Radiotherapy and Oncology 110 (2014) 370–374

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Data sharing

International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining

Erik Roelofs [a,1], André Dekker [a,*,1], Elisa Meldolesi [b], Ruud G.P.M. van Stiphout [a], Vincenzo Valentini [b,1], Philippe Lambin [a,1]

[a] Department of Radiation Oncology (MAASTRO), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center (MUMC+), The Netherlands; [b] Department of Radiation Oncology, Policlinico Universitario Agostino Gemelli, Rome, Italy

ARTICLE INFO

Article history:
Available online 3 December 2013

Keywords:
Medical informatics
Knowledge engineering
Machine learning
Data mining
Open source software

ABSTRACT

Extensive, multifactorial data sharing is a crucial prerequisite for current and future (radiotherapy) research. However, the cost, time and effort to achieve this are often a roadblock. We present an open-source based data-sharing infrastructure between two radiotherapy departments, allowing seamless exchange of de-identified, automatically translated clinical and biomedical treatment data.

Sharing data across institutions is required for multi-institutional radiotherapy research [1]. Besides exchanging data for specific research projects, there is a recognized need to establish a culture of data pooling both within the radiotherapy [2] and the broader cancer community [3]. For the transition from population based treatment options (where "one size fits all") toward personalized medicine we are increasingly depending on decision support systems that require large heterogeneous datasets [4–8]. Randomized controlled trials hardly offer such data with only 3% of adult cancer patients included in trials [9–11]. However, aggregating routinely collected real-time biomedical patient data and innovative "rapid-learning" research techniques allow us to use the knowledge of the masses for the benefit of the individual [3,12,13].

Medical informatics driven research, for instance in the field of predictive modeling, requires a large amount of data to provide sufficient statistical power to act as acceptable decision supporting tools. Furthermore, another substantial amount of data is needed for validation of the models, preferably by external datasets.

This brings up some stringent and challenging demands on the quality as well as the quantity of the data. Data of inferior quality do not improve by pooling it with other data. It can actually worsen the value of good datasets. Work by the Quality Assurance Review Center (QARC; http://www.qarc.org) and a review by the EORTC show the importance of proper Quality Assurance (QA) programs in collaborative efforts and the long history thereof in the field of Radiotherapy [14–16].

Another challenge for data-sharing initiatives in the field of biomedical research is that the investigated data are often multifactorial, comprising of laboratory data, diagnostic and clinical imaging and treatment outcome data, among others. Combining these data securely can be troublesome, even when sharing between departments within the same institution, let alone when between institutions, especially international ones.

Furthermore, in many research projects, dedicated data management staff are required to translate and copy data into trial-specific case report forms and/or dedicated IT staff are needed to de-identify DICOM images or build databases that are suitable for machine learning and data mining techniques. However, without dedicated staff, the sheer amount of time it takes to collect, de-identify and share data often is a roadblock to participation in clinical research. With many research projects not or underfunded, especially in the initiation phase, one requires existing staff to balance other duties with these research requests. This causes the process of data sharing to take a long time, despite the cooperation and willingness of everyone involved.

In this technical report, we describe one way to quickly build a low cost, infrastructure that makes sharing of data easier between two institutions wishing to work together, but having different IT systems. This infrastructure was implemented to share data from the Policlinico Universitario Agostino Gemelli in Rome, Italy (Gemelli) to the MAASTRO Clinic in Maastricht, the Netherlands (MAASTRO) to facilitate research projects such as the Thunder clinical trial (NCT00969657, http://ClinicalTrials.gov) and "knowledge

[*] Corresponding author. Address: Department of Radiation Oncology (MAASTRO), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center (MUMC+), Dr. Tanslaan 12, 6229 ET Maastricht, The Netherlands.
E-mail address: andre.dekker@maastro.nl (A. Dekker).

[1] These authors contributed equally to this work.

0167-8140/$ - see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.radonc.2013.11.001

engineering" research using data-mining and machine learning techniques to develop predictive models for various cancer sites (http://www.predictcancer.org).

Material and methods

Clinical data sources

In general, radiotherapy research requires various types of information:

- Clinical data (e.g. demographics, TNM-stage, date of diagnosis, histopathology, etc.).
- Diagnostic imaging data (e.g. diagnostic and follow-up PET, CT and MR imaging).
- Radiotherapy treatment planning data (e.g. delineation, planning-CT, dose matrix, beam setup, prescribed dose and fractions).
- Radiotherapy treatment delivery data (e.g. cone beam CTs, Orthogonal EPID imaging, delivered fractions).
- Non-radiotherapy treatment data (e.g. surgery, chemotherapy).
- Outcome data (e.g. survival, local control, toxicity).

Typically, in a radiotherapy department, this information is scattered across a number of data sources from a variety of vendors, which do not necessarily share the same patient identification number. In the case of Gemelli, the data sources were as given in Table 1.

Data model

For the research database (DB), a patient-centric data model was designed that would enable queries for both medical data and the existence of imaging data (Fig. 1). A simple data model was deliberately chosen to allow easy identification of core disease characteristics but with most information in the form of lists of performed procedures as well as performed imaging study and series.

De-identification

A coding scheme was employed in which a secure database is maintained that holds the link (Key) between a unique random patient identification code (ID) and all directly identifying data (ID's as used in the clinical data sources, name, birth date etc.). This Key is maintained by local hospital personnel and only accessible from within the firewall of the hospital. In the research DB, the patient is only identified by the research ID. Rather than applying an irreversible anonymization method to the patient data, we used "coding" or "pseudo-anonymization" to enable extending the datasets with additional information at a later stage, which would otherwise be impossible to do.

Some data elements were not de-identified as they were considered to be important for the research while they only carry a small risk of identifying a patient. The de-identification scheme was reviewed and approved by the local ethics authority. The elements that were kept were: exact dates of various procedures (including treatments), exact dates of diagnosis & death, DICOM UIDs and CT and MR imaging of the head.

Clinical terms and translation

To convert Italian to English standard terms, SNOMED Clinical Terms [17] were used as the dictionary. SNOMED CT is considered as the most comprehensive multilingual medical terminology in the world. A separate database was maintained in which local terms were mapped to the SNOMED-CT concepts and both the preferred term and the concept ID were stored.

Research hardware & software

On a research workstation (Windows 7 64-bit, Intel Xeon, 2.53 GHz, 4 GB RAM) the following software was installed: SQL Server 2008 (free Express version, Microsoft, Redmond, WA); Clear Canvas Image Server and Workstation (both free and open source, Clear Canvas, Toronto, Canada); DCMTK DICOM toolkit (free and open source, Offis, Oldenburg, Germany); RSNA Clinical Trial Processor (CTP) (free and open source, RSNA, Oakbrook, IL) and Matlab Compiler Runtime (MCR) engine (free, Mathworks, Natick, MA).

Clear Canvas Image Server is a PACS and was installed with a temporary partition holding identifiable DICOM headers and a research partition holding only de-identified DICOM objects. The Clear Canvas Workstation was used for DICOM import of the Nuclear Medicine department's optical disks. SQL Server 2008 was used to host the mentioned databases as well as the database of the Clear Canvas Image Server.

Finally, for data synchronization a variety of SQL scripts was designed and run through the command line interface. The MCR engine was used to run compiled custom Matlab code in which the DCMTK toolkit was called. The CTP package was used to build a de-identification pipeline (DICOM import → de-identification → DICOM Export) and a file export pipeline (DICOM Import → File Export to shared directory). CTP allows customizable de-identification settings through a web interface or by directly editing an XML configuration file.

Table 1
Data sources at Gemelli.

Data type	Data format	Database	Department	Name, Vendor
Clinical Outcome Non-RT treatment	Text	SQL database	Multiple	Spider, Opengraph (local development)
Diagnostic imaging	DICOM-CT DICOM-MR DICOM-PT	DICOM server	Radiology	Careview, Kodak
	DICOM-PT	Optical Disks	Nuclear Medicine	PET workstation, Philips
RT treatment planning	DICOM-CT DICOM-RTDOSE DICOM-RTIMAGE DICOM-RTPLAN DICOM-RTSTRUCT	DICOM server	Radiotherapy	Aria, Varian
RT treatment delivery	Text DICOM-RTRECORD DICOM-RTIMAGE	Sybase Database DICOM server	Radiotherapy	Aria, Varian