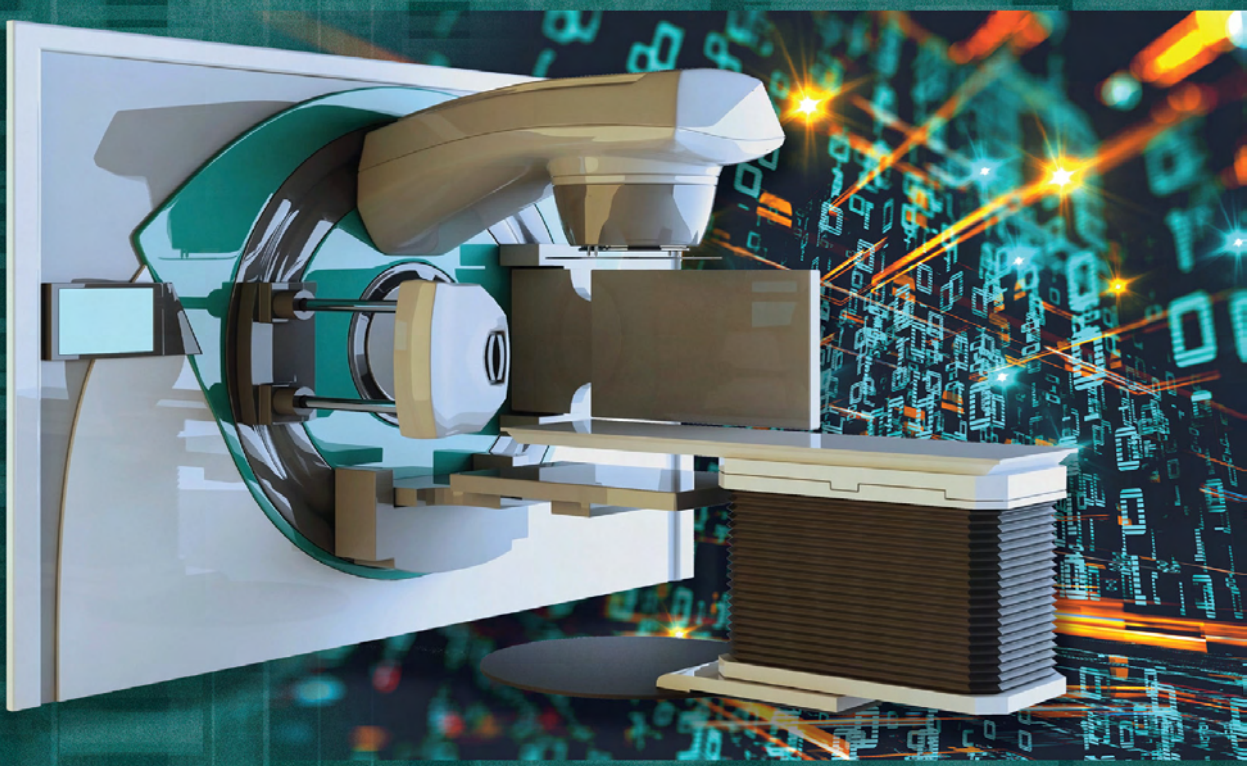
IMAGING IN MEDICAL DIAGNOSIS AND THERAPY

Andrew Karellas and Bruce R. Thomadsen, Series Editors

Big Data in Radiation Oncology



Edited by

Jun Deng

Lei Xing

CRC Press
Taylor & Francis Group

Big Data in Radiation Oncology

Imaging in Medical Diagnosis and Therapy

Series Editors
Andrew Karellas
Bruce R. Thomadsen

Stereotactic Radiosurgery and Stereotactic Body Radiation Therapy
Stanley H. Benedict, David J. Schlesinger, Steven J. Goetsch, Brian D. Kavanagh

Physics of PET and SPECT Imaging
Magnus Dahlbom

Tomosynthesis Imaging
Ingrid Reiser, Stephen Glick

Beam's Eye View Imaging in Radiation Oncology
Ross I. Berbeco, Ph.D.

Principles and Practice of Image-Guided Radiation Therapy of Lung Cancer
Jing Cai, Joe Y. Chang, Fang-Fang Yin

Radiochromic Film: Role and Applications in Radiation Dosimetry
Indra J. Das

Clinical 3D Dosimetry in Modern Radiation Therapy
Ben Mijnheer

Hybrid Imaging in Cardiovascular Medicine
Yi-Hwa Liu, Albert J. Sinusas

Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples
Dev P. Chakraborty

Ultrasound Imaging and Therapy
Aaron Fenster, James C. Lacefield

Dose, Benefit, and Risk in Medical Imaging
Lawrence T. Dauer, Bae P. Chu, Pat B. Zanzonico

Big Data in Radiation Oncology
Jun Deng, Lei Xing

For more information about this series, please visit:

https://www.crcpress.com/Series-in-Optics-and-Optoelectronics/book-series/TFOPTICSOPT

Big Data in Radiation Oncology

Edited by

Jun Deng

Lei Xing

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-1-138-63343-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Deng, Jun (Professor of therapeutic radiology), editor. | Xing, Lei, editor.
Title: Big data in radiation oncology / [edited by] Jun Deng, Lei Xing.
Other titles: Imaging in medical diagnosis and therapy ; 30.
Description: Boca Raton : Taylor & Francis, 2018. | Series: Imaging in medical diagnosis and therapy ; 30
Identifiers: LCCN 2018040966 | ISBN 9781138633438 (hardback : alk. paper)
Subjects: | MESH: Radiation Oncology | Data Mining--methods
Classification: LCC RC270.3.R33 | NLM WN 21 | DDC 616.99/40757--dc23
LC record available at https://lccn.loc.gov/2018040966

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To my wife, Jie, and my children, Daniel and Grace,
Thank you for your love, support, and inspiration.

Jun

In loving memory of my father who passed away from rectal cancer. His spirit lives with me.

Lei

Contents

Series preface

Since their inception over a century ago, advances in the science and technology of medical imaging and radiation therapy are more profound and rapid than ever before. Further, the disciplines are increasingly cross-linked as imaging methods become more widely used to plan, guide, monitor, and assess treatments in radiation therapy. Today, the technologies of medical imaging and radiation therapy are so complex and computer-driven that it is difficult for the people (physicians and technologists) responsible for their clinical use to know exactly what is happening at the point of care, when a patient is being examined or treated. The people best equipped to understand the technologies and their applications are medical physicists, and these individuals are assuming greater responsibilities in the clinical arena to ensure that what is intended for the patient is actually delivered in a safe and effective manner.

The growing responsibilities of medical physicists in the clinical arenas of medical imaging and radiation therapy are not without their challenges, however. Most medical physicists are knowledgeable in either radiation therapy or medical imaging, and expert in one or a small number of areas within their disciplines. They sustain their expertise in these areas by reading scientific articles and attending scientific talks at meetings. In contrast, their responsibilities increasingly extend beyond their specific areas of expertise. To meet these responsibilities, medical physicists periodically must refresh their knowledge of advances in medical imaging or radiation therapy, and they must be prepared to function at the intersection of these two fields. How to accomplish these objectives is a challenge.

At the 2007 annual meeting of the American Association of Physicists in Medicine in Minneapolis, this challenge was the topic of conversation during a lunch hosted by Taylor & Francis Publishers and involving a group of senior medical physicists (Arthur L. Boyer, Joseph O. Deasy, C.-M. Charlie Ma, Todd A. Pawlicki, Ervin B. Podgorsak, Elke Reitzel, Anthony B. Wolbarst, and Ellen D. Yorke). The conclusion of this discussion was that a book series should be launched under the Taylor & Francis banner, with each volume in the series addressing a rapidly advancing area of medical imaging or radiation therapy of importance to medical physicists. The aim would be for each volume to provide medical physicists with the information needed to understand technologies driving a rapid advance and their applications for safe and effective delivery of patient care.

Each volume in the series is edited by one or more individuals with recognized expertise in the technological area encompassed by the book. The editors are responsible for selecting the authors of individual chapters and ensuring that the chapters are comprehensive and intelligible to someone without such expertise. The enthusiasm of volume editors and chapter authors has been gratifying and reinforces the conclusion of the Minneapolis luncheon that this series of books addresses a major need of medical physicists.

This series "*Imaging in Medical Diagnosis and Therapy*" would not have been possible without the encouragement and support of the series manager, Lou Chosen, Executive Editor at Taylor & Francis. The editors and authors, and most of all I, are indebted to his steady guidance of the entire project.

William R. Hendee
Founding Series Editor

Preface

We initially discussed the possibility of publishing a book on big data in radiation oncology while organizing a symposium on the topic in the 2015 ASTRO annual meeting at San Antonio, Texas. After chatting with Lou Han of the Taylor & Francis Group, it became apparent that this was an undertaking that would benefit the community of radiation oncology and cancer research. We were thrilled to receive highly constructive and encouraging comments from two anonymous reviewers, to whom we are grateful, about our book proposal submitted to Taylor & Francis in late 2016. Here we are—in a period of just a little over two years, we were able to bring our idea to print.

The tremendous possibilities that big data can bring to cancer research and management have triggered a flood of activities in the development and clinical applications of the technology. Particularly, with the support of machine learning algorithms and accelerated computation, the field is taking off with tremendous momentum. We strongly believe that data science will dramatically change the landscape of cancer research and clinical practice in the near future.

This book is intended for radiation oncologists, radiation physicists, radiation dosimetrists, data scientists, biostatisticians, health practitioners, and government, insurance, and industrial stakeholders. The book is organized into four main groups: Basics, Techniques, Applications, and Outlooks. Some of the most basic principles and concepts of big data are introduced in the Basics. Following that, techniques used to process and analyze big data in radiation oncology are discussed in some details. Then some clinical applications of big data in radiation oncology are presented with great details. Finally, future perspectives and insights are offered into the use of big data in radiation oncology in terms of cancer prevention, detection, prognosis, and management.

Compared to a handful of similar books, the major features of this book include: (1) a comprehensive review of the clinical applications of big data in radiation oncology; (2) specially designed content for a wide range of readership; and (3) valuable insights into future prospects of big data in radiation oncology from experts in the field.

Being the first of its kind in this much talked-about topic, by no means did we set out to nor could we cover all the related topics in this book. However, we hope that this book will lay the foundations to many future works and hopefully inspire others to get involved in big data analytics.

Acknowledgments

It was rewarding to undertake this book project. When we began this project, we knew that it would take a huge amount of time and efforts to complete. However, we grossly underestimated the amount of support we would need from our colleagues in medical physics, radiation oncology, and beyond. It was the totality of our efforts and the support we received that made this book a reality.

While the ultimate responsibility for the content of this book is ours, we acknowledge with gratitude the generous help from the lead authors and co-authors of all the chapters, including Drs. Sepideh Almasi, Aditya P. Apte, Stanley Benedict, Jean-Emmanuel Bibault, Stephen Boyd, Cort Buchholz, Mildred Cho, Laurence E. Court, Joseph O. Deasy, Andre Dekker, Sunil W. Dutta, Brandon Dyer, Issam El Naqa, Eric Ford, Gregory R. Hart, Vojtech Huser, Aditi Iyer, Andrew Jackson, Jeho Jeong, Yu Jiang, Uwe Just, Alan Kalet, Mechthild Krause, Ying Liang, Yi Luo, Shuangge Ma, Dennis Mackin, Morteza Mardani, Charles S. Mayo, Wazir Muhammad, Bradley J. Nartowt, Jung Hun Oh, Gaurav Pandey, Mark Phillips, Guillem Pratx, Shyam Rao, David A. Roffman, Yi Rong, Barry S. Rosenstein, Ida Schönfeld, Chris Sherman, Zhenwei Shi, Timothy N. Showalter, Tomas Skripcak, Corey W. Speers, Maria Thor, Daniel M. Trifiletti, Esther G.C. Troost, Leonard Wee, Catharine M.L. West, Yong Yang, and Yinyi Ye.

We would also like to thank the people at the Taylor & Francis Group, particularly Lou Han, for continuous and prompt support during this long journey. We can't remember how many times we have approached Lou for thoughtful advice, suggestions, or last-minute help, for which we are deeply indebted. We are also very grateful to Angela Graven, our Project Manager from Lumina Datamatics, who efficiently managed the typesetting and proofreading of all the chapters in the book.

On a more personal level, we would like to thank our families for their gracious love, unwavering support and encouragement that empowered us to complete this book project.

Editors

Jun Deng, PhD, is a professor at the Department of Therapeutic Radiology of Yale University School of Medicine and an American Board of Radiology board-certified medical physicist at Yale New Haven Hospital. He obtained his PhD from the University of Virginia in 1998 and finished his postdoctoral fellowship at the Department of Radiation Oncology of Stanford University in 2001. Dr. Deng joined Yale University's Department of Therapeutic Radiology as a faculty physicist in 2001. He serves on the editorial boards of numerous peer-reviewed journals and has served on study sections of the NIH, DOD, ASTRO, and RSNA since 2005 and as a scientific reviewer for the European Science Foundation and the Dutch Cancer Society since 2015. He has received numerous honors and awards such as Fellow of Institute of Physics in 2004, AAPM Medical Physics Travel Grant in 2008, ASTRO IGRT Symposium Travel Grant in 2009, AAPM-IPEM Medical Physics Travel Grant in 2011, and Fellow of AAPM in 2013. At Yale, his research has focused on big data, machine learning, artificial intelligence, and medical imaging for early cancer detection and prevention. In 2013, his group developed CT Gently®, the world's first iPhone App that can be used to estimate organ doses and associated cancer risks from CT and CBCT scans. Recently, funded by an NIH R01 grant, his group has been developing a personal organ dose archive (PODA) system for personalized tracking of radiation doses in order to improve patient safety in radiation therapy.

Lei Xing, PhD, is the director of Medical Physics Division and the Jacob Haimson Professor of Medical Physics in the Departments of Radiation Oncology and Electrical Engineering (by courtesy) at Stanford University, Stanford, California. His research has been focused on artificial intelligence in medicine, biomedical data science, medical imaging, inverse treatment planning, image-guided interventions, nanomedicine, and molecular imaging. Dr. Xing is on the editorial boards of a number of journals in medical physics and imaging and is a recipient of numerous awards. He is a fellow of American Association of Physicists in Medicine (AAPM) and American Institute for Medical and Biological Engineering (AIMBE).

Contributors

Sepideh Almasi
Department of Radiation Oncology
Stanford University
Stanford, California

Aditya P. Apte
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Stanley Benedict
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Jean-Emmanuel Bibault
Radiation Oncology Department
Georges Pompidou European Hospital
Assistance Publique—Hôpitaux de Paris
and
INSERM UMR 1138 Team 22:
Information Sciences to support
Personalized Medicine
Paris Descartes University
Sorbonne Paris Cité
Paris, France

Stephen Boyd
Department of Electrical Engineering
Stanford University
Stanford, California

Cort Buchholz
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Mildred Cho
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Laurence E. Court
Department of Radiation Physics
The University of Texas MD Anderson Cancer
Center
Houston, Texas

Joseph O. Deasy
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Andre Dekker
Department of Radiation Oncology (MAASTRO
Clinic)
GROW—School for Oncology and Development
Biology
Maastricht University Medical Centre
Maastricht, the Netherlands

Jun Deng
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Sunil W. Dutta
Department of Radiation Oncology
University of Virginia
Charlottesville, Virginia

Brandon Dyer
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Eric Ford
Department of Radiation Oncology
University of Washington
Seattle, Washington

Gregory R. Hart
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Vojtech Huser
Laboratory of Informatics Development
NIH Clinical Center
Washington, District of Columbia

Aditi Iyer
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Andrew Jackson
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Jeho Jeong
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Yu Jiang
School of Public Health
University of Memphis
Memphis, Tennessee

Uwe Just
Department of Radiotherapy and Radiation Oncology
Technische Universität Dresden
Dresden, Germany

Alan Kalet
Department of Radiation Oncology
University of Washington
Seattle, Washington

Mechthild Krause
Department of Radiotherapy and Radiation Oncology
Technische Universität Dresden
Dresden, Germany

Ying Liang
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Yi Luo
Department of Radiation Oncology, Physics Division
University of Michigan
Ann Arbor, Michigan

Shuangge Ma
Department of Biostatistics
Yale University
New Haven, Connecticut

Dennis Mackin
Department of Radiation Physics
The University of Texas MD Anderson Cancer Center
Houston, Texas

Morteza Mardani
Department of Radiation Oncology
and
Department of Electrical Engineering
Stanford University
Stanford, California

Charles S. Mayo
Department of Radiation Oncology
University of Michigan
Ann Arbor, Michigan

Wazir Muhammad
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Issam El Naqa
Department of Radiation Oncology, Physics Division
University of Michigan
Ann Arbor, Michigan

Bradley J. Nartowt
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Jung Hun Oh
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Gaurav Pandey
Department of Genetics and Genomic Sciences
Icahn Institute of Genomics and Multiscale Biology
Icahn School of Medicine at Mount Sinai
New York City, New York

Mark Phillips
Department of Radiation Oncology
University of Washington
Seattle, Washington

Guillem Pratx
Department of Radiation Oncology
Stanford University
Stanford, California

Shyam Rao
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

David A. Roffman
Department of Therapeutic Radiology
Yale University School of Medicine
New Haven, Connecticut

Yi Rong
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Barry S. Rosenstein
Department of Radiation Oncology
and
Department of Genetics and Genomic Sciences
Icahn Institute of Genomics and Multiscale Biology
Icahn School of Medicine at Mount Sinai
New York City, New York

Ida Schönfeld
Department of Radiotherapy and Radiation Oncology
Technische Universität Dresden
Dresden, Germany

Chris Sherman
Department of Radiation Oncology
UC Davis Cancer Center
Sacramento, California

Zhenwei Shi
Department of Radiation Oncology (MAASTRO Clinic)
GROW—School for Oncology and Development Biology
Maastricht University Medical Centre
Maastricht, the Netherlands

Timothy N. Showalter
Department of Radiation Oncology
University of Virginia
Charlottesville, Virginia

Tomas Skripcak
Department of Radiotherapy and Radiation Oncology
Technische Universität Dresden
Dresden, Germany

Corey W. Speers
Department of Radiation Oncology
University of Michigan
Ann Arbor, Michigan

Maria Thor
Department of Medical Physics
Memorial Sloan Kettering Cancer Center
New York City, New York

Daniel M. Trifiletti
Department of Radiation Oncology
University of Virginia
Charlottesville, Virginia

Esther G.C. Troost
Department of Radiotherapy and Radiation Oncology
Technische Universität Dresden
Dresden, Germany

Leonard Wee
Department of Radiation Oncology (MAASTRO Clinic)
GROW—School for Oncology and Development Biology
Maastricht University Medical Centre
Maastricht, the Netherlands

Catharine M.L. West
Division of Cancer Sciences
The University of Manchester
Manchester Academic Health Science Centre
Christie Hospital
Manchester, United Kingdom

Lei Xing
Department of Radiation Oncology
and
Department of Electrical Engineering
Stanford University
Stanford, California

Yong Yang
Department of Radiation Oncology
Stanford University
Stanford, California

Yinyi Ye
Department of Electrical Engineering
and
Department of Management Science and Engineering
Stanford University
Stanford, California

1 Big data in radiation oncology: Opportunities and challenges

Jean-Emmanuel Bibault

Contents

The increasing number of clinical and biological parameters that need to be explored to achieve precision medicine makes it almost impossible to design dedicated trials.[1] New approaches are needed for all populations of patients. By 2020, a medical decision will rely on up to 10,000 parameters for a single patient,[2] but it is traditionally thought that our cognitive capacity can integrate only up to five factors in order to make a choice. Clinicians will need to combine clinical data, medical imaging, biology, and genomics to

achieve state-of-the-art radiotherapy. Although sequencing costs have significantly decreased,[3,4] we have seen the generalization of electronic health records (EHRs) and record-and-verify systems that generate a large amount of data.[5] Data science has an obvious role in the generation of models that could be created from large databases to predict outcome and guide treatments. A new paradigm of data-driven decision making: The reuse of routine health care data to provide decision support is emerging. To quote I. Kohane, "Clinical decision support algorithms will be derived entirely from data … The huge amount of data available will make it possible to draw inferences from observations that will not be encumbered by unknown confounding."[6]

Integrating such a large and heterogeneous amount of data is challenging. In this first chapter, we will introduce the concept of big data and the specificities of this approach in the medical field. We will show the opportunities of data science applied to radiation oncology as a tool for treatment planning and predictive modeling. We will also explain the main requirements for the implementation of a precision medicine program relying on big data.

1.1 WHAT IS BIG DATA?

This section defines big data and introduces a few key concepts the readers need to be familiarized with before they can proceed.

1.1.1 THE FOUR V'S OF BIG DATA

The fours Vs of big data are volume, variety, velocity, and veracity.[7] A comprehensive EHR for any cancer patient is around 8 GB, with genomic data being much larger than all other data combined (volume). Creating a predictive model in radiation oncology requires a significant heterogeneity in the data types that need to be included (variety). The use of big data for medical decision making requires fast data processing (velocity). As sequencing costs have significantly decreased[3,4,8] and computing power has steadily increased, the only factor preventing us from discovering factors influencing disease outcome is the lack of large phenotyped cohorts. The generalization of the use of EHRs gives us a unique opportunity to create adequate phenotypes (veracity).

1.1.2 THE SPECIFICITIES OF MEDICAL DATA

1.1.2.1 Data relevance

Lambin et al. have described in details the features that should be considered and integrated into a predictive model.[9] They include

- Clinical features: patient performance status, grade and stage of the tumor, blood tests results, and patient questionnaires.
- Treatment features: planned spatial and temporal dose distribution, associated chemotherapy. For this, data could be extracted directly from the record-and-verify software for analysis.
- Imaging features: tumor size and volume, metabolic uptake (more globally included into the study field of "radiomics").
- Molecular features: intrinsic radiosensitivity,[10] hypoxia,[11] proliferation, and normal tissue reaction.[12] Genomic studies play a key role in determining these characteristics.

1.1.2.2 Data granularity (Surveillance, Epidemiology and End Results database versus EHRs)

Big data in radiation oncology means studying large cohorts of patients and integrating heterogeneous types of data. Using these types of data through machine learning holds great promises for identifying patterns beyond human comprehension. Oncology is already moving away from therapies based on anatomical and histological features and focusing on molecular abnormalities that define new groups of patients and diseases. This evolution induces an increasingly complex and changing base of knowledge that ultimately will be not usable by physicians. The other consequence of this is that, as we individualize molecular traits, designing clinical trials will become more and more difficult to the point where it will become statistically impossible to achieve sufficient power. The financial and methodological burdens of designing these clinical trials will eventually become unsustainable. EHR use in most institutions

is an elegant and easy way to digitally capture large amounts of data on patient characteristics, treatment features, adverse events, and follow-up. This wealth of information should be used to generate new knowledge. The quality and nature of the data captured is important because poor data will generate poor results ("garbage in, garbage out") and big data should not be seen as a magical box able to answer any question with ease and trust. Clinical trials are designed to avoid confounding factors and gather detailed data that are not always available in EHRs.[13] Several Surveillance, Epidemiology and End Results (SEER) studies have generated fast results on important questions.[14–18] However, when studying radiation treatments, a major limitation of big data is the lack of detailed information on treatment characteristics. Integrating these features straight out of the record-and-verify systems will provide faithful dosimetric and temporal data. Several teams have already published studies using prediction to better adapt radiation treatments.[19–24] None of these approaches have reached clinical daily use. A simple, easy-to-use system would need to be directly implemented into the treatment planning system to provide decision support. The best achievable treatment plan based on a patient's profile would be given to the dosimetrist or physicist. The same system would be used to monitor patients during treatment and notify physicians whenever an adverse event outside of the predicted norm would happen. The data generated by each patient and treatment would be integrated into the model. We are, however, very far from this vision and in order to achieve it several methodological challenges will need to be addressed (e.g., how to capture core radiation oncology data into EHRs, integrate clinical, dosimetric, and biologic data into a single model and validate this model in a prospective cohort of patients).

1.1.2.3 Structured data

In the field of radiation oncology, medical data is already highly structured through the use of oncology information and record-and-verify systems. Data can be easily extracted with the precise features of treatment planning (dosimetry) and delivery; however, this data can have very heterogeneous labels that require time-consuming curation. This is particularly true for anatomical and target volumes labeling. Using routine radiation oncology data requires respecting a set of principles to make it more accessible. These principles, known as the Findable, Accessible, Interoperable, Re-Usable (FAIR) Data Principles,[25] initially developed for research data, are now being extended to clinical trials and routine care data. Data must be Findable, Accessible, Interoperable, and Reusable for research purposes. Behind Findable, Accessible, Interoperable, Re-Usable (FAIR) principles is the notion that algorithms may be used to search for relevant data, to analyze the data sets, and to mine the data for knowledge discovery. EHR data cannot be fully shared, but efforts can be made to make vocabularies and algorithms reusable and enable multi-site collaborations. To achieve that goal, the radiation oncology community must pave the road for semantic frameworks that the sources and the users could agree upon in the future. Besides usual quantitative data (e.g., dose), standard representation of anatomical regions and target volumes is required to study, for example, radiation complications. There are currently several domain-specific software packages for radiation oncology planning: Elekta (MOSAIQ©), Varian (ARIA©), Accuray (Multiplan© and Tomotherapy Data Management System©), and BrainLab (iPlan©). Each of these treatment planning and record-and-verify systems has its own anatomical structure labeling system, and these systems are not consistent across platforms, making it difficult to extract and analyze dosimetric data on a multicenter large scale. Using knowledge management with concept recognition, classification, and mapping, an accurate ontology that is dedicated to radiation oncology structures can be used to unify data in clinical data warehouses, thus facilitating data reuse and study replication in cancer centers.[26]

1.1.2.4 Unstructured data: The challenge of EHRs and the role of Natural Language Processing

Each physician has a specific way of reporting and writing medical notes. To leverage this kind of data, natural language processing (NLP) is required in order to make sense of stored files and extract meaningful data. NLP is a part of machine learning that can help in understanding, segmenting, parsing, or even translating text written in a natural language.[27] It can be used to repurpose electronic medical records (EMR) to automatically identify postoperative complications,[28] create a database from chest radiographic reports,[29] or even rapidly create a clinical summary from data collected for a patient's disease.[30] This kind of technology will be essential for big data analytics in radiation oncology, mostly for clinical information.

1.1.3 FROM BIG DATA AND DARK DATA TO SMART DATA

Radiation oncology is one of the most interesting fields of medicine for big data analytics because treatment planning and delivery data are very structured. However, this type of data is rarely used for analytics (dark data). Data integration approaches are necessary in order to effectively curate clinical unstructured data and this highly structured data. Collecting and repurposing these data into an automatic smart data system will be necessary before any medical use can be made.

1.2 OPPORTUNITIES OF BIG DATA IN RADIATION ONCOLOGY: DATA-DRIVEN DECISION MAKING

This part will highlight a few examples of the potential of big data applications in radiation oncology and cite the main studies that have already used data mining methodologies for technical or clinical questions.

1.2.1 ACCELERATING TREATMENT PLANNING

1.2.1.1 Contouring

The contouring of a large number of organs at risk before treatment planning is very time-consuming. Although manual segmentation is currently viewed as the gold standard, it is subject to interobserver variation and allows fatigability to come into play at the risk of lowering accuracy. A potential way to spare time would be automatic segmentation, with numerous industrial and homemade solutions being developed. Very few of them have been evaluated in clinical practice. Most of the existing solutions use atlases as a basis for automatic contouring. In 2016, DeepMind, a Google-owned startup, announced a project to use deep learning for automatic structures segmentation in head and neck cancer through a partnership with the National Health Service (NHS) in the United Kingdom.[31]

1.2.1.2 Dosimetry optimization

Machine learning has been used to predict radiation pneumonitis after conformal radiotherapy,[32] local control after lung stereotactic body radiation therapy (SBRT),[33] and chemoradiosensitivity in esophageal cancer.[34] In these studies, dose–volume histograms were used as predictive factors. They were also used to predict toxicity after radiotherapy for prostate cancer[35–37] and lung cancer.[38,39] Future treatment planning systems will need to directly integrate machine learning algorithms in order to automatically predict efficacy or toxicity to help the physician choose the optimal dosimetry.[19–24]

1.2.2 EVALUATING NEW TREATMENT TECHNIQUES

Big data studies can help in evaluating new treatment techniques. It is highly unlikely that we will see studies comparing three-dimensional (3D) conformal radiotherapy and intensity-modulated radiotherapy (IMRT). However, IMRT is now used in almost all contexts, even if it was only proven superior to 3D for head and neck cancer.[40] For future treatment technology improvements, big data studies could be used to generate hypothesis that will need to be ideally validated in a prospective trial.

1.2.3 PERSONALIZED RADIATION ONCOLOGY

1.2.3.1 Predicting disease progression and treatment response

Predictive modeling is a two-step process involving qualification followed by validation. Qualification consists of demonstrating that the data are indicative of an outcome. Once predictive or prognostic factors have been identified, they should be validated on a different data set. Once a model has been qualified and validated, further studies must be conducted in order to assess whether treatment decisions relying on the model actually improve the outcome of patients.

Kang et al. have proposed seven principles of modeling[41] in radiation oncology:

1. Consider both dosimetric and non-dosimetric predictors.
2. Manually curate predictors before automated analysis.
3. Select a method for automated predictor selection.
4. Consider how predictor multicollinearity is affecting the model.