimum dose, etc.) and one optional curve data vector.
2. A curve data vector: aggregation of curve data points.
3. A point: aggregation of two coordinates.

### *Standard i2b2 Techniques for Managing Structured Data*

Although aggregations techniques are widely used in EHR for structuring and displaying data (such as DVH), the i2b2 observations table does not provide aggregations.[29] It is

**Contribution des faisceaux**

☑ Afficher tous les plans ☐ Masquer le coefficient ☐ Masquer les champs ☐ Afficher tous les points réf.

| Plan | Champ | UM | Coefficient [UM/Gy] | Dose au point de réf. CTV LARYNX[Gy] | Dose au point de réf. IIbas,III bilat[Gy] | Dose au point de réf. IV bilat[Gy] |
|---|---|---|---|---|---|---|
| GD FX | OAG FAP | 107 | 170.5856 | 0.630 | 0.630 | |
| | OAG | 28 | 104.5112 | 0.270 | 0.270 | |
| | LAT D | 29 | 106.2987 | 0.270 | 0.270 | |
| | LAT D FAP | 109 | 173.2256 | 0.630 | 0.630 | |
| | Dose planifiée par fraction | | | 1.800 | 1.800 | |
| | **Dose planifiée** | | | **39.600** | **39.600** | **0.000** |
| MEDSCL | ANT | 84 | 116.2425 | | | 0.720 |
| | POST | 82 | 113.2285 | | | 0.720 |
| | OAG | 21 | 116.8513 | | | 0.180 |
| | OAD | 20 | 111.8082 | | | 0.180 |
| | Dose planifiée par fraction | | | | | 1.800 |
| | **Dose planifiée** | | | **0.000** | **0.000** | **45.000** |
| MEP RED 1 | Dose planifiée par fraction | | | | | |
| | **Dose planifiée** | | | **0.000** | **0.000** | **0.000** |
| MEP RED 2 | Dose planifiée par fraction | | | | | |
| | **Dose planifiée** | | | **0.000** | **0.000** | **0.000** |
| MEP SCL | Dose planifiée par fraction | | | | | |
| | **Dose planifiée** | | | **0.000** | **0.000** | **0.000** |
| RED 1 | LAT D FAP | 161 | 179.2742 | 0.900 | 0.900 | |
| | OAG FAP | 160 | 177.3649 | 0.900 | 0.900 | |
| | Dose planifiée par fraction | | | 1.800 | 1.800 | |
| | **Dose planifiée** | | | **10.800** | **10.800** | **0.000** |
| RED 2 | OAG FAP | 155 | 287.2294 | 0.540 | | |
| | OAG | 40 | 109.7957 | 0.360 | | |
| | OAD | 40 | 109.8425 | 0.360 | | |
| | OAD FAP | 155 | 286.5487 | 0.540 | | |
| | Dose planifiée par fraction | | | 1.800 | | |
| | **Dose planifiée** | | | **18.000** | **0.000** | **0.000** |
| SCL | OAG | 99 | 109.5854 | | | 0.900 |
| | OAD | 96 | 106.8296 | | | 0.900 |
| | Dose planifiée par fraction | | | | | 1.800 |
| | **Dose planifiée** | | | **0.000** | **0.000** | **5.400** |

**Fig. 3** Partial view of the "Dose details" screen of the vendor's radiation therapy application.

possible to mitigate this limitation with two referenced techniques:

- Structuring the concepts ontology.[30]
- Use of concept modifiers.[31]

We considered that none of these approaches was suitable for the storage of the DVH because:

- The DVH vector is of variable size.
- The numerous repetition of X and Y "modifier-observations" would have led to an overhead of storage resources (a DVH curve contains often more than 600 points).

*A New JSON Document-Based Approach for Managing Radiation Therapy Structured Data in i2b2*

The JSON format was then chosen for the contextual and DVH data because it allows flexibility in the storage while preserving data consistency and indexing features by JSON dedicated packages.

JSON is widely used for storing objects in document-oriented databases[32] and NoSQL databases: CouchDB[33] provide schema-less feature with JSON-based items. Combined with parallel computation and incremental maintenance features, JSON databases offer valuable scalability performances. For this reason, they have been used for the storage of genomic data for research purposes.[34,35] However, in these contexts, JSON objects are not integrated into the i2b2 core database but stored in a dedicated database (CouchDB). Our approach is somewhat different since we store the JSON objects in the i2b2 core database so that they can be queried together with other data (demographic data, biology, drugs, etc.).

The contextual data and the DVH curves were then converted into JSON strings and stored in the OBSERVATION_BLOB column of the i2b2 facts table (OBSERVATION_FACT). This column may contain JSON data with a maximum size of 4 gigabytes-1 (in the ORACLE 11 g database).

The description of the radiation therapy observations in i2b2 is summarized in ►Table 1.

## The HEGP Generic Load Process

For each data source imported into the staging area, a set of ORACLE views have been designed to format the data and populate the i2b2 observations and concepts tables. Therefore, for the radiation therapy data source an additional set of ORACLE views were designed (see step "F" in ►Fig. 1). These ORACLE views were integrated into the HEGP generic load

| Contributions des faisceaux | | | | | | | Dose au point de réf. CTV LARYNX | Dose au point de réf. IIbas,III bilat | Dose au point de réf. IV bilat |
|---|---|---|---|---|---|---|---|---|---|
| Course | Plans | Nb fract. | Champs | N° | UM | Coefficient | | | |
| C1-LARYNX | GD FX | 22 | OAG FAP | 1 | 107 | 170,58 | 0,630 | 0,630 | |
| | | | OAG | 2 | 28 | 104,51 | 0,270 | 0,270 | |
| | | | LAT D | 3 | 29 | 106,30 | 0,270 | 0,270 | |
| | | | LAT D FAP | 4 | 109 | 173,22 | 0,630 | 0,630 | |
| | | | Dose planifiée par fraction | | | | 1,800 | 1,800 | |
| | | | **Dose planifiée** | | | | **39,600** | **39,600** | |
| | MEDSCL | 25 | ANT | 1 | 84 | 116,24 | | | 0,720 |
| | | | POST | 2 | 82 | 113,23 | | | 0,720 |
| | | | OAG | 3 | 21 | 116,85 | | | 0,180 |
| | | | OAD | 4 | 20 | 111,81 | | | 0,180 |
| | | | Dose planifiée par fraction | | | | | | 1,800 |
| | | | **Dose planifiée** | | | | | | **45,000** |
| | RED 1 | 6 | LAT D FAP | 1 | 161 | 179,27 | 0,900 | 0,900 | |
| | | | OAG FAP | 2 | 160 | 177,36 | 0,900 | 0,900 | |
| | | | Dose planifiée par fraction | | | | 1,800 | 1,800 | |
| | | | **Dose planifiée** | | | | **10,800** | **10,800** | |
| | RED 2 | 10 | OAG FAP | 1 | 155 | 287,22 | 0,540 | | |
| | | | OAG | 2 | 40 | 109,79 | 0,360 | | |
| | | | OAD | 3 | 40 | 109,84 | 0,360 | | |
| | | | OAD FAP | 4 | 155 | 286,54 | 0,540 | | |
| | | | Dose planifiée par fraction | | | | 1,800 | | |
| | | | **Dose planifiée** | | | | **18,000** | | |
| | SCL | 3 | OAG | 2 | 99 | 109,58 | | | 0,900 |
| | | | OAD | 3 | 96 | 106,83 | | | 0,900 |
| | | | Dose planifiée par fraction | | | | | | 1,800 |
| | | | **Dose planifiée** | | | | | | **5,400** |

**Fig. 4** Business Objects dashboard created to validate the "Dose details" data integration into Informatics for Integrating Biology and the Bedside (i2b2). This dashboard is replicating the vendor's application screen.

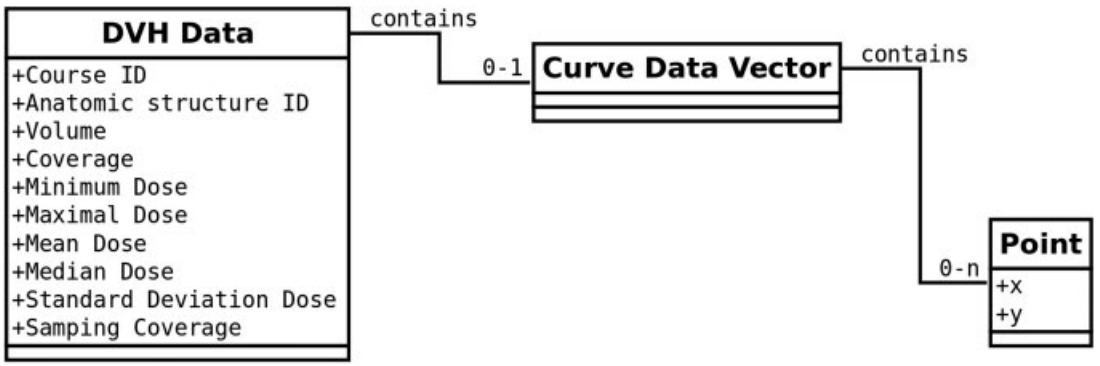


**Fig. 5** Modelization of the dose-volume histogram (DVH) data integrated into the Hôpital Européen Georges Pompidou (HEGP) Informatics for Integrating Biology and the Bedside (i2b2) clinical data warehouse (CDW).

process designed with the Talend Open Studio software suite[4] (see step "J" in ►Fig. 1).

### Validation of the Radiation Therapy Data Imported in the i2b2 CDW

All the validation steps were conducted by a computer scientist (E.Z.) and a radiation therapy specialist (J.E.B.).

### Validation of Prescribed Doses, Received Doses, and Activities Durations

A basic statistic test was performed to compare the data exported with the SQL procedures from the ARIA backup database and the data stored in the i2b2 repository: the count and average values (for the 262 patients' cohort) of prescribed doses, received doses, and activities durations

**Table 1** Description of the i2b2 storage content for the radiation therapy data

| OBSERVATION_FACT columns | Received dose | Prescribed dose | Activity scheduling | DVH Curve |
|---|---|---|---|---|
| ENCOUNTER_NUM | Encounter/stay sequential number | | | |
| PATIENT_NUM | Patient sequential number | | | |
| CONCEPT_CD | 'RTX: ACTUALDOSE' | 'RTX: PRESCRIBEDDOSE' | 'RTX:ACTIVITY' | 'RTX:' + standardized name of the anatomical structure |
| PROVIDER_ID | '@' | | | |
| START_DATE | PlanSetup.HstryDateTime | RTPlan.HstryDateTime | Scheduled Start Time | Structure.HstryDateTime |
| INSTANCE_NUM | 1 | | | |
| VALTYPE_CD | 'N' | | | 'T' |
| TVAL_CHAR | 'E' | | | Original name of the anatomical structure in the Radiation therapy software |
| NVAL_NUM | Sum(RefPointHstry.ActualDose) + RefPointLog.DoseDelta | RTPlan.PrescribedDose | The duration of the activity in minutes | |
| END_DATE | = START_DATE | | | |
| OBSERVATION_BLOB | {"CourseId"":C1 RECTUM," "PlanSetupId"":RECTUM.0," "RefPointId"":ISO RECTUM," "TotalDoseLimit"":46," "DailyDoseLimit"":2," "SessionDoseLimit"":2," "ActualDose"":46"} | {"CourseId"":C1 RECTUM," "PlanSetupId"":RECTUM," "RefPointId"":PELVIS," "PrescribedDose"":45"} | Type of activity + text comment | {"volume"":50.3174409637117,""coverage"":1," "minDose"":16.962 Gy,""maxDose"":48.520 Gy," "meanDose"":44.449 Gy," "samplingCoverage"":0.999752750762631," "medianDose"":45.973 Gy,""stdDev"":4.80722597080866," "curveData": [[0,50.317440963709], [0.1,50.317440963709], [0.2,50.317440963709],..., [48.4,0.37290217957042], [48.5,0.013851501133903]], "CourseId"":C1 RECTUM,""StructureId"":aire iliaque g"} |
| SOURCESYSTEM_CD | 'ARIA' | | | |

Abbreviation: i2b2, Informatics for Integrating Biology and the Bedside.
Note: The "curveData" vector field in the OBSERVATION_BLOB column is truncated for readability purpose in the above example.

were computed (1) by using the SQL procedures on the ARIA backup database and (2) in the i2b2 CDW repository. For each of these three items, the values in (1) and (2) were identical.

### Validation of the "DVH Curve" Domain

The "DVH curve" validation use case consisted of displaying in the R environment the DVH curves of randomly selected patients.

To achieve that, we first enabled the following extensions in R:

- DBI
- rJAVA
- RJDBC
- Rjson

Then, we designed a R script, built on three basic steps (as shown in ►Table 2):

1. A connection is first created to the CDW.
2. A simple SQL query is fetching DVH data for a given patient (based on his encounter number).
3. The resulting graph is then created by transforming JSON formatted DVH data into native R objects.

The output of this R script is presented in ►Fig. 6.

## Results

### A Pipeline for Integrating Radiation Therapy Data into i2b2

We developed a pipeline to integrate the radiation therapy data into the HEGP i2b2 instance and evaluated it on a cohort of 262 patients. The volumetry of the integrated data are shown in ►Table 3 and the overview of the pipeline is presented in ►Fig. 1.

### A New Radiation Therapy Ontology

A DVH is a curve modeled as a vector of points linked to an anatomical structure and, to enable query on specific structures of DVH in the CDW, an open ontology dedicated to ROSs was designed: The ROSs ontology (http://bioportal.bioontology.org/ontologies/ROS) has 417 classes, with a maximum of 14 children classes (average = 5) and is available as a Web Ontology Language online. The integration of the ROS ontology in the i2b2 Web client is presented in ►Fig. 7.

### A Hybrid Approach for Storing DVH Curve in i2b2

We have proposed a new format for storing DVH curves in i2b2 by using a document-based technique with JSON in the OBSERVATION_BLOB column of the i2b2 fact table without modifying the underlying i2b2 storage model.

### Business Intelligence Tools for Validating Extracted Data Set

The radiation therapy BO Universe created for this project is using 25 tables from the source data model and these tables are linked to each other with 28 relations. The tables and the links in the BO Universe were specifically designed to validate the "Dose details" and "Activities scheduling" data set but they can be useful to create other dashboards for various purposes in other projects.

## Discussion

### Toward Scalable Solutions Based on the Open i2b2 Standard

Mechanism for obtaining high quality, routine care, disease-specific, sharable data, is required in translational research. One solution is to leverage the i2b2 open source software to create new ontology-based modules that provides research

**Table 2** R script used to display DVH curves extracted from the CDW for a given patient defined by his encounter number

| | |
|---|---|
| A connection to the CDW is created with the JDBC driver | drv <- JDBC("oracle.jdbc.OracleDriver,"classPath = "/path/to/ojdbc6–11 g.jar,"" ")con <- dbConnect(drv, "jdbc:oracle:thin:@host:port:sid," "user," "password") |
| A simple SQL query is used to fetch only DVH data for a given patient (defined by his encounter number nnnnnn) | data <- dbGetQuery(con, "select tval_char, observation_blob from I2B2DEMODATA.OBSERVATION_FACT WHERE encounter_num = nnnnnn and concept_cd like 'RTX:%' AND concept_cd not in ('RTX:ACTUALDOSE', 'RTX:PRESCRIBEDDOSE', 'RTX:ACTIVITY')") |
| A graph is created in R by transforming JSON formatted DVH data into native R objects | attach(data)<br>curveData <- apply(data[2] 1, fromJSON)<br>colors = rainbow(dim(data)[1])<br>for (i in 1:dim(data)[1]) {<br>dose <- sapply(curveData[[i]]$curveData, '[', 1)<br>volume <- sapply(curveData[[i]]$curveData, '[', 2)<br>if (i == 1) {<br>plot(dose,volume,col = colors[i],type = "l," lty = 1)}<br>else {lines(dose,volume,col = colors[i], lty = 1)}<br>}<br>legend("bottomleft," legend = data[,1], col = colors, lty = 1, cex = 0.7) |

Abbreviations: CDW, clinical data warehouse; DVH, dose-volume histogram; JDBC, Java Database Connectivity; JSON, JavaScript Object Notation.

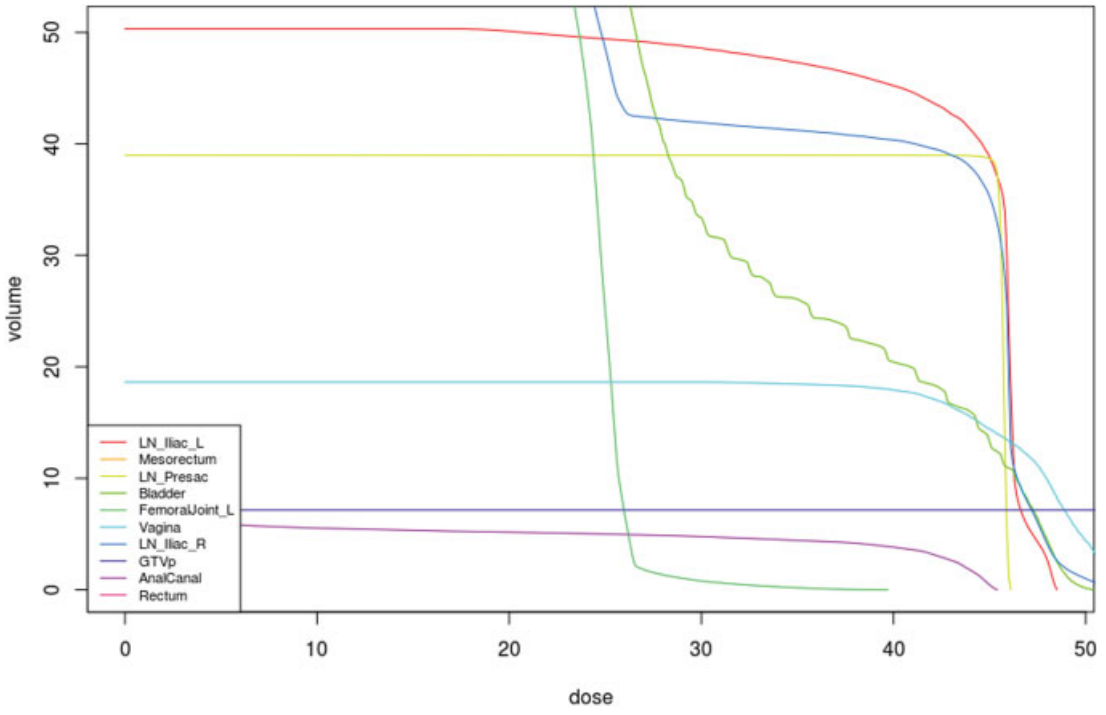


**Fig. 6** Output of the R script displaying radiation therapy data extracted from the Hôpital Européen Georges Pompidou (HEGP) Informatics for Integrating Biology and the Bedside (i2b2) clinical data warehouse (CDW).

investigators with the whole spectrum of clinical data. Examples of previous works focused on pediatric chronic disease registries[36] and cancer-related genomic data.[37,38] We have leveraged the i2b2 platform to store radiation therapy data, including detailed information such as the DVH, thanks to the i2b2 open standard.

### Shared Algorithms

We were able to integrate radiation oncology data into i2b2. The availability of the VARIAN/ESAPI was a key for success of integrating the DVH data. The integration of the "Dose details" and "Activities scheduling" data was much more time consuming.

We developed this pipeline in the context of AP-HP, the largest hospital group in Europe, with five radiation oncology departments treating 8,000 patients each year. We focused on the system used at HEGP, but the model can be generalized to other treatment planning and record-and-verify systems if the data model is similar. As a matter of fact, the SQL and the C# procedures used to extract the radiation therapy data from the vendor's application should be redesigned to fit a different data source model. The BO Universes that are connected to the radiation therapy backup database should also be redesigned since they are linked to the vendor's application data model. Unless vendor's applications have a unified API to access internal data, the ETL process is often specific to the source which makes this kind of task very time consuming. In the presented work, the only generic step occurs between the staging area and the i2b2 repository where a set of ORACLE views formats each different data sources of the staging area into i2b2 standardized data flows.

**Table 3** Volumetry of the radiation therapy data integrated in the HEGP i2b2 CDW with the initial 262 patients' sample

| i2b2 concept | Number of observations | Number of distinct patients | Total size of JSON objects |
|---|---|---|---|
| Prescribed dose | 791 | 246 | 75.2 kilobytes |
| Actual dose | 739 | 252 | 119 kilobytes |
| Activity | 7,631 | 262 | 197.2 kilobytes |
| DVH | 1,644 | 103 | 17.8 megabytes |
| Total | 10,805 | 262 | 18.2 megabytes |

Abbreviations: i2b2, Informatics for Integrating Biology and the Bedside; CDW, clinical data warehouse; DVH, dose-volume histogram; HEGP, Hôpital Européen Georges Pompidou; JSON, JavaScript Object Notation.

## Document-Based Storage in the i2b2 Architecture

### Compatibility of the JSON Document-Based Storage with the i2b2 Architecture

The i2b2 data repository (also referred to as the Clinical Research Chart cell) is a component of the broader i2b2 architecture that also defined a Web client for clinical users and a set of other cells that enable additional features such as Natural Language Processing, Correlation Analysis Plugin,

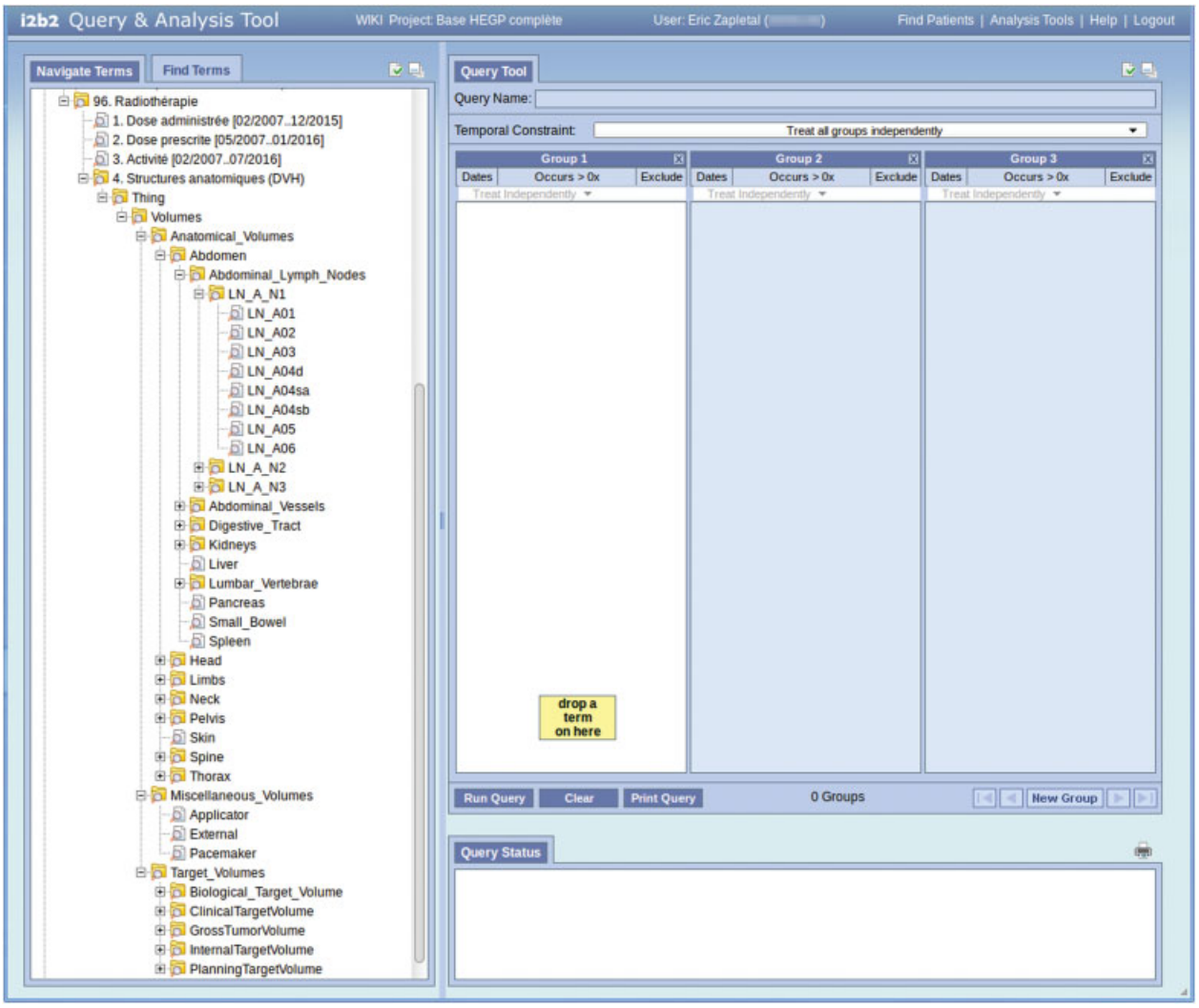


**Fig. 7** The Radiation Oncology Structures as displayed in the Informatics for Integrating Biology and the Bedside (i2b2) Web client.

etc. Therefore, any modifications in the i2b2 data repository storage model should take into account a compatibility aspect with the other cells and especially with the Web client since it is certainly the most used cell. This component is able to natively display numeric or string data fetched from the i2b2 repository but it should be modified to display the DVH curves. We did not start this task yet since we are still committed to achieve the loading of the complete radiation therapy data in the repository. However, the Web client features can be extended by using different techniques. The most straightforward one is based on the plugins extension mechanism of the i2b2 Web client.[39] Other more complex solutions could be derived from the SmartR project which is an open source platform for interactive presentations for the translational research data.[40]

### Estimation of the Target Volumetry for the DVH Data

For our 262 patients' cohort, the size (in character string length) of the JSON objects (i.e., all the contextual data plus the DVH curve data) is ~18.4 megabytes. With the hypothesis of a constant DVH allotment among the 14,000 patients in the radiation therapy software, the total size of the JSON DVH objects should be around 1 gigabyte as shown in ►Table 4.

As a comparison, the total size of all the clinical reports objects (that are already stored in the HEGP CDW) is around 17 gigabytes.

**Table 4** Estimation of the JSON objects volumetry for the entire radiation therapy database

| Domains | JSON objects size | |
|---|---|---|
| | Sample | Entire database |
| | 262 patients | 14,000 patients |
| DVH | 18.2 megabytes | 950.7 megabytes |
| Dose | 194 kilobytes | 10.1 megabytes |
| Activities | 197 kilobytes | 10.3 megabytes |
| Total | 18.2 megabytes | 971.1 megabytes |

Abbreviations: DVH, dose-volume histogram; JSON, JavaScript Object Notation.

With this estimation of the expected target volumetry, we think it is not required to use a native NoSQL framework for the management of the radiation therapy data.

### Validation of the Data Set

The validation of the imported data set is a key step in every ETL process, especially when the data are complex such as in the radiation therapy context.

By using the BO platform, we were able to validate each element of the data set in a "real life" use case by comparing the BO reports with the user's application screen. Validating the data set only by looking at the table content would have been unsatisfactory.

The simplicity of the R script needed to display the DVH curves from the CDW content is an indication that the JSON format is well suited for storing the DVH data.

### Perspectives

There are still several short- and long-term goals in this project. The "Couch Deviation" domain is not covered yet and we must find a strategy for fetching these data. We validated our approach on a cohort of 262 patients. It is now being extended to the whole VARIAN/ARIA data. Furthermore, the integration process requires a new offline VARIAN/ARIA copy database for dose details and activities scheduling update. We are also setting up new projects using this new data. For example, we are currently analyzing response to radiation therapy in T2–4 N0–1 rectal cancer patients, and we have integrated in our model genomic, clinical, and radiation therapy data. The work presented in this article is a significant step of the integration pipeline.

The method developed for this pilot project will be scaled up and used to integrate the data generated in all five AP-HP radiation oncology departments into the central AP-HP CDW (6.5 million patient records stored as of February 2017). Common data models and shared algorithms will reinforce (1) the central role played by the i2b2 CDW, and (2) the ability to mine cancer data and discover new markers in precision radiation therapy.

In the recent years, a set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have promoted the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles.[41] Behind FAIR principles is the notion that shared algorithms, tolls, and workflows are needed to search for relevant data sources, to analyze the data sets, and to mine the data for knowledge discovery. The researchers wanting to share and reuse data, methods, and scientific results will benefit from the application of the FAIR principles. As research in oncology is moving toward more data-intensive science, one of the grand challenges is to facilitate knowledge discovery by assisting researchers in their access to, integration and analysis of all data derived from routine care databases. We have developed the ROS ontology and the integration pipeline presented in this article to provide a semantic framework in radiation therapy that the sources and the users could agree upon to facilitate and accelerate data-driven cancer research.

## Conclusion

We have been able to integrate and reuse multimodal radiation therapy data for a preliminary study in the i2b2 platform. These data cover three functional domains:

- Dose details (delivered doses and prescribed doses)
- Activities scheduling (start time and duration of treatments)
- DVH curves

We used the standard i2b2 storage paradigm (EAV) for the doses and the activities scheduling by creating new radiation therapy concepts and by associating the scalar values (doses or dates) to the new concepts. For the contextual data and the DVH curve data, we used JSON formatted strings which may be easily converted into operational objects in frameworks daily used by researchers (such as R). A new domain ontology has also been created to annotate the DVH observations in a consistent and standardized manner. Some artifacts designed for the validation purposes (as the BO Universe) could also be used for various projects.

## Clinical Relevance Statement

We have leveraged the i2b2 platform to store radiation therapy data, including detailed information such as the DVH to create new ontology-based modules that provides research investigators with a wider spectrum of clinical data.

## Multiple Choice Questions

1. When designing a database for enabling queries of various clinical research projects which data model is best suited?

   a. The same data model as the patient healthcare record
   b. A set of specific data models dedicated to each patient data sources (demographic data, biology results, prescriptions, clinical reports, etc.)
   c. A set of specific data models dedicated to each clinical research projects
   d. A generic data model storage with no "source-oriented" nor "project-oriented" features

   **Correct answer:** The correct answer is option d. The resources needed to extract the patient data from their production environment are often so high that it may only be balanced out by the fact that the extracted data would be available for many other uses. Any source- or project-specific solutions would be an obstacle to the reuse of the data in a long-term perspective. Moreover, the patient health care record data model is optimized for storing data for health care-oriented task (patient past and current treatments queries, nurses planning displays, drug prescriptions controls, etc.) but it is not optimized for statistical queries such as "how many patients have been given this drug for these symptom?" The only acceptable data model is a generic model that can handle data coming from various sources and for various uses.

2. What is the most important benefit of having radiation therapy data in a clinical data warehouse?

a. Enabling physicists of radiation therapy departments to compute statistics on their data
b. Enabling clinicians to have access to radiation therapy departments' data
c. Enabling researchers to combine radiation therapy data with other categories of data
d. Enabling hospital managers to evaluate activities of radiation therapy departments

**Correct answer:** The correct answer is option c. The rationale of a CDW is to gather data from different clinical sources into a coherent repository where data are accessed through standardized axis of queries such as (1) patient axis ("what data belong to this patient?"), (2) encounter axis ("what data belong to that encounter?"), (3) time axis ("what data belong to that time frame?"), (4) provider axis ("what data have been produced by that provider?"), and (5) concept axis ("what data match this concept?"). By aggregating data from different sources in a coherent manner, the CDW enables queries with a wider scope than business oriented softwares dedicated to a specific data source. On the other hand, the CDW could not compete with this specific software when the query focuses on the health care process and activities (radiation therapy dose calculus, internal activities statistics, etc.).

## Protection of Human and Animal Subjects

This study from which the data were extracted was approved by the IRB and ethics committee CPP Ile-de-France II (IRB Committee # 00001072, study reference # CDW_2015_0024). Patients consent to participate to the study was implicit if refusal was not expressly stated. The HEGP CDW has been declared to the French CNIL regulatory commission for data privacy (# 1695855 v 0 ; 2013/08/28).

## Conflict of Interest

None.

## Acknowledgment

The authors are thankful to Arnaud Bernard, Alain Fauconnet, and Odile Taugourdeau for their valuable support regarding access to radiation therapy material.

## References

1 Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014;52:28–35
2 Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a Research Patient Data Repository. AMIA Annu Symp Proc 2006;20;06:1044
3 Degoulet P, Marin L, Lavril M, et al. The HEGP component-based clinical information system. Int J Med Inform 2003;69(2-3): 115–126
4 Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform 2010;160 (Pt 1):193–197
5 Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience. Int J Med Inform 2017; 102:21–28
6 Rance B, Canuel V, Countouris H, Laurent-Puig P, Burgun A. Integrating heterogeneous biomedical data for cancer research: the CARPEM infrastructure. Appl Clin Inform 2016;7(02): 260–274
7 Kohane IS, Drazen JM, Campion EW. A glimpse of the next 100 years in medicine. N Engl J Med 2012;367(26):2538–2539
8 Bibault J-E, Giraud P, Burgun A. Big Data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett 2016;382(01):110–117
9 Meldolesi E, van Soest J, Damiani A, et al. Standardized data collection to build prediction models in oncology: a prototype for rectal cancer. Future Oncol 2016;12(01):119–136
10 Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006;20;06:1040
11 i2b2: Informatics for Integrating Biology & the Bedside. Available at: https://www.i2b2.org/work/i2b2_installations.html. Accessed September 5, 2017
12 pubmeddev. No items found - PubMed - NCBI. [cited 2017 Dec 11]. Available at: https://www.ncbi.nlm.nih.gov/pubmed/?term=i2b2+radiation+therapy
13 Katz A, Kang J. Stereotactic body radiation therapy for low- and intermediate-risk of prostate cancer: disease control and quality of life at 8 years. Int J Radiat Oncol 2015;93(3, Supplement):E187
14 Murphy SN, Herrick C, Wang Y, et al. High throughput tools to access images from clinical archives for research. J Digit Imaging 2015;28(02):194–204
15 SchemaSpy. Available at: http://schemaspy.sourceforge.net/. Accessed April 10, 2017
16 Varian Developers Forum. CodePlex. Available at: https://variandeveloper.codeplex.com/Wikipage?ProjectName=variandeveloper. Accessed April 11, 2017
17 Talend Data Integration. ETL Software for Enterprise Data Integration. Talend Real-Time Open Source Data Integration Software. Available at: https://www.talend.com/products/data-integration/. Accessed April 12, 2017
18 BI & Analytics Platform | SAP BusinessObjects. SAP. Available at: https://www.sap.com/product/analytics/bi-platform.html. Accessed April 10, 2017
19 R: The R Project for Statistical Computing. Available at: https://www.r-project.org/. Accessed April 12, 2017
20 CRC_Design - CRC_Design.pdf-. Available at: https://www.i2b2.org/software/files/PDF/current/CRC_Design.pdf. Accessed April 14, 2017
21 SNOMED CT - Summary | NCBO BioPortal. Available at: https://bioportal.bioontology.org/ontologies/SNOMEDCT. Accessed April 28, 2017
22 National Cancer Institute Thesaurus - Summary | NCBO BioPortal. Available at: https://bioportal.bioontology.org/ontologies/NCIT. Accessed April 28, 2017
23 Common Terminology Criteria for Adverse Events - Summary | NCBO BioPortal. Available at: https://bioportal.bioontology.org/ontologies/CTCAE. Accessed April 28, 2017
24 UMLS Metathesaurus Fact Sheet. Available at: https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. Accessed April 28, 2017
25 Radiation Oncology Ontology - Summary | NCBO BioPortal. Available at: http://bioportal.bioontology.org/ontologies/ROO. Accessed April 28, 2017
26 Grégoire V, Ang K, Budach W, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiother Oncol 2014;110(01):172–181
27 Rusch VW, Asamura H, Watanabe H, Giroux DJ, Rami-Porta R, Goldstraw P. The IASLC lung cancer staging project: a proposal for a new international lymph node map in the forthcoming seventh

edition of the TNM classification for lung cancer. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer 2009;4(05):568–577
28 Bibault J-E, Zapletal E, Rance B, Giraud P, Burgun A. Labeling for Big Data in radiation oncology: the Radiation Oncology Structures ontology. PLoS One 2018;13(01):e0191263
29 Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Med Res Methodol 2009;9:70
30 Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016;63:277–294
31 Modifiers in i2b2 Data Model - i2b2 Developer's Forum - i2b2 Community Wiki. Available at: https://community.i2b2.org/wiki/display/DevForum/Modifiers+in+i2b2+Data+Model. Accessed April 12, 2017
32 Modeling and Querying Data in MongoDB. Available at: http://www.ijser.org/paper/Modeling-and-Querying-Data-in-MongoDB.html. Accessed April 12, 2017
33 Apache CouchDB. Available at: http://couchdb.apache.org/. Accessed April 27, 2017
34 Murphy SN, Avillach P, Bellazzi R, et al. Combining clinical and genomics queries using i2b2 - three methods. PLoS One 2017;12(04):e0172187
35 Gabetta M, Limongelli I, Rizzo E, Riva A, Segagni D, Bellazzi R. BigQ: a NoSQL based framework to handle genomic variants in i2b2. BMC Bioinformatics 2015;16:415
36 Natter MD, Quan J, Ortiz DM, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc 2013;20(01):172–179
37 Segagni D, Ferrazzi F, Larizza C, et al. R engine cell: integrating R into the i2b2 software infrastructure. J Am Med Inform Assoc 2011;18(03):314–317
38 Segagni D, Tibollo V, Dagliati A, et al. The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. Stud Health Technol Inform 2011;169:887–891
39 Web Client Plug-in Developers Guide - i2b2 Web Client - i2b2 Community Wiki. Available at: https://community.i2b2.org/wiki/display/webclient/Web+Client+Plug-in+Developers+Guide. Accessed February 1, 2018
40 Herzinger S, Gu W, Satagopam V, et al; eTRIKS Consortium. SmartR: an open-source platform for interactive visual analytics for translational research data. Bioinformatics 2017;33(14):2229–2231
41 Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018