Cancer Registry and Big Data Exchange

Citation for published version (APA):

Shi, Z., Wee, L., & Dekker, A. (2019). Cancer Registry and Big Data Exchange. In *Big Data in Radiation Oncology* C R C Press. <https://doi.org/10.1201/9781315207582-11>

Document status and date:

Published: 20/03/2019

DOI:

[10.1201/9781315207582-11](https://doi.org/10.1201/9781315207582-11)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Cancer registry and big data exchange

Zhenwei Shi, Leonard Wee, and Andre Dekker

Contents

11.1	Introduction	154
11.2	Cancer Registry	155
11.2.1	Different Types of Cancer Registry	155
11.2.1.1	Hospital-Based CRs	156
11.2.1.2	Population-Based CRs	156
11.2.2	Examples of Cancer Registries	156
11.2.2.1	Surveillance, Epidemiology, and End Results	156
11.2.2.2	National Radiation Oncology Registry	156
11.2.2.3	Registro Tumori Ospedaliero (RTO)	157
11.2.2.4	Dutch Surgical Colorectal Audit	157
11.2.3	Data Sources of CRs	157
11.2.4	Receiving and Reporting Information in CRs	161
11.2.4.1	Receiving Information	161
11.2.4.2	Reporting Information	164
11.2.5	Assessment of Data Quality Within CRs	164
11.3	From CR to Big Data in Radiation Oncology	164
11.3.1	Barriers of CRs	164
11.3.2	Big Data in Radiation Oncology	166
11.4	Big Data Exchange in Radiation Oncology	168
11.4.1	Big Data Collection	168
11.4.2	Standard and Framework for Data Exchange	169
11.4.3	Data Pooling Architectures	169
11.4.3.1	Centralized Architecture	169
11.4.3.2	Decentralized Architecture	170
11.4.3.3	Hybrid Architecture	170
11.4.4	Data Interoperability	171
11.4.4.1	Semantic Web	171
11.4.4.2	Ontology	172
11.4.5	Data Exchange	173
11.4.5.1	Send Data Out	173
11.4.5.2	Send Questions in	174
11.4.5.3	Centralized Versus Distributed Learning Architecture	174

11.5	Barriers of Big Data Exchange among Multicenters	176
11.5.1	Administrator Barrier	176
11.5.2	Ethical Barrier	177
11.5.3	Political Barrier	177
11.5.4	Technical Barrier	177
11.6	Conclusion	177
	References	178

11.1 INTRODUCTION

A vast stream of data is generated by the routine operations of modern cancer diagnosis and oncologic treatments. Usually, the data is stored electronically but it tends to be scattered across different disciplines (e.g., radiation oncology, radiology, medical oncology, surgery), different data storage platforms (e.g., electronic medical record [EMR], Picture Archiving Communication System [PACS]) and in a wide variety of formats (e.g., Digital Imaging and Communication in Medicine [DICOM], ASCII, PDF) (Deng 2014). In addition, data in cancer care has the particular properties of large volume and complex dependencies between data elements, which is creating growing difficulties for conventional methods of data handling. By handling, we refer to collection, storage, update, and exchange. Although the variety and volume of big data continues to grow exponentially within the field of oncology (Chen et al. 2014), it has not been easy to exploit this rich vein of data to improve patient safety and health outcomes (McNutt et al. 2016).

Data, as a whole, in the clinical routine is one of the most valuable, but most underutilized, assets within radiotherapy and oncology studies (Roelofs et al. 2013). Rapid Learning Health Care (RLHC) (Lambin et al. 2013a) envisions virtuous cycles of rapid knowledge generation and knowledge utilization within health care, by combining routine clinical practice with prospective research studies using a big data exchange paradigm. The resulting interconnected web of data repositories across departments and institutions becomes a global resource to mine for new knowledge, generate hypotheses for novel treatments, and test the effectiveness of interventions in real-world settings. RLHC also leads to a wide range of distributed software applications that can exploit FAIR (findable, accessible, interoperable, and reusable) data repositories (Wilkinson et al. 2016) to create predictive outcome models for clinical decision support (Lambin et al. 2013b) and to discover predictive digital signatures in biomedical images (i.e., radiomics) (Aerts et al. 2014).

To be effective in clinical decision support, predictive models must be able to estimate the probability of a given outcome over a range of clinical situations. Therefore, an approach that integrates the data of many patients over different treatment settings is essential. Predictive outcome models have the potential to improve quality of life, identify patients at high (or low) risk, and to prolong the survival of patients with cancer (Dehing-Oberije et al. 2009, 2010, 2011; Oberije et al. 2014). Some predictive outcome models for various cancer sites can be found at <http://www.predictcancer.org>. These prediction models support the practice of personalized radiotherapy that is tailored to the individual risk profile of the patient. If multiple treatment possibilities exist where clinical evidence is equivocal, patient and physician preferences for certain outcomes may play a role. The collaborative consultation process between a patient and their treatment physician is known as shared decision making (Elwyn et al. 2012; Stacey et al. 2014). Reliable model-based predictions of potential treatment responses are a prerequisite for information trade-offs between the risks of harm and strength of treatment in the shared decision-making paradigm. In keeping with RLHC, observational data on quality-of-life, patient-reported outcomes, and decision regret should put back into the web of clinical knowledge, so that continually updated models are better able to inform physician's and patients' decisions.

We have prefaced this chapter with a consideration of the wide-ranging opportunities for clinical innovation and improved outcomes that could be possible with comprehensively voluminous, multimodality, and multi-institutional data. Central to this ambition is the integration of all the data that presently remains underutilized in closed, unconnected private repositories. The remainder of this chapter is organized as follows: first, we explore the paradigm of data centralization in the form of a

AQ 1

cancer registry (CR). This, whether large or small, represents the current orthodoxy for collection, storage, analysis, and distribution of oncology data. Second, as the multiplicity, volume, and dimensionality of cancer data continues to grow rapidly, we shall argue that a new paradigm for data exchange is urgently needed. We thus describe practical and feasible big data architectures that can be applied to overcome the challenges of poorly connected data repositories in radiation oncology. Third, the remaining barriers impeding universal rollout of big data exchange architecture in radiation oncology field will be examined, as will the unique characteristics of certain data architectures that may be more naturally adapted to overcoming these barriers. The primary aim of this chapter is to support the understanding of readers about big data exchange architectures that facilitate rapid learning in clinical practice and scientific research in the field of radiation oncology.

11.2 CANCER REGISTRY

A CR refers to architectures that are capable of systematically capturing, storing, analyzing, and reporting data on patients with cancer (Santos 1999). Data on cancer occurrence and properties is registered in CRs. A cancer registrar is a person with over-arching responsibility to capture complete, accurate, and timely data on cancer patients. Participation in cancer registries is typically, but not solely, mandated by local legislations.

Data within a CR typically consists of demographic information, medical history, diagnostic investigations, outcomes of therapy, and some follow-up details of patients. CRs may have a variety of functionalities. For example, data within a CR can be used to

- a. Assess treatment outcomes (e.g., survival and toxicity)
- b. Assess disease survivorship aspects such as quality of life
- c. Evaluate efficacy and/or economic impacts of diseases and their interventions
- d. Provide follow-up investigation and guidance
- e. Allocate health resources at regional or country level
- f. Compare quality of treatment and outcomes between care providers
- g. Report on cancer incidence.

11.2.1 DIFFERENT TYPES OF CANCER REGISTRY

A CR may have different specific purposes and functionalities as described above, which primarily depend on local demands and requirements. Two primary types of CR are commonly used worldwide: hospital-based and population-based CRs. Table 11.1 shows a brief description of these two types of CR.

Table 11.1 Brief description of each type of CR

TYPES	HOSPITAL –BASED CR	POPULATION-BASED CR
Purposes	<ul style="list-style-type: none"> • Improvement of cancer care • Administration of cancer information • Clinical research • Training and education 	<ul style="list-style-type: none"> • Cancer prevention and early detection • Determine cancer incidence and trends • Academic research • Assessment of population cancer outcomes
Details	<ul style="list-style-type: none"> • Maintains data on all cancer patients diagnosed and/or treated at a particular hospital • Provides medical audit-style assessment of outcomes within a particular hospital • Support institutional registries with common standard protocols and integrated data 	<ul style="list-style-type: none"> • Records all new cases in a well defined population (e.g., geographic area) with an emphasis on epidemiology and public healthcare • Informs cancer agencies and organizations of cancer statistics in specific populations • Informs cancer researchers about an unbiased group of cases that can be selected for studies

11.2.1.1 Hospital-based CRs

Hospital-based CRs are established to record the information of cancer patients collected within a specific care setting, such as a major hospital or cancer clinic that aims to offer readily accessible, complete, accurate, and timely information of patients with cancer (Young 1991). This may include, for example, data about diagnosis, treatments, and outcomes. Hospital-based CRs generate reports on the number of cancers observed within a particular hospital per year by site, sex, and age. These reports are very useful for clinical research by comparing the frequency of certain types of cancer within a single hospital to the total number of cancer cases (Santos 1999). Furthermore, it can lead to several potential applications for epidemiological research by (1) providing information on approaches of diagnosis, stage distribution, outcomes to treatment, and overall survival at the hospital level and identifying potential drawbacks of treatments; and (2) predicting future needs for services, equipment, and human resources within a particular cancer center.

Generally, the endpoints of a hospital-based CR are geared toward quality, management, and caseload planning goals. However, many hospital-based CRs are required to submit data to a centralized disease-specific CR as well. For this purpose, the hospital-based CR usually has to collect data elements that are useful for the central CR but may be not immediately useful for the hospital. Therefore, data elements within a hospital-based CR often cover a wider range than data within a population-based CR (Young 1991).

11.2.1.2 Population-based CRs

Population-based CRs are concerned with collecting data on all new patients with cancer arising in a well-defined community (Santos 1999; Sadjadi et al. 2003; Parkin 2006; Coleman et al. 2011). The primary purpose of population-based CRs is to report cancer incidence and produce analytic findings about cancer in a defined population. In addition, population-based CRs are benefit of assessment and control of cancer care. Thus, the focus of population-based CRs is on epidemiology research and public health care.

AQ 2

AQ 3

Furthermore, population-based CRs can monitor the occurrence of cancer and prevalence of diseases; thus, they are of importance in planning and evaluating region-based or population-based programs on cancer control through

1. Standardizing treatment priorities and predicting resources needed in the future
2. Examining the effectiveness and appropriateness of screening programs in the community
3. Comparing health care providers in terms of practice and quality
4. Evaluating cancer care population outcomes through survival statistics.

Because it is not always possible to strictly define a catchment population, a hospital-based CR is not necessarily able to provide assessment and statistics on the cancer occurrence in the defined population (Young 1991), which is a major difference between hospital- and population-based CRs.

11.2.2 EXAMPLES OF CANCER REGISTRIES

There are many running CRs in the world. In this section, we will describe a small collection of them, which are from three countries: the United States, Italy, and the Netherlands.

11.2.2.1 Surveillance, epidemiology, and end results

The Surveillance, Epidemiology, and End Results (SEER) program is a source of information on cancer incidence and survival in the United States. The data is published on its Web site (<https://seer.cancer.gov/about/overview.html>). SEER currently captures and publishes cancer data on incidence and survival from the cancer data sources that cover around 28% of the American population.

The data captured by the SEER Program comprises demographics information, main tumor site, tumor morphology and cancer stage, treatment at first course, and follow-up details. The SEER program is the only program that collects the data on cancer stage at diagnosis and on survival data of cancer patients in the United States. The data is updated annually and published in reports as a public service. The U.S. Census Bureau provides periodic data on the population for the SEER program to use to compute the cancer ratio. Many practitioners and members of the public have used the data reported by the SEER program.

11.2.2.2 National radiation oncology registry

The National Radiation Oncology Registry (NROR; <http://www.roinstitute.org/What-We-Do/NROR/Index.aspx>) is a collaborative initiative of the Radiation Oncology Institute (ROI) and the American

Society of Radiation Oncology (ASTRO) through guidance and data support from other major stakeholders in oncology. The NROR captures related data on cancer patients' treatment delivery and outcomes; the data are used to improve cancer care. The overarching purposes of the NROR are to (1) compare cancer patients who have similar cancer state or profiles, (2) identify suitable treatment and possible drawbacks in cancer care, and (3) build a CR for health care in a defined population.

The national registry comprises standardized aggregated data on therapies for specific types of cancers. An analysis of the outcomes achieved would yield invaluable benchmarking measures, help define best practices, evaluate the comparative effectiveness of treatments, and identify gaps in quality. However, due to funding limitations, the NROR is no longer active.

11.2.2.3 Registro tumori ospedaliero (RTO)

AQ 4

The Italian Association of Cancer Registries (AIRTUM), established in Florence, Italy, in 1997, aims to coordinate multiple cancer registries in Italy. For more details, see <http://www.registri-tumori.it/cms/en>. The linkage created by the association supports research, editorial output, and methodological development of the various member registries. The association is connected to equivalent bodies in other countries at European and global levels. Statistics about the distribution of cancer in the areas covered by the member registries covers

- a. Incidence—the number of new cancer cases per year
- b. Prevalence—the number of Italians who have a particular cancer
- c. Mortality—the various different causes of death for Italians registered in the CR
- d. Trends—whether the number of cancer cases has been increasing or decreasing with respect to preceding years
- e. Survival—how long Italians survive after treatments for cancer
- f. Comparison of registries—whether the impacts of cancer are uniform across Italy
- g. International comparisons—how the situation in Italy compare with the rest of the world

11.2.2.4 Dutch surgical colorectal audit

The Dutch Surgical Colorectal Audit (DSCA) (Van Leersum et al. 2013) was established by the Association of Surgeons of the Netherlands (ASN) in 2009 and aims at surveillance, assessment, and improvement of colorectal cancer treatment. For more details, see <https://www.dica.nl/dlca>.

All Dutch hospitals that perform bowel cancer and rectal cancer surgery, participate in the Web-based quality registration. At present, more than 60,000 treatments have been registered. This registration allows quality benchmarking, where hospitals compare the quality of their cancer care with those of others. The comparisons are statistically corrected for differences in level of care (local case-mix) and random sampling, which renders the analysis meaningful and “fair.” The system has been able to propose possible improvements to individual colorectal surgeons, while the national professional association facilitates and monitors these improvements.

One of the main contributions of DSCA is that it leads to effective multicenter surgical collaboration. Because the ASN has an important role in audits, all colorectal surgeons in the Netherlands have participated in the collaboration.

11.2.3 DATA SOURCES OF CRs

Data sources of CRs are external data resources available to the registry that are used for the collection and verification of cancer-related information. According to the relation between the data elements and the goals of CRs, there are two types of data sources: primary and secondary data sources (Gliklich et al. 2014). Primary data sources refer to the data collected for the immediate goals of the CR. The data collection from primary data sources can promote elements of data quality such as completeness, validity, and reliability. This data collection is implemented via a standard protocol. The protocol is intended to enforce the same procedures and data format used in all CRs and patients, which ultimately benefits data analysis, tracking, and integration. Due to the auditability of collected data, the entered data can typically be traced back to an individual patient. Finally, the quality of primary data sources is usually better than secondary data sources because of automatic quality control procedures or follow-up checks made by data managers.

The initial purposes of secondary data sources are not for cancer registration (e.g., data generated in routine medical practice and insurance claim forms). The data in secondary data sources is usually stored electronically and can be accessed through appropriate permissions (Gliklich et al. 2014).

To ensure that few cancer cases are missed and that the quality of data (i.e., dimensionality, completeness, accuracy, timeliness) remains high, CRs usually collect the data of patients with cancer from multiple sources. For instance, data sources of a population-based CR often refer to cancer centers, general practitioners, screening programs, coroners' recording systems, health insurance companies, and other CRs. However, as a disadvantage, the use of multiple sources of information raises concerns about receiving multiple notifications of the same cancer patient. To avoid this problem, the data of the same patient existing within multiple data sources should be linkable, thereby eliminating duplicate registry.

It might be generally presumed that it is simple to maintain a population-based CR when sub-registries (e.g., hospital-based CRs) can openly transfer identifiable data such that the population-based CR only needs to integrate the incoming data. However, this is not always possible in real-world scenarios. In practice, the population-based CR still needs to collect overlapping data from numerous data sources. The reasons are twofold. First, patients with an eligible condition might never attend a contributing hospital, therefore the population-based CR needs to use multiple sources to prevent eligible cases being missed. Second, patients may attend more than one hospital over different parts of their treatment pathway. The use of multiple sources is then of benefit of identifying duplicate registrations or missing registrations of the same patient. Although it is not always possible to collect all data from all data sources in practice, the aim is still to use as many cancer data sources as possible. As described in (Gliklich et al. 2014), advantages and disadvantages of key sources are presented in Table 11.2.

Table 11.2 Advantages and disadvantages of key data sources

DATA SOURCE	ADVANTAGES	DISADVANTAGES
Patient report	<ul style="list-style-type: none"> • Unique perspective based on patient experience of disease and treatment. • Information on treatments not necessarily prescribed by clinicians. • Obtaining information about intended compliance. • Useful when timing of follow-up may not be concordant with timing of clinical encounter. • Can capture patient and/or caregiver outcomes. 	<ul style="list-style-type: none"> • Literacy, language, physician access or other barriers that may result in under-enrolment of some sub-populations. • Validated data collection equipment may need to be established. • Patients may refuse to participate in follow-up study. • Limited confidence on clinical information and utilization information.
Clinician report	<ul style="list-style-type: none"> • More specific information than available from coded data or medical record. • Tends to be more objective and focused on impacts on care delivery. 	<ul style="list-style-type: none"> • Clinicians are highly conscious of administrative burden. • Potential inconsistencies in capture of patient signs, symptoms, use of non-prescribed therapy.
Medical chart abstraction	<ul style="list-style-type: none"> • Information on routine medical care, with more clinical context than coded claims. • Potential for comprehensive view of patient medical and clinical history. 	<ul style="list-style-type: none"> • The underlying information is not always collected in a systematic way. For example, a diagnosis of bacterial pneumonia by one physician may be based on a physical exam and patient report of symptoms, while another physician may record the diagnosis only in the presence of a confirmed laboratory test.

(Continued)

Table 11.2 (Continued) Advantages and disadvantages of key data sources

DATA SOURCE	ADVANTAGES	DISADVANTAGES
	<ul style="list-style-type: none"> • Use of abstraction and strict coding standards (including handling missing data) increases the quality and interpretation of data abstracted. 	<ul style="list-style-type: none"> • It is difficult to interpret missing data. For example, absence of a specific symptom in the visit record would not indicate whether the symptom was truly absent or that the physician did not actively inquire about this specific symptom or set of symptoms. • Data abstraction is more (technical) resource intensive. • Complete medical and clinical history may not be available (e.g., new patient to clinic).
Electronic health records (EHRs)	<ul style="list-style-type: none"> • Information on routine medical care and practice, with more clinical context than coded claims. • Potential for comprehensive view of patient medical and clinical history. • Effective access to medical and clinical data. • Use of data transfer and coding standards (including handling of missing data) will increase the quality of data abstracted. 	<ul style="list-style-type: none"> • Underlying information from clinicians is not collected using uniform decision rules. (See example under “Medical chart abstraction.”) • Consistency of data quality and breadth of data collected varies across sites. • Difficult to handle information uploaded as image files into the EHRs (e.g., scanned clinician reports) vs. direct entry into data fields. • Historical data capture may require manual chart abstraction prior to implementation date of medical records system. • Complete medical and clinical history may not be available (e.g., new patient to clinic). • EHR systems vary widely. If data come from multiple systems, the registry should plan to work with each system individually to understand the requirements of the transfer.
Institutional databases	<ul style="list-style-type: none"> • Diagnostic and treatment information (e.g., pharmacy, laboratory, blood bank, radiology). • Resource utilization data (e.g., days in hospital). • May incorporate cost data (e.g., billed and/or paid amounts from insurance claims submissions). 	<ul style="list-style-type: none"> • Important to be knowledgeable about coding systems used in entering data into the original systems. • Institutional or organizational databases vary widely. The registry should plan to work with each system individually to understand the requirements of the transfer.

(Continued)

Table 11.2 (Continued) Advantages and disadvantages of key data sources

DATA SOURCE	ADVANTAGES	DISADVANTAGES
Administrative databases	<ul style="list-style-type: none"> • Useful for tracking health care resource utilization and cost-related information. • Range of data includes anything that is reimbursed by health insurance, generally including visits to physicians and allied health providers, most prescription drugs, many devices, hospitalization(s), if a lab test was performed, and in some cases, actual lab test results for selected tests (e.g., blood test results for cholesterol, diabetes). • In some cases, demographic information (e.g., gender, date of birth from billing files) can be uploaded. • Potential for efficient capture of large populations. 	<ul style="list-style-type: none"> • Represents clinical cost drivers vs. complete clinical diagnostic and treatment information. • Important to be knowledgeable about the process and standards used in claims submission. For example, only primary diagnosis may be coded and secondary diagnoses not captured. In other situations, value-laden claims may not be used (e.g., an event may be coded as a “nonspecific gynecologic infection” rather than a “sexually transmitted disease”). • Important to be knowledgeable about data handling and coding systems used when incorporating the claims data into the administrative systems. • Can be difficult to gain the cooperation of partner groups, particularly in regard to receiving the submissions in a timely manner.
Death indexes	<ul style="list-style-type: none"> • Completeness—death reporting is mandated by law in many countries, such as the United States. • Dependable alternative source for mortality tracking (e.g., if a patient was lost to follow-up). • National Death Index (NDI) — centralized database of death records from State vital statistics offices; database updated annually. • NDI causes of death relatively reliable (93–96%) compared with State death certificates. • Social Security Administration’s (SSA) Death Master File—database of deaths reported to SSA; database updated weekly. 	<ul style="list-style-type: none"> • Time delay—indexes depend on information from other data sources (e.g., State vital statistics offices), with delays of 12 to 18 months or longer (NDI). It is important to understand the frequency of updates of specific indexes that may be used. • Absence of information in death indexes does not necessarily indicate “alive” status at a given point in time. • Most data sources are country specific and thus do not include deaths that occurred outside of the country. • As of November 2011, Death Master File no longer includes protected State records.

(Continued)

Table 11.2 (Continued) Advantages and disadvantages of key data sources

DATA SOURCE	ADVANTAGES	DISADVANTAGES
Existing registries	<ul style="list-style-type: none"> • Can be merged with another data source to answer additional questions not considered in the original registry protocol or plan. • May include specific data not generally collected in routine medical practice. • Can provide historical comparison data. • Reduces data collection burden for sites, thereby encouraging participation 	<ul style="list-style-type: none"> • Important to understand the existing registry protocol or plan to evaluate data collected for element definitions, timing, and format, as it may not be possible to merge data unless many of these aspects are similar. • Creates a reliance on the other registry. Other registry may end. • Other registry may change data elements (which highlights the need for regular communication). • Some sites may not participate in both. Must rely on the data quality of the other registry.

11.2.4 RECEIVING AND REPORTING INFORMATION IN CRs

Because a CR is an organization for systematically processing the data of patients with cancer, it is very important to determine (1) the data items that a CR will receive from various data sources, and (2) the data items that a CR will report to other organizations. Received information by CRs refers to the information (e.g., demographic information, medical history, diagnostic discoveries, therapeutic information, follow-up details) collected by CRs from multiple sources or reported by certain institutions. On the other hand, reported information by CRs is simply concerned with summary analyses generated by the CRs, which can be deemed the functional output of the CRs. Because CRs can be different from each other in terms of aims, functions, local needs, and requirements, receiving and reporting information might also be different. The details of receiving and reporting information of CR are as follows.

11.2.4.1 Receiving information

The data elements collected by a CR are directly related to its aims and functionalities. The primary purpose of a hospital-based CR is administration of patients with cancer in a particular site. On the other hand, the main goal of a population-based CR is to produce statistics about cancer occurrence in a defined population. Thus, choosing data elements for a CR requires considering many aspects such as data reliability, the necessity in analyzing treatment responses, and even the cost of data collection (Gliklich et al. 2014).

11.2.4.1.1 Basic data elements

Although we have to determine the purposes and functionalities of a CR before specifying the data elements, some basic common elements exist in most CRs. The basic data elements that have been described in (Gliklich et al. 2014) are listed in Table 11.3, with the caveat that many directly identifying data elements will often be coded or partitioned in a secure part of the registry for privacy purposes.

11.2.4.1.2 Optional data items

Adding additional collection elements will increase the complexity and cost of registration (MacLennan 1991). Therefore, when designing registration forms and before performing registration, one should first

Table 11.3 Examples of possible basic data elements

Registrar information	<ul style="list-style-type: none"> • Registrar contact information • Contact information (e.g., address, telephone and email) of another individual who can be reached for follow-up
Patient information	<ul style="list-style-type: none"> • Patient identifiers (e.g., name, age, date of birth, place of birth, Social Security number) • Permission/consent • Source of enrollment (e.g., provider, institution, phone number, address, contact information) • Enrollment criteria • Sociodemographic characteristics, including race, gender, and age or date of birth • Education and/or economic status, insurance, etc. • Place of birth • Location of residence at enrollment • Source of information • Country, State, city, county, ZIP Code of residence.

consider whether a CR really needs certain data elements and whether it can sustain the cost associated with collecting these data element. Apart from the basic data elements, other data elements may be needed based on the specific design and purpose of a registry. Table 11.4 shows a collection of possible data elements that are described in (Gliklich et al. 2014).

Data element selection is primarily dependent on the goals of a CR, the approaches used for data collection, and the resources available. Many cancer CRs have failed because they attempted to capture too many data elements. The focus of the CR should be on the quality of data rather than the quantity. As we described in Section 2, those successful and productive CRs only collected a limited amount of information for each patient.

Table 11.4 Examples of optional data elements

PRE-ENROLMENT HISTORY	
Medical history	<ul style="list-style-type: none"> • Morbidities/conditions • Onset/duration • Severity • Treatment history • Medications • Adherence • Health care resource utilization • Diagnostic tests and results • Procedures and outcomes • Emergency room visits, hospitalizations (including length of stay), long-term care, or stays in skilled nursing facilities • Genetic information • Comorbidities
Environmental exposures	<ul style="list-style-type: none"> • Places of residence • Hazardous occupations? • Exposure to occupational hazards?
Patient characteristics	<ul style="list-style-type: none"> • Development (pediatric/adolescent) • Functional status (including ability to perform tasks related to daily living), quality of life, symptoms

(Continued)

Table 11.4 (Continued) Examples of optional data elements

PRE-ENROLMENT HISTORY	
	<ul style="list-style-type: none"> • Health behaviors (alcohol, tobacco use, physical activity, diet) • Social history • Marital status • Family history • Work history • Employment, industry, job category • Social support networks • Economic status, income, living situation • Sexual history • Foreign travel, citizenship • Legal characteristics (e.g., incarceration, legal status) • Reproductive history • Health literacy • Social environment (e.g., community services) • Enrollment in clinical trials (if patients enrolled in clinical trials are eligible for the registry)
Provider/system characteristics	<ul style="list-style-type: none"> • Geographical coverage • Access barriers • Quality improvement programs • Disease management, case management • Compliance programs • Information technology use (e.g., computerized physician order entry, e-prescribing, electronic medical records)
Follow-up/ Outcomes	<ul style="list-style-type: none"> • Safety: adverse events (see Chapter 12) • Quality measurement/improvement: key selected measures at appropriate intervals • Effectiveness and value: intermediate and endpoint outcomes; health case resource use and hospitalizations, diagnostic tests and results. Particularly important are outcomes meaningful to patients, including survival, symptoms, function, and patient-reported outcomes, such as health-related quality-of-life measures. • Natural history: progression of disease severity; use of health care services; diagnostic tests, procedures, and results; quality of life; mortality; cause/date of death • Economic status • Social functioning
Other potentially important information	<ul style="list-style-type: none"> • Changes in medical status • Changes in patient characteristics • Changes in provider characteristics • Changes in financial status • Residence • Changes to, additions to, or discontinuation of exposures (medications, environment, behaviors, procedures) • Changes in health insurance coverage • Sources of care (e.g., where hospitalized) • Changes in individual attitudes, behaviors

11.2.4.2 Reporting information

The most important purpose of a CR is to perform statistics on cancer occurrence, treatment and outcomes in a particular region or population (Powell 1991). Therefore, collation, examination, and explanation of the captured data are the main reporting tasks of a CR.

Information reported by CRs is often presented by means of cancer incidence reports, practice and treatment outcomes reports, and scientific publications. Results and conclusions are usually documented in reports and subsequently published to users. Generally, the reports contain background information about registration, procedures of registration, population of covering, data quality (e.g., completeness and validity), and results of analysis. The population-based CR should perform basic statistics that are primarily about the distribution of the tumor in the community. The data and findings may be displayed in various types of format such as tabular and graphical forms, by which the readers can draw their own conclusions according to their interests.

11.2.5 ASSESSMENT OF DATA QUALITY WITHIN CRs

CRs have evolved beyond a data provider that reports cancer incidence within a well-defined population (Parkin 2006). By linking sufficient resources, a CR is useful in many aspects of the cancer control domain, such as identification of causes of specific cancer, assessment of screening programs, and improvement of cancer care (Armstrong 1992; Parkin, 2008).

The functionalities of a modern CR and its capacity to perform cancer control activities are highly dependent on the quality of data within the CR. Three dimensions of data quality have been introduced in the earlier publication (Storm 1996): comparability, completeness, and validity. As described in (Bray and Parkin 2009), timeliness is another key indicator of data quality for CRs.

In order to assure data quality, quality control plays as an important role. Theoretically, it is possible that a CR can collect very high-quality data (similar to clinical trials) without extensive quality control processes, but this is seldom the case in real-world scenarios. There is no large-scale database that can be perfect in regard to completeness and validity. Therefore, routine quality control is a necessary step to identify the area needing improvement. The quality control can help with data interpretation and may further indicate a need for procedural changes (Navarro et al. 2010).

11.3 FROM CR TO BIG DATA IN RADIATION ONCOLOGY

11.3.1 BARRIERS OF CRs

We have seen that CRs can act as a valuable oncology data resource at several geographic scales and between multiple collaborating cancer centers. It remains a complex and costly process to maintain a CR, and the constraints on its architecture are obvious, that is, it is difficult to be scaled up to manage big data in radiation oncology. A centralized data repository needs to coordinate uniform collection at many different data sources and manage exchanges between different types of storage formats. This requires every contributing party to first agree on, and then rigidly adhere to, the same data collection instrument and the same internal data structure. The types of statistical analysis is constrained by the limited data fields collected and further restricted by the operational objective(s) of the CR. Although some attempts have been made by certain CRs to make data available for research, the elements generally have to be extracted as specialized queries by a data manager at the CR.

A large amount of human resources (e.g., registrars and data managers) and infrastructure (e.g. servers, user interfaces) is required to build and maintain a CR, adding significantly to its financial cost. Furthermore, rigidly structured collaborations among many departments are needed to support a CR. For example, a population-based CR requires first that local hospitals, cancer centers, and other institutions extract and upload (automatically or entering by hand) very specific data collection forms pertaining to the condition of interest. Then the collected data has to be processed and audited internally in the CR. Finally, with a significant time latency, results from a rigidly prescribed statistical analysis are reported to higher authorities and made available to the public.

Crucially, one of the most significant barriers to universal adoption of CRs pertains to how big data can be flexibly and securely shared among a large network of cancer institutes or research programs to answer

a broad range of clinically relevant questions. First, the data is neither readily findable nor discoverable; no general query process is available to physicians and researchers to ask if data pertaining to their clinical question reside in the CRs.

Second, data in CRs are generally inaccessible to physicians and researchers due to concerns over patient privacy, data sharing, or intellectual property rights. Where an entity outside the CR might be permitted to request some data, internal resources of the CR are required to program a data extraction query specific to each request received. For instance, researchers in a U.S. cancer center may be interested in the overall survival outcomes of lung cancer patients in a reasonably similar cohort in the Netherlands. Assuming the investigators in the United States even know what data fields to ask for, they would not be permitted to structure any external query that might address questions such as the following: (1) Is the case-mix of Dutch and U.S. populations approximately comparable? (2) What types of oncological and surgical interventions do Dutch lung cancer patients receive? (3) Do the follow-up protocols in both populations overlap to a sufficient degree to even contemplate a comparative study? In Dutch law, it is not allowed to share the data even for the purpose of clinical academic research. Patient privacy laws prevent effective data sharing, even in a highly geographically localized setting (e.g., two neighboring cancer hospitals in the Netherlands).

Third, even if the administrative, legal, and political barriers that prevent data sharing can be overcome, the technical challenge remains of reading data across incompatible computer systems that reside in many different formats. The exchange of “data dictionaries” between any two centers may permit a level of syntactic interoperability, but the parties also need to know the exact structure of how the data elements are actually stored in memory in order to construct a query. Exchange protocols such as Health Level-7 (HL7) and Fast Healthcare Interoperability Resources (FHIR) attempt to offer a degree of interoperability. However, each data transaction must be specifically customized to each collaboration site, rather than creating a single universal query that can work across all collaboration sites.

Therefore, one may not only consider a CR as a particular organization or entity (such as a hospital or a governance authority), but rather the CR must be considered as an architectural archetype for progressively more centralized data collection, storage, management, analysis, and dissemination. At each layer of centralization, the variety and contextual richness of the data is incrementally lost, such that only the broadest and least-detailed summary statistics can be reported at the uppermost level (for example, a national cancer registry). The reduction in dimensionality and contextual depth is compensated by an increasing universality in population coverage.

Such a trade-off in CR architecture may be visualized as in Figure 11.1. At the first level, individual institutions (e.g., clinics and hospitals) are the point of generation of data that lies closest to the patient. Data collection is generally immediate and highly multidimensional (e.g., clinician notes, nursing

AQ 5

Figure 11.1

observations, diagnostic imaging scans, treatment information, follow-up observations). The data is typically summarized and related to the parameters that may address governance, quality, and cost information. One level further, say at the level of a geographic region, each population-based CR may collect the data from every hospital in the region, but only for the conditions of specific interest to itself. The regional population-based CRs covers many more cancer cases and aggregates certain types of data elements, but this also results in increasing the fragmentation of clinical data as well as even further loss of dimensionality. Finally, population-based CRs typically summarize their data further to a global CR. Following the same process as the previous step, the global CR enables more universal coverage of cancer cases but with reduced data elements. The height of each layer represents the variety of the data (e.g., richness and dimensionalities). The bottom “layer” is thickest, indicating that it consists of the most varieties of data elements. In contrast, the upper “layer” is thinnest, indicating that it includes fewer dimensionalities of the data. Such data collection architecture can result in three serious issues:

1. The central CR cannot collect all the data, part of which are very important for modern cancer treatments and control (e.g., images data), from those data sources.
2. It usually spends much expense and takes a long time to report data to central CRs and collect data from data sources.
3. There is no complete network between different data sources, which seriously impedes linking, integrating, and exchanging data.

The volume of data in oncology is large and increasing rapidly, and it is already well beyond the processing capability of humans. Some companies (e.g., Google) are able to show that handling this volume of data is no longer a technical hurdle. A major problem is that most of the data in oncology is still unstructured, that means the clinical knowledge implicit in the data cannot be explicitly mined by the use of machines. Without major disruption to the way data is stored and the manner in which data is used to generate clinical knowledge, essential insights for personalized medicine and truly participative decision support systems (that involve predictive modeling of treatment effects) will remain beyond our reach.

11.3.2 BIG DATA IN RADIATION ONCOLOGY

With the advances in information technology, computing power, and digital medical equipment, the amount and types of data elements generated from just a single patient during routine cancer treatment are increasing in terms of observational, biological, genetic, imaging, and omics data. However, the data is usually collected at various points of care and then stored in varying formats inside separate databases. As discussed in the preceding sections, this presents significant challenges when trying to aggregate, analyze, and disseminate this data through traditional methods such as CRs. To employ these data to improve cancer treatment and health care of patients, an architecture specifically designed for big data is needed.

Big data research refers to the collection and analysis of a large volume of data elements and interrelationships that are difficult to discover through traditional methods. Big data approaches have been used in many areas of medicine, including comparison of innovative techniques (Chen 2014) and treatment modalities (Aneja and Yu 2014) in the field of radiation oncology.

A large volume of data has been generated within the radiation oncology field—mainly due to the frequent use of innovative techniques (e.g., medical imaging) in diagnostic and therapeutic procedures—during the routine practice of modern radiotherapy treatment (Lustberg et al. 2017). The data within radiation oncology displays all of the most significant hallmarks of big data, that is

1. the use of data-intensive imaging modalities (*volume*),
2. imaging generates a large amount of data per time interval (*velocity*),
3. there is an increasingly diverse spectrum of modalities available (*variety*), and
4. objectivity and quality of collected data vary greatly, which can influence accurate analysis (*veracity*).

As illustrated in Figure 11.2, a CR architecture is generally able to cover almost 100% of cancer cases (*volume*), but only about 3% of potentially prognostic factors are collected, along with a moderate data loss rate of approximately 20%. A different data-generating paradigm in oncology (i.e., clinical trials), typically enrolls around 3% all eligible cancer cases (Murthy et al. 2004; Movsas et al. 2007; Grand and O'Brien; 2012). However, due to stringent quality assurance and strict protocolization, nearly all of the factor of interest are recorded (*variety*) for only a relatively low (perhaps around, 5%) missing data rate.

Figure 11.2

What we conceive of as “big data” in oncology is represented by the entire graphic in Figure 11.2, where a vast volume and high throughput of data is continuously being generated by routine operations of modern cancer diagnosis and treatments. It is clear that neither CRs nor clinical studies adequately cover the full range of variety and volume of the clinical information available. However, as indicated in the figure, the consistency of data capture in the real world is generally low, thereby resulting in a high rate of missing data (perhaps around 80%). Further, unlike the rigid protocols required by clinical trials and registry submissions, interobserver biases and divergent interpretations are additional data quality issues. With the increasing use of automation and electronic data capture technology in oncology clinics, we expect that the potential bias and variation in the data (i.e., *veracity*) will continue to improve.

There is strong motivation within the oncology field to exploit hitherto underutilized real-world data. Improvements in treatment outcomes guided by big data utilization, in combination with registry studies and clinical trials, are widely seen as the most effective avenue of delivering value-based health care (Larson 2013; Murdoch and Detsky 2013; Khoury and Ioannidis 2014; Schneeweiss 2014). Big data therefore lies at the core of a personalized approach to medicine that leads to increased value of treatment for a given financial outlay. The question of how to use real-world big data to improve cancer control and reduce treatment-related toxicity has gained increasing traction over time. Closely related questions of interest concern improving data collection coverage in real clinical settings, imputation of missing data, and validation of outcome predictions in a wide variety of clinical settings.

As one of its primary objectives, big data research is expected to find the multiple clinical biological, and treatment variables that are related to treatment outcomes (e.g., overall survival, toxicity) (Rosenstein et al. 2016). This benefits the creation of better predictive models that promote the advances of personalized therapies for each individual patient (e.g., delivering more aggressive therapies where needed and less aggressive treatments when appropriate). In order to develop reliable and robust prognostic outcome models, data sharing and exchange is required between multiple institutions. The reasons are threefold:

1. Obviously, each institution has a limited capability of data collection. The amount of new patients who are diagnosed or treated in a clinic may range from hundreds to thousands per year. Thus, it is reasonable to estimate that the amount of cancer cases stored in-house for an individual clinic's past one or two decades within a clinic should be 20 to 200 thousands. Modern analytics methods (e.g., machine learning) are poised to satisfy the promise of identifying and guiding the response to variables influencing treatment outcomes of patients. However, the value of these analytics methods are highly dependent on the volume of data used for learning. As an illustrative example for size of training data, the hotly debated artificial intelligence (AI) computer program in 2016, AlphaGo

(designed by Alphabet Incorporated's Google DeepMind in London) was initially trained to mimic human gameplay from a historical games database containing approximately 30 million moves (Metz 2016). The data volume within an individual institution is often not sufficient to build such reliable and robust predictive outcome models through modern machine learning algorithms. Thus, data integration from multiple centers is necessary to develop realistic, understandable, and robust predictive outcome models.

2. Data collected within a particular hospital usually refer to local patients with some specific types of cancer. For example, the incidences of esophageal cancer vary widely among countries, with approximately half of all cases occurring in China. For other countries, it is difficult to collect sufficient data elements of esophageal cancer data for both volume and dimensionalities due to the very small number of patients with esophageal cancer. To generate prognostic outcome models, data exchange between two or more institutions seems a feasible and efficient approach.
3. As the predictive models are trained through local cohorts, it may be robust to predict the treatment outcome (e.g., overall survival) of unseen cancer cases within a local population. However, the predictive performance may be poor when applied to other populations, which impedes the usability and extension of the predictive model. Therefore, external validation is always necessary to measure the performance of a predictive model.

Because of the reasons explained above, data exchange between multiple clinics is a necessary procedure in radiation oncology field. Hence, there is a need to develop robust data exchange architectures instead of traditional methods to handle big data within the radiation oncology field. The future data architectures must be able to (1) scale to process ever-increasing amounts of data (*volume*); (2) have the throughput capacity to deal with high rates of data generation, in particular from imaging modalities (*velocity*); (3) process many different types of data into a form that is amenable to machine-based analysis (*variety*); and (4) intercept issues of data quality (e.g., bias, nonreproducibility and abnormality) (*veracity*).

11.4 BIG DATA EXCHANGE IN RADIATION ONCOLOGY

11.4.1 BIG DATA COLLECTION

Within radiation oncology, multiple types of data are routinely generated in the clinic from a variety of sources, which is the basis for big data research and provides an opportunity to improve cancer care.

One important source of big data is the patient demographics and clinical baseline factors obtained at the very beginning of the radiation oncology process. This information includes information about family history and personal health status. Crucially, this form of big data also consists of clinical observations and a baseline for treatment-related outcomes (especially comorbidities before treatment) by which the effectiveness of cancer interventions are evaluated.

Furthermore, an increasingly important source of big data, especially in regard to volume and variety, is the data stream produced by radiological and diagnostic imaging modalities. In radiation oncology, this routine data generation involves CT, PET, and MRI. The volume of image-based data is increasing rapidly due to the use of daily verification imaging with the patient lying in the intended treated position (e.g., for cone beam CT). It is no surprise that the largest (by volume) repositories of data in radiation oncology is the PACS.

Radiotherapy treatment planning is a highly sophisticated and computationally intensive process that generates big data in the form of organ delineations, beam geometry, radiation energy, collimation settings, and spatial dose distribution. This data generally resides within the radiotherapy Treatment Planning System (TPS). With the growing trend toward adaptive radiotherapy responding to real-time imaging at the point of treatment delivery, this volume of data is also set to grow rapidly.

Last, a rapidly developing data source is the result of digital pathology and high-throughput specimen analysis from medical laboratories. This includes genomics, proteomics, metabolomics, histologies, and hematologies.

Table 11.5 provides an overview of many of the possible radiotherapy research data types.

Table 11.5 Radiotherapy research data types

DATA TYPE	DATA EXAMPLES
Baseline clinical data	Demographics information, TNM-stage, date of diagnosis, histopathology
Diagnostic imaging data	Diagnostic CT, MR and PET imaging
Radiotherapy treatment planning data	Delineation sets, planning-CT, dose matrix, beam set-up, prescribed dose and fractions
Radiotherapy treatment delivery data	Cone beam CTs, orthogonal EPID imaging, delivered fractions
Non-radiotherapy treatment data	Surgery, chemotherapy
Outcome data	Survival, local control, distant failure, toxicity, quality of life
Follow-up imaging data	Follow-up CT, MR and PET imaging
Biological data	Sample storage, shipping, tracing and lab results
Additional study conduct data	Study design, protocol, eligibility criteria

Source: Skripcak, T. et al., *Radiother. Oncol.*, 113, 303–309, 2014.

11.4.2 STANDARD AND FRAMEWORK FOR DATA EXCHANGE

Efforts have been made to standardize data exchange between medical information archival systems with the DICOM standard and HL7 interoperability standard.

DICOM (Mildenberger et al. 2002) refers to a standard in medical imaging, which is supported by all imaging systems in the medical field and used widely. DICOM has two identities: a type of file format and a network communication protocol. First, medical image systems generate DICOM files containing patient information (e.g., name, identifier, sex, date of birth) and then acquires information of medical image systems and corresponding settings. The images are stored in DICOM files. Second, the DICOM protocol can be employed to exchange data (e.g., image or patient information) between different systems that are connected to the network within the hospital. DICOM has shown its ability to improve data exchange in medicine field. Radiotherapy data are commonly exchanged using a subset of DICOM often referred to as DICOM-RT.

HL7 (Dolin et al. 2001, 2006) refers to a widely accepted standard-setting organization that provides standards to define the protocol, language, and data type used for information communication among different systems. The most used version of HL7 is version 2 with which only a limited and not semantically rich data can be exchanged. HL7 version 3 had a much wider scope but is generally considered a failed standard due to its complexity and limited uptake. HL7 FHIR is the most recent standard and is receiving a lot of positive attention from the community and has resulted in real-world implementations by medical vendors.

11.4.3 DATA POOLING ARCHITECTURES

As described above, data integration is a necessary procedure of modern studies in the radiation oncology field. A data pooling architecture is used for data processing, storage, management, and exchange within an individual institution or between multiple institutions. In this section, we will describe three data pooling architectures (i.e., centralized, decentralized, and hybrid architecture), of which the infrastructures are shown in Figure 11.3.

11.4.3.1 Centralized architecture

A centralized architecture has complete physical control over the data pooled in a centralized repository. There is no direct real- or near real-time connection among participating institutions and operations (e.g., push/pull transactions and auditing occur in a central server). Although the architecture of the centralized model is simple, it can raise several issues including privacy and anonymization, duplication of data, mapping the local data to the central data model (usually resulting in manual data entry/copying),

Figure 11.3

and intellectual property (IP) rights (Skripcak et al. 2014). The most significant advantage of the centralized model is that the data is stored in a centralized repository, which makes data management and access easy.

The CR architecture described in Section 2 can be considered as an instance of the centralized model. A centralized CR collects data of patients with cancer from different sub-data-sources (e.g., hospitals, cancer centers, other institutions) through regular standards. Afterward, the centralized CR reports the analysis based on these data to the public and local government. However, an institution is not usually allowed to access the data within another institution due to data privacy of patients or local laws. It means there is usually no direct connection between these data sources.

11.4.3.2 Decentralized architecture

A decentralized architecture enables data exchange to occur among multiple institutions without any mediators, which is project-based via direct communication (Skripcak et al. 2014). However, the infrastructure required to enable data exchange must first be established at each site and comply with a standard exchange protocol. Shared data may be persistent (i.e., stored after exchange) or volatile (i.e., nothing is stored after exchange).

11.4.3.3 Hybrid architecture

A hybrid architecture attempts to combine the strengths of centralized and decentralized architectures. Data is transferred through direct communication across multiple sites. The difference between the hybrid and decentralized architectures is that the information on infrastructure, data representation format, controlled terminologies, and other required metadata are stored in a central server, making the

maintenance and modification of data exchange settings easier. Another advantage of a hybrid architecture is that big data generated in a local context is conceptually centralized but stored locally (i.e., inside the hospital or clinic). This hides the local complexities that differ for every clinic, below the level of multicenter sharing. This can happen using the agreed-upon terminology. We prefer decentralized and hybrid architectures for data exchange across multiple centers based on the present IT systems within hospitals (Skripcak et al. 2014).

11.4.4 DATA INTEROPERABILITY

Data interoperability has an important role in data exchange among multicenters: It is concerned with the capacity of a system to read and understand data transferred from another system. To implement complex and comprehensive data analyses, the data sources are required to be made fully interoperable across multiple information technology (IT) systems. Data interoperability consists of two main subprinciples: (1) to enable data exchange between multiple institutions, all institutions need to have syntactic interoperability in reference to establishing uniform data formats and exchange protocols. In other words, data representation for writing and reading information should be identical among all institutions, (2) syntactic and semantic interoperability should be in place, as described by Valentini et al. (2012). The aim of data semantic interoperability is to make data reliable and understandable by machines (Skripcak et al. 2014).

In a real-world scenario, it is difficult to achieve data interoperability because of privacy of patient data, local policy, and even technical issues. To enable data interoperability among different IT platforms, certain technologies (e.g., Semantic Web and ontology) have been applied to big data exchange in the radiation oncology field.

11.4.4.1 Semantic web

The Semantic Web (also known as “Linked Data”) is an extension of the Web via many standards by the World Wide Web Consortium (W3C). The standards boost the development of data formats and communication protocols on the Web. Among the various data formats in the Semantic Web, Resource Description Framework (RDF) is the most fundamental format and is commonly used.

Figure 11.4

The rationale behind the RDF data model is that any arbitrary statement about resources within the web can be represented by a simple triple (i.e., subject, predicate, and object). Any levels of complexity in the descriptions of resources are possible using multiple lines of triples. The subject and object here can be considered as two resources. The predicate is the property of the subject and represents the relation between the subject and object. For example, a patient's survival age, biological sex, and type of carcinoma can be described in the RDF format. Figure 11.4 shows the virtual representation of this ontology.

Storing data in the Semantic Web-based triple store is of great advantage for data access, usage, and exchange when compared with a relational database (e.g., SQL). A situation that arises often is data exchange between two hospitals (A and B), which use different relational databases (i.e., the name of each column may be different as well as the internal linkages within the respective databases). Given that each database likely comprises many thousands of rows, the problem of integrating these two relational tables is a serious problem and requires both parties to have in-depth knowledge of the other's relational data structure. Querying a nonexistent field in the other party's database may cause the query to crash. Therefore, one cannot add, delete, or otherwise move records around without informing each other. The need for such knowledge necessarily precludes the ability to preserve privacy and data confidentiality. In contrast, it is trivially simple to integrate data that are stored in triple stores. Because everything is stored as statements with only three pieces of information per statement, a single additional line with a reference from one database into a unique resource identifier in the other is typically all that is needed. In triple stores, there is no constraint that a data structure must be known in advance for a query to work (the query returns null instead of crashing the system), and therefore lines can be added, removed, or ordered in any fashion without affecting the query.

Consider, as an example, that the data on a patient is stored in a relational table called "Patient_1" with several columns and that all the radiomics features of the same patient are stored in another separate table called "Radiomics_Patient_1." Linking the two tables is often not an easy task, which means that we have to add new columns of each radiomics feature in "Patient_1" table and enter the value. However, if these the data of patient and his/her radiomics features are stored in separate RDF triple stores, data integration is trivial. To link the two triple stores, we just need one single sentence to define the triple: "Patient_1 (subject) has Radiomics Feature (predicate) Radiomics_Patient_1 (object)" via unique resource identifiers (URI) that represent patient_1 and his/her radiomics features. Now, the two databases are linked in the entirety of their data elements. For this triple, the class entities are defined in the Radiation Oncology Ontology (ROO) and the Radiomics Ontology (RO).

Based on current developments in technology, the Semantic Web is a more feasible and flexible choice for data representation in the radiation oncology field. As the data is represented through ontologies, the Semantic Web enables seamless linkage of data that are stored in different data platforms. Another benefit of the Semantic Web is as that searching and accessing data can be done through web technologies that are known to be extremely scalable. Recently HL7 FHIR has also been specified in the RDF format.

11.4.4.2 Ontology

An ontology refers to a terminology dictionary that defines the commonly used entities and relationships between entities in a particular domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. There are several ontologies available in radiation oncology field such as the ROO (<https://www.cancerdata.org/roo>), the National Cancer Institute Thesaurus (NCIT) ontology (Sioutos et al. 2007), and the International Classification of Diseases (ICD) ontology. When linking an entity to an ontology, the unique code in the corresponding ontology will be used to replace the value stored within the local database. The main purposes of using ontologies are as follows:

1. One of the primary purposes of using ontology is to share the common meaning of information among humans or machines (Musen 1992; Gruber 1993). For instance, the literal representation of a patient's biological sex may be "female" or "male." When linking to the NCIT ontology, codes C16576 and C20197 are used to represent "female" and "male," respectively. One of the advantages is that a machine only needs to recognize the ontology codes that represent the common concept (e.g., biological sex) and ignore the literal representations stored within local databases (e.g., "f/m", "female/male" and "0/1").

2. Knowledge defined in one domain can be reused in another domain through the use of ontologies. For instance, assume different domains all needed to represent the concept of treatment of oncology including radiotherapy, chemotherapy, chemo radiotherapy, and surgery. If some group built an ontology on treatment of oncology, this knowledge can be simply reused in the others' domains. In addition, it is common to reuse a general ontology (e.g., NCIT ontology).
3. The differentiation of the domain knowledge from the operational knowledge is another important purpose. As an example for radiotherapy, we can define as domain knowledge that a planning target volume is based on margins around the clinical target volume while still allowing local operational knowledge from a trial or institution to describe if and how the margins were applied.
4. If a declarative explanation of the term is available, it is possible to analyze domain knowledge. For example, a radiation ontology specifying that radiation on the chest may result in nonbacterial radiation pneumonitis might be used in another context to prescribe steroids instead of antibiotics in these patients.

11.4.5 DATA EXCHANGE

As mentioned above, different types of data have been routinely generated in a clinic from a variety of sources, which can be of benefit to cancer care. In order to ultimately reach FAIR data (Wilkinson et al. 2016), data exchange among different data sources is a necessary and important step. Therefore, architectures of data exchange should be built based on the rules such as efficiency, safety, and veracity. For the purpose of data exchange, two manners (i.e., manual and automatic data exchange) are used for both internal and external data exchange. This operation can be considered as “send data out,” which will be described in Section 4.5.1. However, “send data out” is not suitable or efficient for exchanging big data between multicenters in one or more countries because of reasons such as IP rights, local policy, and patient data privacy. Instead of sending data out, it is better to keep data inside hospitals and “send questions in.” This method is known as “distributed learning,” of which the details will be described in Section 4.5.2. Finally, the comparison between centralized and distributed multicenter architectures for big data exchange will be described in Section 4.5.3.

11.4.5.1 Send data out

11.4.5.1.1 Internal exchange

In many situations, text data of patients with cancer are generated in a hospital, such as (1) personal information forms on demographic information and medical histories filled by patients, (2) diagnosis and treatment notes created by doctors, and (3) follow-up details. These data can be in free text or structured questionnaires, which are entered in departmental (e.g., Oncology Information System) and/or hospital systems (e.g., electronic health record [EHR] system) for storage, management, and analysis. The data are exchanged, usually through the EHR platform, from one department to another within a hospital.

In addition, some types of data generated in the process of radiotherapy can be transferred to computers automatically, including imaging scans (e.g., CT, PET, MRI), tumor delineations, and treatment plans. One of the common properties of these data is that they are generated by electronic equipment (e.g., scanners and computers). The medical imaging data is usually transferred using the DICOM protocol to a central imaging archive (e.g., PACS). Another type of data that is often digitally exchanged is laboratory results.

11.4.5.1.2 External exchange

For data exchange among multiple centers, one commonly used approach is based on multicenter clinical trials. Data are often transferred via mail, fax, and email between two or more sites or to a central location. Although email is popular and convenient for receiving and sending information, it may result in three serious issues when it involves medical data exchange: (1) missing data and data security, (2) there are no standards for data exchange via email, and (3) transfer efficiency will be very low for a large volume of data (e.g., image data).

Another approach of external data exchange is through a Web-based application. For example, Openclinica (<https://www.openclinica.com>) has been developed for clinical data federation. Indeed, the use of Web-based products for data exchange have some benefits such as reliability and flexibility, although their use still can result in privacy-related issues.

Unlike in clinical research, data exchange for health care is not well developed and regular mail and faxes are still often the main manners of information exchange. Hence, there is a need for a more efficient and powerful approach for big data exchange among multiple centers.

11.4.5.2 Send questions in

As described above, data sharing between multiple centers often raises privacy-related issues. A better manner is to keep data in hospitals and “send questions in” (Lindell and Pinkas 2000; Wiessler 2013; Damiani et al. 2015) rather than sharing data. The primary goal of data sharing is to mine knowledge from others’ data. If there is an approach that can answer the research questions without allowing data outside of the hospitals, sending data out is unnecessary. The distributed solution allows advanced data analysis (e.g., knowledge sharing). Mathematical models are trained on local databases and shared with other hospitals. Because models contain only the “answer” to the question while the research data are kept within the hospital, using a mathematical model avoids privacy-related issues. Only some aggregate parameters are transferred between multicenters to reach the global convergence (consensus) of the mathematical model. This approach is known as “distributed learning” (Lambin et al. 2013a).

In addition, this distributed learning can be implemented on a Web-based learning environment (e.g., Varian Learning Portal). The learning platform can be considered as the master that merges knowledge models learned from different participating sites and continuously updates the model when more data are available (Skripcak et al. 2014).

11.4.5.3 Centralized versus distributed learning architecture

Modern medical research has to process an increasing number of data generated from many fields such as medical imaging, genomic, and proteomics. However, the reality is that an individual hospital only has data on a limited number of patients, which may be not sufficient to medical research. From the experience of machine learning in other fields, we know one needs a sufficient number of events to build a reliable predictive model for cancer treatments. In general, the more data collected from different sources, the more robust a predictive model is. Thus, cooperation between two or more hospitals is needed to collect more data regarding patients with cancer. The architectures of centralized and distributed learning among multiple centers has been described by van Soest et al. (2015).

AQ 10

Figure 11.5 shows the general overview for the centralized multicenter architecture. This approach allows participating sites to build the institutional architectures based on local policies. In addition to the entry points of all institutions, there are two key components within in a centralized learning architecture: (1) a central machine learning server is the place where learning occurs; and (2) a central collection point is responsible to perform the horizontal accumulation of data between all sites. As an example to explain the learning process, first, participating Site A sends an algorithm to the central machine learning server. Second, the central server implements calculation of this algorithm on the centralized data repository. Finally, the results are sent back to Site A after the calculation is completed.

Figure 11.5

Figure 11.6

On the other hand, the distributed learning architecture between multiple centers is different in terms of the places where computations happen (Meldolesi et al. 2014; Damiani et al. 2015; van Soest et al. 2015). Figure 11.6 displays the general overview of distributed learning architecture. We can see that the local unit has been added and the central federation point has been removed. In this architecture, the responsibility of the central machine learning server is only coordination. First, a site submits an algorithm or query to the central machine learning server. The algorithm or query is then split into several small sub-tasks. Second, the small sub-tasks are packed and sent to local machine learning units within each site. They will query the local data that is stored in RDF triple stores and the sub-algorithms are implemented in the local sites. The local application learns a model from local data. Third, the central machine learning server will merge all the results that have been computed on the local machine learning units of all sites (van Soest et al. 2015). Finally, if the preset criteria are met, a final model is generated. If not, the central machine learning server will send the models to all sites for re-learning until the preset convergence criteria have been reached.

The significant difference between distributed and centralized machine learning architecture is the transfer of data versus the transfer of model weights. When performing centralized learning, data leaves the hospital and is sent to the machine learning system. In contrast, data is kept within the hospital when we perform distributed learning. In this setting, the volume of data that is needed to be transferred is decreased compared with the centralized learning architecture; however, the transfer efficiency per task is increased. Boyd et al. (2011) and Wu et al. (2012) have given a complete explanation of how distributed machine learning algorithms work in their publications.

As proof that the concepts covered in Section 11.4 are indeed practical, a real-world machine learning project on distributed databases, known as Computer Assisted Theragnostics (CAT), has been proposed as illustrated schematically in Figure 11.7. First, data (e.g., image, genomics) are collected from a variety of data sources within each site and stored in local databases. Second, these data within each site are converted into the RDF data format and stored in the Semantic Web-based triple stores. Third, as shown in Figure 11.7, researchers in (for example) Rome can send research questions to (for example) Oxford via a global learning server. Then, learning happens on the local learning server in Oxford. After finishing the local learning, the results are sent back to the global learning server. Finally, researchers in Rome can query the global server for the answers to the research questions they proposed. This pipeline enable the communication between two or more institutions that have participated in the CAT project. Data exchange via the CAT architecture leads to a few benefits as:

Figure 11.7

- AQ 11** 1. Because the information shared in the CAT pipeline is the result (i.e., answers of research questions), not the real data, data exchange and knowledge mining occur without leaving data outside the local hospitals, avoiding data privacy-related issues.
2. All the data items stored in different databases within each hospital are standardized by the same ontology, leading to linked data.
3. In the example above, researchers in Rome only needed to send the algorithms or query (known as a “research question”), which is a small package (around hundreds of kilobytes), compared with sending the real data (a large volume).

AQ 12 A few projects (e.g., I2B2 [<https://www.i2b2.org>], EuroCAT [Deist et al. 2017], Validation of high TEchnology (VATE) [Meldolesi et al. 2014]) have demonstrated the feasibility of distributed multicenter machine learning. In addition, if we implement the procedures of distributed learning correctly, the results can be the same with as the centralized learning approach (Wiessler, 2013). Furthermore, it is possible to improve the robustness of prognostic models through validating on external data sets (Dekker et al. 2013), which has been demonstrated based on the results of the EuroCAT project (Deist et al. 2017).

11.5 BARRIERS OF BIG DATA EXCHANGE AMONG MULTICENTERS

This section will describe the barriers of big data exchange among multiple institutions in the radiation oncology community. The exchange barriers involves four primary aspects: administrative, ethical, political, and technical barriers. The details are as follows.

AQ 13 11.5.1 ADMINISTRATOR BARRIER

Two main administrative barriers impede data exchange among multiple institutions: data completeness and bias. First, it is usually not possible to collect every data element of an individual patient and not all

data elements are applicable, resulting in many missing data for the patient. This means there are no data to exchange. Second, system settings vary widely across sites. The bias in routine operation, protocols, and equipment settings can all result in differences between data across different sites. Data bias highly impedes data integration among multiple sites.

11.5.2 ETHICAL BARRIER

The ethical issues refer to data privacy and the reuse of research data. There may be a large difference in privacy explanation, application of confidentiality, and legislation between counties and even between ethical committees (Skripcak et al. 2014).

For both centralized and distributed multicenter infrastructures, privacy preservation is a major topic that must be considered. If correctly implemented as the applications described in preceding section, the distributed solution is generally more secure than the centralized solution because only a few parameters are transferred among multiple centers rather than the real data, although we cannot draw the conclusion that the issues on privacy preservation have been overcome via the distributed solution. Actually, there are no standard methods to solve privacy-related issues currently. Various stakeholders will always have to find a balance between the value of information and anonymity of participating patients.

AQ 14

11.5.3 POLITICAL BARRIER

In some scenarios, people do not want to share their data to others because of the issues related to the culture and local policy. Thus, there is a need for more high-quality published research articles that completely prove the benefits of data exchange (e.g., efficiency, robustness, security) to try to persuade data holders to participate in the collaborative research community and subsequently share their data.

11.5.4 TECHNICAL BARRIER

Even if these administrative, ethical, and political issues are solved, the technical barriers such as interoperability between clinical departments and lack of a uniform standard of data collection may still impede data exchange among multiple institutions. First, obtaining internal clinical IT systems interoperability is important for the generation of local anonymized data sets, which enables a universal access to integrate research data. These data are managed by an institutional data warehouse and can be findable through the corresponding semantic models (ontologies). However, it is often difficult to reach this interoperability in real-world scenarios because of the differences in their support of internationally standardized protocols, formats, and semantics. Although these problems can be solved, they often require an investment in resources that is not available in the operational setting.

AQ 15

Second, because the radiotherapy terminologies dictionary and ontologies are still under development, it is difficult to ensure that an element has a unique term and definition (Skripcak et al. 2014). As an example described above, different hospitals may use various representations to describe the biological sex of a patient such as “female and male,” “f and m,” or “0 and 1.” While performing data accumulation between two hospitals, two representations may exist in one aggregated data set. The best way to overcome this issue is to link clinical variables to ontologies that can provide the standard definitions of these variables. The biological sex of a patient “female” and “male” are represented by NCIT codes C16576 and C20197, respectively. Thus, the meaning of a clinical variable is only related to its ontological code rather than the literal representation.

AQ 16

11.6 CONCLUSION

In this chapter, we have seen that how different types of CR work as an organization for cancer data collection, management, storage, analysis, and exchange. However, the architecture of a CR is not robust and it is infeasible for it to handle the data in radiation filed currently, as the particular properties of big data (i.e., volume, velocity, variety, and veracity). In addition, it remains a complex and costly process to maintain a CR, which needs a large amount of human resources (e.g., registrars and managers) and infrastructure (servers and user interfaces).

AQ 17

One of the most important challenges to the universal adoption of CRs is how big data can be flexible and securely shared among a large network of cancer institutes or research programs to answer a broad range of clinically relevant questions. Thus, there is a need to build robust data exchange architectures to handle big data within the radiation oncology field instead of using traditional methods (e.g., CR). “Send data out” is the commonly used approach of data exchange via mail, fax, email, or Web-based applications although it may result in privacy-related data issues. The best manner to avoid privacy-related issues, while exploiting multicenter data, is to avoid sending data outside of the hospitals. This can be achieved through applying the method “send questions in,” which is also known as distributed learning. The distributed approach only transfers a subprocess machine learning algorithm to a specific hospital and sends the results back to the sender rather than transferring real data. It means that data/knowledge exchange occurs without allowing data outside of the hospital. Many collaborative projects (e.g., I2B2, EuroCAT, VATE) have demonstrated the feasibility of this distributed learning architecture in handling big data in radiation oncology field. If the procedures are implemented correctly, it can produce the same results as a centralized learning architecture.

REFERENCES

- Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, Bussink J et al. (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 5:4006.
- Aneja S, Yu JB (2014) Comparative effectiveness research in radiation oncology: Stereotactic radiosurgery, hypofractionation, and brachytherapy. *Semin Radiat Oncol* 24:35–42.
- Armstrong BK (1992) The role of the cancer registry in cancer control. *Cancer Causes Control* 3:569–579.
- Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning* 3:1–122.
- Bray F, Parkin DM (2009) Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *Eur J Cancer* 45:747–755.
- Chen AB (2014) Comparative effectiveness research in radiation oncology: Assessing technology. *Semin Radiat Oncol* 24:25–34.
- Chen M, Mao SW, Liu YH (2014) Big data: A survey. *Mobile Netw Appl* 19:171–209.
- Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U et al. (2011) Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): An analysis of population-based cancer registry data. *Lancet* 377:127–138.
- Damiani A, Vallati M, Gatta R, Dinapoli N, Jochems A, Deist T, Van Soest J, Dekker A, Valentini V (2015) Distributed learning to protect privacy in multi-centric clinical studies. In: *Conference on Artificial Intelligence in Medicine in Europe*, pp. 65–75: Springer.
- Dehing-Oberije C, Aerts H, Yu S, De Ruyscher D, Menheere P, Hilvo M, van der Weide H, Rao B, Lambin P (2011) Development and Validation of a Prognostic Model Using Blood Biomarker Information for Prediction of Survival of Non–Small-Cell Lung Cancer Patients Treated With Combined Chemotherapy and Radiation or Radiotherapy Alone (NCT00181519, NCT00573040, and NCT00572325). *International Journal of Radiation Oncology* Biology* Physics* 81:360–368.
- Dehing-Oberije C, De Ruyscher D, Petit S, Van Meerbeeck J, Vandecasteele K, De Neve W, Dingemans AM et al. (2010) Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. *Radiother Oncol* 97:455–461.
- Dehing-Oberije C, Yu S, De Ruyscher D, Meersschout S, Van Beek K, Lievens Y, Van Meerbeeck J et al. (2009) Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. *Int J Radiat Oncol Biol Phys* 74:355–362.
- Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, Eble M, Bulens P, Coucke P, Dries W (2017) Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: EuroCAT. *Clin Transl Radiat Oncol* 4:24–31.
- Dekker A, Nalbantov G, Oberije C, Wiessler W, Eble M, Dries W, January L, Bulens P, Krishnapuram B, Lambin P (2013) PD-0496: Multi-centric learning with a federated IT infrastructure: application to 2-year lung-cancer survival prediction. *Radiother Oncol* 106:S193–S194.
- Deng J (2014) Big data in radiation oncology: Challenges and opportunities. *Cancer Sci Res Open Access* 1:1–2.
- Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison JE (2001) The HL7 clinical document architecture. *J Am Med Inform Assoc* 8:552–569.
- Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A (2006) HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 13:30–39.

- Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, Cording E, Tomson D, Dodd C, Rollnick S, Edwards A, Barry M (2012) Shared decision making: A model for clinical practice. *J Gen Intern Med* 27:1361–1367.
- Gliklich RE, Dreyer NA, Leavy MB (2014) Data sources for registries. In: *Registries for Evaluating Patient Outcomes: A User's Guide* (Gliklich, R. E. et al., eds), pp. 127–144. Rockville, MD.
- Grand MM, O'Brien PC (2012) Obstacles to participation in randomised cancer clinical trials: A systematic review of the literature. *J Med Imaging Radiat Oncol* 56:31–39.
- Gruber TR (1993) A translation approach to portable ontology specifications. *Knowl Acquis* 5:199–220.
- Khoury MJ, Ioannidis JP (2014) Medicine. Big data meets public health. *Science* 346:1054–1055.
- Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CM, Carvalho S et al. (2013a) Rapid learning health care in oncology—An approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol* 109:159–164.
- Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, Roelofs E et al. (2013b) Predicting outcomes in radiation oncology-multifactorial decision support systems. *Nat Rev Clin Oncol* 10:27–40.
- Larson EB (2013) Building trust in the power of “big data” research to serve the public good. *JAMA* 309:2443–2444.
- Lindell Y, Pinkas B (2000) Privacy preserving data mining. In: *Annual International Cryptology Conference*, pp. 36–54. Springer.
- Lustberg T, van Soest J, Jochems A, Deist T, van Wijk Y, Walsh S, Lambin P, Dekker A (2017) Big data in radiation therapy: Challenges and opportunities. *Br J Radiol* 90:20160689.
- MacLennan R (1991) Items of patient information which may be collected by registries. In: *Cancer Registration: Principles and Methods* vol. 95 (Jensen, O. M., ed), pp. 43–63. IARC, Lyon, France.
- McNutt TR, Moore KL, Quon H (2016) Needs and challenges for big data in radiation oncology. *Int J Radiat Oncol Biol Phys* 95:909–915.
- Meldolesi E, van Soest J, Alitto AR, Autorino R, Dinapoli N, Dekker A, Gambacorta MA, Gatta R, Tagliaferri L, Damiani A (2014) VATE: VALidation of high TEchnology based on large database analysis by learning machine. *Colorectal Canc* 3(5):435–450.
- Metz C (2016) In major AI breakthrough, Google system secretly beats top player at the ancient game of go.
- Mildenberger P, Eichelberg M, Martin E (2002) Introduction to the DICOM standard. *Eur Radiol* 12:920–927.
- Movsas B, Moughan J, Owen J, Coia LR, Zelefsky MJ, Hanks G, Wilson JF (2007) Who enrolls onto clinical oncology trials? A radiation patterns of care study analysis. *Int J Radiat Oncol* 68:1145–1150.
- Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. *JAMA* 309:1351–1352.
- Murthy VH, Krumholz HM, Gross CP (2004) Participation in cancer clinical trials: Race-, sex-, and age-based disparities. *JAMA* 291:2720–2726.
- Musen MA (1992) Dimensions of knowledge sharing and reuse. *Comput Biomed Res* 25:435–467.
- Navarro C, Martos C, Ardanaz E, Galceran J, Izarzugaza I, Peris-Bonet R, Martinez C, Spanish Cancer Registries Working Group (2010) Population-based cancer registries in Spain and their role in cancer control. *Ann Oncol* 21(Suppl 3):iii3–iii13.
- Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, van Baardwijk A et al. (2014) A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: A step toward individualized care and shared decision making. *Radiother Oncol* 112:37–43.
- Parkin DM (2006) The evolution of the population-based cancer registry. *Nat Rev Cancer* 6:603–612.
- Parkin DM (2008) The role of cancer registries in cancer control. *Int J Clin Oncol* 13:102–111.
- Powell J (1991) Data sources and reporting. In: *Cancer Registration: Principles and Methods*, vol. 95 (Jensen, O. M., ed), pp. 29–42. IARC, Lyon, France.
- Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P (2013) Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol* 108:174–179.
- Rosenstein BS, Capala J, Efstathiou JA, Hammerbacher J, Kerns SL, Kong FM, Ostrer H et al. (2016) How will big data improve clinical and basic research in radiation therapy? *Int J Radiat Oncol* 95:895–904.
- Sadjadi A, Malekzadeh R, Derakhshan MH, Sepehr A, Nouraei M, Sotoudeh M, Yazdanbod A et al. (2003) Cancer occurrence in Ardabil: Results of a population-based cancer registry from Iran. *Int J Cancer* 107:113–118.
- Santos SI (1999) The role of cancer registries. In: *Cancer Epidemiology, Principles and Methods* (Santos, S. I., ed), pp. 385–403.
- Schneeweiss S (2014) Learning from big health care data. *N Engl J Med* 370:2161–2163.
- Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW (2007) NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 40:30–43.
- Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, Buttner D et al. (2014) Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasets. *Radiother Oncol* 113:303–309.

AQ 19

AQ 20

Stacey D, Legare F, Col NF, Bennett CL, Barry MJ, Eden KB, Holmes-Rovner M et al. (2014) Decision aids for people facing health treatment or screening decisions. *Cochrane Db Syst Rev* 1.

Storm HH (1996) Cancer registries in epidemiologic research. *Cancer Causes Control* 7:299–301.

AQ 21 Valentini V, Schmoll H-J, van de Velde CJ (2012) *Multidisciplinary Management of Rectal Cancer: Questions and Answers*. Springer Science & Business Media.

Van Leersum NJ, Snijders HS, Henneman D, Kolfschoten NE, Gooiker GA, ten Berge MG, Eddes EH et al. (2013) The Dutch surgical colorectal audit. *Eur J Surg Oncol* 39:1063–1070.

AQ 22 van Soest JP, Dekker AL, Roelofs E, Nalbantov G (2015) Application of machine learning for multicenter learning. In: *Machine Learning in Radiation Oncology*, pp. 71–97. Springer.

Wiessler W (2013) PO-0886: Privacy-preserving, multi-centric machine learning across hospitals and countries: Does it work? *Radiother Oncol* 106:343.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018.

Wu Y, Jiang X, Kim J, Ohno-Machado L (2012) Grid Binary LOGistic REgression (GLORE): Building shared models without sharing data. *J Am Med Inform Assn* 19:758–764.

AQ 23 Young JL (1991) The hospital-based cancer registry. In: *Cancer Registration: Principles and Methods*, vol. 95 (Jensen, O. M., ed), pp. 177–184: IARC, Lyon, France.

Author Query Sheet

Chapter No.: 11

Query No.	Queries	Response
AQ 1	Please check whether the expansions provided for “EMR” is correct	
AQ 2	Please revisit the sentence “In addition, population-based CRs are...” for clarity and correct if necessary.	
AQ 3	Please check whether the edits made to the sentence “Furthermore, population-based CRs....” retain the intended meaning.	
AQ 4	Please explain this heading “11.2.2.3 Registro Tumori Ospedaliero (RTO)”. The RTO is not mentioned elsewhere. Should the AIRTUM be named in the heading instead?	
AQ 5	Please provide captions for Figures 11.1 through 11.7.	
AQ 6	Please check whether the edits made to the sentence “It is no surprise....” retain the intended meaning.	
AQ 7	Please check whether the edits made to the sentence “Data interoperability....” retain the intended meaning.	
AQ 8	Please verify if the edits made in the sentence “In a real-world scenario, it is...” maintain the intended meaning.	
AQ 9	Please explain what is meant by “federation” in the sentence “For example, Openclinica...” Do you mean “accumulation”?	
AQ 10	Please verify edits to the sentence “In addition to the entry points...” maintain the intended meaning.	
AQ 11	Please verify edits to the sentence “Because the information...” retain the intended meaning.	
AQ 12	Please provide the expansion for “EuroCAT” in the sentence “A few projects...”	
AQ 13	Should the heading “ADMINISTRATOR BARRIER” be “ADMINISTRATIVE BARRIER”?	
AQ 14	Please verify edits to the sentence “If correctly implemented as the...” maintain the intended meaning.	
AQ 15	Please check whether the edits made to the sentence “Even if these....” retain the intended meaning.	
AQ 16	Please check whether the edits made to the sentence “While performing data....” retain the intended meaning.	
AQ 17	Please revisit the sentence “However, the...” for clarity and correct if necessary.	
AQ 18	Please verify edits to this sentence ““Send data out” is the commonly...” maintain the intended meaning.	
AQ 19	Please provide the publisher details for the ref “Gliklich (2014)”.	
AQ 20	Please provide the publisher details and publisher location for the ref “Santos (1999)”.	
AQ 21	Please provide the publisher location for the ref “Valentini et al. (2012)”.	
AQ 22	Please provide the editor details and publisher location for the ref “van Soest et al. (2015)”.	
AQ 23	Please define “ML” in Figures 11.5 and 11.6 captions. It is not abbreviated in the text.	