Title

Image Guided Radiotherapy Safety: Automated Error Detection and Human Factors of Clinical Integration

Author

Petragallo, Rachel Marie

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Image Guided Radiotherapy Safety: Automated Error Detection and Human Factors of Clinical Integration

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Physics and Biology in Medicine

by

Rachel Marie Petragallo

2024

ABSTRACT OF THE DISSERTATION

Image Guided Radiotherapy Safety: Automated Error Detection and Human Factors of Clinical Integration

by

Rachel Marie Petragallo

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2024

Professor James Michael Lamb, Chair

The introduction of image guided radiation therapy (IGRT), defined as regular patient imaging during the course of a radiation therapy treatment regimen, has resulted in significant improvements in both the quality and safety of radiotherapy treatments. Accurate visualization of the target and nearby normal tissue has allowed for the reduction of planning target volume (PTV) margins, leading to increased normal tissue sparing. Daily image guidance can also increase the safety of radiation therapy treatments, as the patient is imaged prior to each treatment fraction to verify target location and compensate for inter-fraction anatomical changes. However, the introduction of daily imaging into the clinical workflow has been coupled with a simultaneous introduction of so-called “IGRT errors.” IGRT errors arise from inaccuracies in registering the patient’s daily setup images with the simulation images acquired prior to

treatment. Such errors could arise due to technical challenges with the image registration algorithms themselves, problems with applying the image registration algorithms to a particular set of patient images, or human mistakes made while interpreting the results of an image registration. While IGRT has increased the precision of radiotherapy treatments, it can lead to treatments that are "precisely wrong."

As a preliminary step to mitigate IGRT errors, we propose the development of novel tools for the automatic detection of IGRT errors. Specifically, we develop a convolutional neural network (CNN)-based model for detecting the rare but serious IGRT error of off-by-one vertebral body misalignments in radiation therapy treatments targeting the thoracic spine. We develop a second CNN-based model for detecting the more generic IGRT error of translational shifts of 1 cm from treatment isocenter in all anatomic regions. We apply both models to retrospective image data from patients aligned using daily image guidance in order to detect previously unreported IGRT errors and near miss events, and to understand where in the clinical workflow such incidents originated.

Finally, we understand that new evidence-based tools can only be effective if they are successfully integrated into the clinical environment. A rigorous implementation science approach is a necessary step to integrating novel technologies and reducing the well-documented lag time from research to practice. We study the barriers and facilitators to use of both automated tools that are commercially available as well as automated tools still in development. We use a survey study to evaluate medical dosimetrists' perceptions of auto-contouring and automated treatment planning tools and their perceived barriers to regular clinical use of such tools. To better understand how a new automated tool designed to assist in the IGRT review portion of

weekly chart checks could be integrated clinically, we use a novel thematic analysis approach to analyze the current weekly chart check workflow from the perspective of the clinical medical physicist.

The dissertation of Rachel Marie Petragallo is approved.

Nzhde Agazaryan

Matthew Sherman Brown

Daniel Abraham Low

James Michael Lamb, Committee Chair

University of California, Los Angeles

2024

DEDICATION

To my village:

Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D	Two dimensional
3D	Three dimensional
AAPM	American Association of Physicists in Medicine
AC	Auto-contouring
AI	Artificial intelligence
AMC	Academic medical center
ASTRO	American Society for Radiation Oncology
ATP	Automated treatment planning
AUC	Area under the curve
CAMPEP	Commission on Accreditation of Medical Physics Education Programs
CBCT	Cone beam computed tomography
CMC	Community medical center
CNN	Convolutional neural network
CT	Computed tomography
DICOM	Digital imaging and communications in medicine
DL	Deep learning
DRR	Digitally reconstructed radiograph
FIF	Field-in-field
FMEA	Failure modes and effects analysis
GTV	Gross tumor volume
GUI	Graphical user interface
HBMC	Hospital-based medical center
IGRT	Image guided radiotherapy
IMRT	Intensity modulated radiation therapy
IRB	Institutional review board
IT	Information technology
ITK	Insight segmentation and registration toolkit

KBP	Knowledge-based planning
MP3.0	Medical Physics 3.0
MPPG	Medical physics practice guideline
MU	Monitor unit
PTV	Planning target volume
QA	Quality assurance
ROC	Receiver operating characteristic
RO-ILS	Radiation Oncology Incident Learning System
SBRT	Stereotactic body radiation therapy
SSD	Source-to-surface distance
TG	Task group
UCLA	University of California, Los Angeles
UTAUT	Unified theory of acceptance and use of technology

ACKNOWLEDGEMENTS

First and foremost, an enormous thank you to my advisor, Dr. James Lamb. Thank you for being both a professional mentor and a personal mentor. You have been a constant source of support, encouragement, and overwhelming kindness throughout my entire graduate career. Thank you for consistently making time to teach me and for encouraging me to find and pursue my own passions. Thank you for welcoming me into your lab and for always making me feel that it was a place I belonged. Thank you for caring so deeply about each and every one of your students and being so committed to their success. I am incredibly grateful to have worked with someone who exemplifies integrity, leadership, respect, compassion, and professionalism in every aspect of their work.

I would also like to thank the other members of my dissertation committee: Dr. Nzhde Agazaryan, Dr. Matthew Brown, and Dr. Daniel Low. To Dr. Agazaryan, thank you for always being available to answer questions and share your vast clinical expertise. I am incredibly grateful for your constant support and for the insights gained through every interaction with you. To Dr. Matthew Brown, thank you for providing a unique perspective on the technical aspects of my work and the challenges to clinical adoption. Your commitment to providing all students in our department with countless opportunities to work with and learn from others is unparalleled, and I know that I'm far from the only one who has benefitted greatly from your focus on collaboration. To Dr. Daniel Low, thank you for always taking the time to thoughtfully critique my work and for providing me with incredibly detailed feedback throughout my graduate career. I am so grateful for your passion and enthusiasm for teaching all students, not just your own.

I am beyond thankful for my time in the Physics and Biology in Medicine graduate program. To Dr. Michael McNitt-Gray, thank you for fostering such an excellent department culture. I am constantly amazed by all the roles you hold and how you excel in every single one of them. Thank you for the personal and professional leadership you demonstrate in your role as program director. To Reth Im and Alondra Correa-Bautista, thank you for all you do to keep the program running smoothly. To my fellow graduate students, thank you for your support, your commiseration at times, and most importantly your camaraderie. Thank you to Morgan, Pav, Peter, Claudia, and Louise for your incredible friendships and just the right amount of distraction. To my labmates Dishane, John, Yasin, and Justin: thank you for constantly offering research suggestions, encouragement, and friendship. I cannot adequately express how grateful I am for the opportunity to collaborate with such intelligent and thoughtful peers.

To my friends, my family, and my friends who have become my family: thank you for being my village through graduate school, but in reality for so much longer than that. Thank you to Sarah, Lauren, Claire, and Parisa for believing in me and cheering me on through every hurdle. I'm incredibly lucky to have such accomplished, intelligent, funny, and kind friends to look up to. Thank you to Doug, Joyce, Michael, Christopher, and Ellie for welcoming me so easily into your family and for celebrating each and every milestone with me. All of you have an enthusiasm for life that is infectious, and I'm so grateful for the countless wonderful memories over the years. Thank you to my sister Taye for making me laugh harder than anyone, and to my mom Lisa for being an example of resilience. I am truly humbled by the support I've been lucky enough to receive over the years, and I treasure each and every one of my relationships with all of you.

Finally, to Matthew, thank you for your unwavering support in my professional career and in my personal life. You've been the biggest cheerleader when things are going well and the biggest support when they are not. You've been there for all the moments, big and small, for so long now that I truly can't remember what it's like to not have you there. Thank you for your willingness to embark on any adventure and for being a constant source of pure joy in my life. My home is wherever I'm with you.

Chapter 2 is a version of Petragallo R, Bertram P, Halvorsen P, Iftimia I, Low DA, Morin O, Narayanasamy G, Saenz DL, Sukumar KN, Valdes G, Weinstein L, Wells MC, Ziemer BP, Lamb JM. Development and multi-institutional validation of a convolutional neural network to detect vertebral body mis-alignments in 2D x-ray setup images. *Med. Phys.* 2023; 50(5): 2662-2671. doi:10.1002/mp.16359

Chapter 3 is a version of Petragallo R, Charters JA, Lamb JM. Using geometrically-realistic images to update a convolutional neural network to detect off-by-one vertebral body misalignments in 2D x-ray setup images [in preparation].

Chapter 5 is a version of Petragallo R, Luximon DC, Neylon J, Ritter T, Lamb JM. A convolutional neural network-based retrospective search for previously unreported radiation events in a stereoscopic planar x-ray setup image database [in preparation].

Chapter 6 is a version of Petragallo R, Bardach N, Ramirez E, Lamb JM. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists. *J. Appl. Clin. Med. Phys.* 2022; 23(5): e13568. doi:10.1002/acm2.13568

Chapter 7 is a version of Petragallo R, Luximon DC, Neylon J, Bardach N, Ritter T, Lamb JM. Clinical physicists' perceptions of weekly chart checks and the potential role for automated image review assessed by structured interviews. *J. Appl. Clin. Med. Phys.* 2024; 25(5): e14313. doi:10.1002/acm2.14313

VITA

EDUCATION

University of California, Los Angeles October 2020
M.S. in Physics and Biology in Medicine

University of Utah May 2015
B.S. in Physics
B.S. in Applied Mathematics

PUBLICATIONS

Petragallo R, Luximon DC, Neylon J, Bardach N, Ritter T, Lamb JM. Clinical physicists' perceptions of weekly chart checks and the potential role for automated image review assessed by structured interviews *J. Appl. Clin. Med. Phys.* 2024; 25(5): e14313. doi:10.1002/acm2.14313

Charters JA, Luximon D, **Petragallo R**, Neylon J, Low DA, Lamb JM. Automated detection of vertebral body misalignments in orthogonal kV and MV guided radiotherapy: application to a comprehensive retrospective dataset. *Biomed. Phys. Eng. Express.* 2024; 10(2): 025039. doi:10.1088/2057-1976/ad2baa

Petragallo R, Bertram P, Halvorsen P, Iftimia I, Low DA, Morin O, Narayanasamy G, Saenz DL, Sukumar KN, Valdes G, Weinstein L, Wells MC, Ziemer BP, Lamb JM. Development and multi-institutional validation of a convolutional neural network to detect vertebral body mis-alignments in 2D x-ray setup images. *Med. Phys.* 2023; 50(5): 2662-2671. doi:10.1002/mp.16359

Luximon DC, Ritter T, Fields E, Neylon J, **Petragallo R,** Abdulkadir Y, Charters J, Low DA, Lamb JM. Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. *Med. Phys.* 2022; 49(10): 6410-6423. doi:10.1002/mp.15927

Petragallo R, Bardach N, Ramirez E, Lamb JM. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists. *J. Appl. Clin. Med. Phys.* 2022; 23(5): e13568. doi:10.1002/acm2.13568

Tward JD, O'Neil B, Boucher K, Kokeny K, Lowrance WT, Lloyd S, Cannon D, Stephenson RA, Agarwal N, Farr T, **Petragallo R**, Sherar NZ, Kunz I, Hofer A, Courdy S, Shrieve DC, Dechet C. Metastasis, mortality, and quality of life for men with NCCN high and very high risk localized prostate cancer after surgical and/or combined modality radiotherapy. *Clin. Genitourin. Cancer*. 2020; 18(4): 274-283. doi:10.1016/j.clgc.2019.11.023

ABSTRACTS AND PRESENTATIONS

Petragallo R, Charters J, Lamb JM. A CNN-based retrospective search for previously unreported treatment errors in a stereoscopic planar x-ray setup image database. Oral presentation at the 65th AAPM Annual Meeting; July 27th, 2023; Houston, TX.

Petragallo R, Bardach N, Luximon DC, Neylon J, Ritter T, Lamb JM. The shifting weekly chart check paradigm. E-poster presentation at the 65th AAPM Annual Meeting; July 25th, 2023; Houston, TX.

Petragallo R, Bertram P, Halvorsen P, Iftimia I, Low DA, Morin O, Narayanasamy G, Saenz DL, Sukumar KN, Valdes G, Weinstein L, Wells MC, Ziemer BP, Lamb JM. A multi-institutional, convolutional neural network-based approach to the detection of vertebral body mis-alignments in planar x-ray setup images. Oral presentation at the 64th AAPM Annual Meeting, Early-Career Investigator Symposium; July 11th, 2022; Washington, D.C.

Petragallo R, Charters J, Kunz J, Low DA, Morin O, Saenz DL, Salter B, Valdes G, Ziemer BP, Lamb JM. Multi-institutional validation of a convolutional neural network-based approach to the detection of vertebral body misalignments in planar x-ray setup images. Snap Oral presentation at the 63rd AAPM Annual Meeting; July 25th, 2021; Virtual Meeting.

Petragallo R, Low DA, Lamb JM. A convolutional neural network-based retrospective search for previously unreported treatment errors in a planar x-ray setup image database. E-poster presentation at the 63rd AAPM Annual Meeting; July 27th, 2021; Virtual Meeting.

Petragallo R, Lamb JM. A comparison of convolutional neural networks and logistic regression for the detection of vertebral body misalignments during radiation therapy. E-poster presentation at the 62nd AAPM Annual Meeting; July 12th, 2020; Virtual Meeting.

AWARDS

Second Place Early-Career Investigator July 2022
AAPM Annual Meeting

Best Medical Travel Fellowship July 2022
Best Medical/ AAPM

Second Place Norm Baily Award May 2022
AAPM—Southern California Chapter

CHAPTER 1: INTRODUCTION

1.1 Introduction to radiation therapy

Radiation therapy is a type of cancer treatment that uses high energy particles to deliver high doses of radiation to cancer cells within the body. Cancer cells are killed as a result of DNA damage caused by the radiation[1]. This type of treatment is both safe and effective, and it is estimated that 50% of new cancer diagnoses worldwide would benefit from radiation therapy[2–4]. Delivering a high dose of radiation to the tumor is crucial, but it is also the goal of radiation therapy to spare nearby healthy tissue as much as possible[5].

The overarching goal of radiation therapy has always been to maximize dose to the target while simultaneously minimizing dose to nearby normal tissue. Recent technological advancements, specifically new imaging modalities and advanced linear accelerators[6], have enabled providers to design precise radiotherapy treatment plans to accomplish this goal. Modern linear accelerators equipped with retractable multi-leaf collimators allowed for the development of an advanced type of radiation therapy termed intensity modulated radiation therapy (IMRT). The improved dose conformity and steep dose gradients achievable with IMRT necessitated the subsequent development of improved imaging in order to accurately localize patient anatomy[7]. The highly conformable nature of IMRT coupled with new advanced imaging techniques gave rise to image guided radiation therapy, or IGRT. IGRT is defined by daily imaging of the patient prior to radiation delivery. This frequent imaging allows for the patient to be accurately positioned, and for adjustments to be made immediately in response to anatomical changes[8].

1.2 Image guidance during radiation therapy

Although radiation therapy as a treatment modality for cancer has been employed for well over 100 years now, it is only within the past few decades that the routine use of IGRT has become mainstream. Several review papers detail the rapid rise in clinical use of IGRT and summarize the various technologies, along with the advantages and disadvantages of each[9–12]. IGRT was quickly adopted into routine clinical use, with the vast majority of radiation treatments in the United States currently using some form of image guidance[13,14]. Image guidance can increase both the quality and the safety of radiation therapy treatments, and the technology became a defining feature of modern radiotherapy soon after its inception[15,16].

The benefits of IGRT have been well documented in the literature. IGRT enables the radiotherapy treatment team to achieve high accuracy in the patient setup[17], and this high accuracy subsequently allows for a reduction to be made in the planning target volume (PTV) margins used during treatment planning[18]. Geometric variability of the target setup is reduced through the consistent use of daily imaging[19–21], ultimately meaning that the prescribed dose is more likely to be delivered as intended[22]. IGRT has been clinically implemented for a wide variety of anatomical sites and radiotherapy applications, including pelvis[23], prostate[24], rectal[25], lung[26], head and neck[27], respiratory motion management[28], and stereotactic body radiotherapy (SBRT)[29]. Regular clinical use of IGRT has been shown to lead to both higher tumor control rates and reduced toxicity to nearby normal tissue[30,31], in addition to a significant reduction in the frequency of gross treatment errors[32]. IGRT has become a fundamental component of modern radiation therapy and has advanced the ultimate goal of precise and accurate treatments[33].

1.3 Limitations of image guided radiation therapy

The introduction of IGRT into routine clinical practice has not been without its own set of unique challenges and clinical considerations. Daily imaging does result in increased dose to the patient[15], although this is typically negated by the PTV margin reduction that can be achieved. This increased dose is perhaps more relevant for pediatric patients, and careful consideration of the potential long-term effects is necessary[34]. IGRT also comes with increased expenses in terms of capital, maintenance costs, and already limited human resources[8]. The use of daily imaging has the potential, perversely, to increase the likelihood of errors resulting from misinterpreting the images that are directly guiding a patient's treatment[10]. Daily imaging also has the potential to provide false reassurance if used inappropriately or without a robust quality assurance program in place[22]. IGRT can increase the quality and safety of radiation therapy, but the American Society for Radiation Oncology (ASTRO) cautioned that it must be deployed in a robust and safe manner[35]. Ultimately, failure to understand and appropriately use IGRT can result in a treatment that is "precisely wrong"[36].

1.4 Characterization of the ExacTrac image guidance system

1.4.1 Technical specifications

ExacTrac (Brainlab AG, Feldkirchen, Germany) is one specific type of image guidance currently available for clinical use. The imaging system consists of two floor-mounted kilovoltage x-ray tubes that project obliquely through the patient onto two corresponding flat-panel detectors mounted on the ceiling (**Figure 1.1**). After acquisition of the two stereoscopic x-rays, the six degree of freedom fusion software compares the 2D images with digitally

reconstructed radiographs (DRRs) generated from the patient's planning computed tomography (CT) scan at various translational and rotational shifts. The fusion algorithm locates the pair of generated DRRs that show maximum similarity to the acquired x-rays and calculates the translational and rotational shifts that should be performed on the patient. A more detailed description of the ExacTrac system and image registration algorithm can be found in the literature[37,38]. ExacTrac has demonstrated submillimeter alignment accuracy[39,40], overall spatial accuracy on the order of 1.24-1.35 mm[41], and mean deviations of less than 1 degree in all three rotational directions[42]. Use of daily ExacTrac imaging can allow for the patient's position to be reproduced within 1 mm from one treatment day to the next[43]. Furthermore, it has been shown that this high level of accuracy in locating the treatment isocenter position is achievable even when an initial patient setup has a large error[44].

Figure 1.1: Geometry of the ExacTrac IGRT system.

1.4.2 Predominant uses

Due to the high positioning accuracy that can be obtained with ExacTrac, it is frequently used for imaging prior to and during high-dose SBRT treatments. SBRT involves delivering dose to the patient over a small number of fractions (typically 1-5), with a high dose delivered per fraction. Because of the high dose per fraction, accurate target localization and patient immobilization during treatment is crucial. Early immobilization techniques were often invasive and uncomfortable for the patient, involving surgically implanted metal devices. Numerous studies have found that the accuracy of target localization obtained with ExacTrac imaging is comparable to that obtained with these invasive stereotactic frames[45–47], leading to a new approach of frameless SBRT treatments. Care must still be taken to properly immobilize the patient during treatment in order for frameless SBRT to be successful[48], but when done properly this technique achieves very high local tumor control that is well tolerated by patients[49].

1.4.3 Known limitation: thoracic spine

While patient setup using ExacTrac has been shown to be highly accurate, the system does have some critical limitations, especially with regard to radiation therapy targeting the thoracic spine. In this region, implanted markers are still more effective in patient positioning, reducing the deviation from the planned isocenter by up to half a millimeter[50]. Intra-fraction motion can be of particular concern as well, with one study finding that vertebral anatomy can vary as much as 3 mm between measurements, and that this movement could occur in as little as 5 minutes[51]. Perhaps most critically, intensity-based 2D-3D image registration methods such as that employed by ExacTrac are particularly susceptible to local minima in the algorithm's cost function[52]. Practically speaking, this means that the algorithm has the potential to lock on to an

adjacent vertebral body instead of the one intended for treatment depending on the accuracy of the initial patient setup. Because adjacent vertebral bodies in the thoracic spine look similar, this particular error can be difficult to visually verify.

1.5 Errors in medicine

In 1999, the Institute of Medicine released a report estimating that between 44,000 and 98,000 patients die in hospitals each year in the United States due to preventable errors[53]. The renowned psychologist James Reason theorized that accidents like those occurring in the healthcare system result from two distinct types of errors: active, or unsafe acts directly linked to an event, and latent, or systemic conditions and practices leading to an event[54]. Furthermore, he argued that latent human errors are more significant than technical failures[55], a sentiment echoed by the finding that up to 80% of accidents in high risk industries can be attributed to operator error[56]. This work ultimately led Reason to propose his "Swiss cheese model"[57] (**Figure 1.2**) to explain how system failures occur, even with multiple barriers in place to mitigate against accidents.

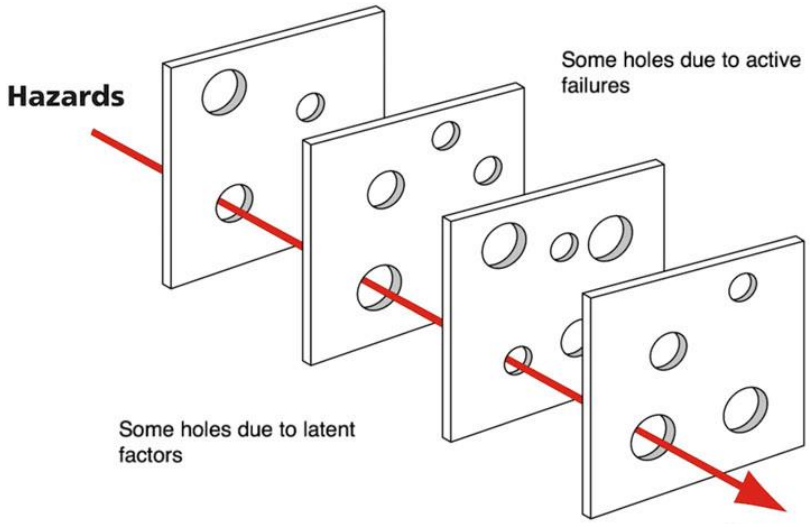

Figure 1.2: Reason's Swiss cheese model of system failures.

1.6 Errors in radiation oncology

The field of radiation therapy is far from immune from system failures of the type described by Reason. One study estimated that 269 potential failure modes exist in the planning and delivery of radiation therapy[58]. Redundant safety checks exist at almost every stage of the radiation oncology clinical workflow, drastically reducing the chances for errors to reach the patient and cause harm. Even so, studies have found a high incidence of treatment errors in radiation oncology due to incorrect patient setup[59,60]. Such errors can have devastating, at times even fatal, consequences, as was highlighted in a 2010 New York Times exposé[61].

Misalignment of radiation treatments to the patient anatomy, even when using IGRT, remains an important source of error. The Radiation Oncology Incident Learning System (RO-ILS) was established in 2014 as an inter-institutional error reporting system. In a 2018 report, 396 incidents were investigated, of which 40 were classified as "wrong shift instructions given to therapists," and 34 as "wrong shift performed at treatment.[62]" The Quarter 3 2016 RO-ILS summary report describes a case of a treatment misaligned by 3 cm and opined: "This is one of approximately 28 IGRT events documented this quarter. Clearly we need to increase the attention paid to how to decrease the number of IGRT-related issues throughout the field.[63]"

1.7 Clinical significance of patient setup errors

The importance of ensuring accurate patient setup prior to radiation therapy treatment cannot be overstated. Patient setup and dose delivery errors that occur during the course of a treatment fraction are difficult to detect, meaning that there are likely quality issues in radiation oncology that are not currently well-studied. The patient positioning step of the treatment workflow is both high severity and high risk, and was ranked in the top 20% most hazardous

steps by an AAPM Task Group Report[64]. Additional studies have supported this finding, documenting a high incidence of treatment errors in radiation therapy caused directly by incorrect patient setup[59,60,65]. One study found that geometric misses of the target had the highest error probability, and once again tied this failure mode back to improper patient setup[66]. While many patient-specific factors have been shown to influence the reproducibility of the patient setup[67], the specific type of radiation treatment used also plays a role. As radiation therapy has become more precise with the development of IMRT and, subsequently, IGRT, these precise treatments are more affected by small patient misalignments than more traditional (and more simple) methods of radiation delivery[68]. Yamashita et al. found that a target margin of up to 8 mm was necessary to account for the various types of patient setup error present in patients being treated for esophageal cancer[69], highlighting the impact these errors can have on precise radiation delivery. Incorrect patient setup has been shown to lead to both a significant decrease in PTV coverage[70,71] and increases in the dose delivered to nearby normal tissue[71].

1.8 Incident learning in radiation oncology

Radiation therapy remains a remarkably safe treatment modality for patients, with error rates per delivery fraction of well under 1% reported in the literature. Some highly publicized accounts of egregious radiation therapy errors[61,72–75] have caught public attention, even though the vast majority of the rare errors that do occur have little to no observable clinical consequences. Still, knowledge of the exact rate and types of radiation mis-administration is critical information for all members of the treatment team. While some previous studies have attempted to define the error rate within radiation therapy treatments, such studies are limited in that they rely on self-identified and self-reported errors[76]. RO-ILS, jointly sponsored by ASTRO

and AAPM, represents the most comprehensive effort to date within the field to collect data on errors. Incident learning systems such as RO-ILS are widely regarded as invaluable tools for improving both the safety and the quality of patient treatments[77]. Literature in recent years has highlighted some successes with implementing incident learning systems in various radiation oncology clinics[78,79], and has synthesized some key themes regarding incident learning systems. These include an emphasis on incident learning (rather than incident reporting) and the value of a non-punitive department culture[80], the importance of information sharing to prevent future incidents and facilitate safety improvements[81], and the value of viewing near-misses as valuable "free lessons"[82]. The explosion of technological advancements in radiation therapy in recent years has reduced the likelihood of certain errors but simultaneously introduced new avenues for other errors[83], a trend that shows no indication of abating. Understanding and mitigating such errors is critical for the safe delivery of radiation treatments to the patient. While incident learning systems represent a vital component of the safety policies and procedures, their efficacy is limited in that they cannot be used to truly measure department error rates[84] due to the aforementioned known issue of under-reporting. There is still a pressing need to better understand the true frequency of errors, particularly those which are most likely to lead to patient harm[85].

1.9 Automation to prevent radiation therapy errors

Because errors in radiation therapy can have such severe ramifications, it is of the utmost importance to introduce multiple safety barriers into each stage of the treatment workflow. Such safety barriers can take many forms, as illustrated in **Figure 1.3**, but automated checks and forcing functions have been shown to be more effective at error prevention than policies and

procedures or staff training[86]. Various automated tools have been proposed for different types of error prevention, including eliminating the common error pathway of providing incorrect shifts to therapists[87], treatment plan quality control[88], and the use of cameras to detect treatment of the incorrect patient or anatomic site[89]. Recently, several papers have summarized the advancements made in applying convolutional neural networks (CNNs) to problems of medical image analysis[90–92]. The excellent performance of CNNs on image analysis problems led to the novel concept of using IGRT as an additional means of ensuring patient safety[93]. Subsequent studies have shown promise in using CNNs during IGRT for the narrow tasks of treatment target localization[94,95] and image registration[95,96], but the concept of using IGRT itself as a safety barrier has been relatively understudied to date.

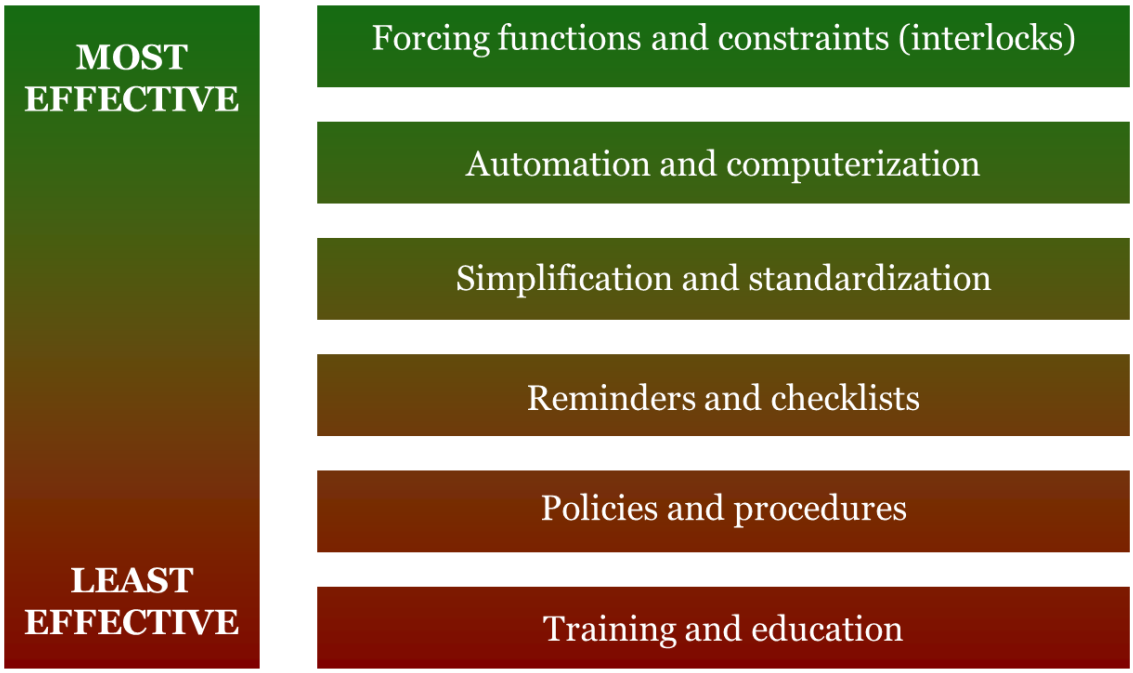

Figure 1.3: Hierarchy of hazard mitigation effectiveness, adapted from Hendee et al.

1.10 Clinical adoption of novel technologies

The challenges of translating basic science research into clinical practice have been well documented[97,98]. On average, there is a lag time of 17 years for research evidence to reach routine clinical practice[99]. In comparison to funding for basic science research, very little funding is awarded for studying best practices to shorten this lag time[100]. However, it is clear that a focused effort on understanding the barriers to implementation is crucial for ensuring that advances in research translate to advances in patient care[101]. This identification of barriers is essential for maximizing the clinical adoption of new evidence-based tools and techniques[102].

In order to address the gap between research and clinical practice, the field of implementation science was born. Implementation science is fundamentally based on the concept that the successful implementation of evidence-based practices requires an evidence-based approach to their actual implementation[103]. The factors preventing successful uptake of a new technology can vary widely, but often depend on contextual factors significantly more than the proven effectiveness of the new innovation[104]. Various groups have proposed strategies for successfully implementing implementation science[105–107]. In general, these strategies focus on human factors as a key component to increasing the likelihood that a new technology is successfully integrated into a given organization.

Implementation science has proven to be quite successful in facilitating the adoption of evidence-based practices into routine clinical use, ultimately leading to improved patient care[106,108]. It has become an established and indeed widely accepted field in healthcare broadly[109]. The incorporation of strategies from human factors research is a small but significant shift within implementation science focusing on healthcare[110,111], and is especially critical in

light of the fact that so much of successful implementation science hinges on human behavior. Implementation science has been well-studied in other healthcare fields, and its application to specific radiation oncology tools is a natural extension of the underlying theory.

1.11 Overview and specific aims

The goal of this dissertation is threefold: to develop novel automated tools for the detection of patient misalignments in image guided radiotherapy, to apply these automated tools to archived image data in order to better understand IGRT error rates, and finally to analyze best practices for introducing automated tools more broadly into the current radiation oncology clinical workflow. The individual specific aims are as follows:

1. Develop and validate automated tools for detecting both off-by-one vertebral body misalignments and gross patient misalignments in ExacTrac IGRT.
2. Utilize automated methods to identify misalignments in archived ExacTrac patient setup images that represent previously unreported IGRT errors or radiation therapy incidents.
3. Evaluate the barriers and facilitators to clinical implementation of both automated treatment planning tools and automated tools intended to address and prevent IGRT errors.

Chapter 2 addresses Specific Aim 1 by proposing the use of an automated tool for the detection of off-by-one vertebral body misalignments in 2D ExacTrac images developed using a multi-institutional collaboration of image data. **Chapter 3** describes improvements made to this model in the form of incorporating the stereoscopic geometry of the ExacTrac system into the training data. **Chapter 4** again addresses Specific Aim 1 by proposing the development and validation of an automated tool to detect gross patient misalignments in a wide variety of anatomical treatment

sites all imaged using the ExacTrac IGRT system. **Chapter 5** provides a solution to Specific Aim 2 by using the models for detecting patient misalignments described in **Chapter 3** and **Chapter 4** to perform a retrospective search for previously unreported treatment incidents, hereby allowing for an independent quantification of the IGRT error rate. **Chapter 6** addresses Specific Aim 3 by using a survey study of medical dosimetrists to identify their perceived barriers to use of both auto-contouring and automated treatment planning tools in their routine clinical workflow. **Chapter 7** again addresses Specific Aim 3 by using thematic analysis to gain insight into the current practices employed by clinical medical physicists during their weekly chart checks, and identify openings for automation to assist in the IGRT review portion of such checks.

CHAPTER 2: DEVELOPMENT AND MULTI-INSTITUTIONAL VALIDATION OF AN AUTOMATED TOOL FOR DETECTING VERTEBRAL BODY MISALIGNMENTS

2.1 Introduction

2.1.1 Off-by-one vertebral body misalignments

Off-by-one vertebral body errors are particularly insidious due to the translational symmetry of the vertebral column[112]. Redundant safety checks are in place during the simulation imaging, treatment planning, setup imaging, and treatment delivery phases, with the overall goal of catching potential errors and correcting them before treatment is actually delivered to the patient. Identification of error pathways[64,113,114], the use of established safety protocols[115,116], and redundant safety checks[117] are common approaches to radiation therapy error mitigation. However, ample evidence demonstrates inadequate compliance with safety protocols in practical healthcare situations. Recent studies have reported surgical safety checklist compliance of 52-80%[118–120]. In a 2017 report of adverse events in Minnesota[121], the most commonly identified root cause category was "Rules/Policies/Procedures", and within that category the most commonly reported factor was "policies/procedures are in place, but not followed". Mallet et al[122] studied eight cases of wrong site/procedure/patient events occurring from 2008-2010 at an academic medical center. They stated "the Rules, Policies, and Procedures category contained the highest number of failure modes (22) and was present in all 8 wrong-person, procedure, and site events analyzed. *Every failure mode within this category reflects an incomplete or improper use of the 3 steps involved with the Universal Protocol* (emphasis added)."

2.1.2 IGRT errors in ExacTrac

In this dissertation we focus on what has been termed an "IGRT Error", in which human error contributes to misalignment of a radiotherapy treatment even with the use of image guidance. Setup imaging prior to each fraction of radiation is an important part of the safety protocols in place to ensure that the patient is correctly aligned and that the radiation is delivered to the appropriate target. However, IGRT still depends on manual verification of patient alignment, a process that can be improved upon with the introduction of a quantitative image alignment metric[123]. One study found that patient setup errors accounted for almost half of all documented incidents at their institution over a 10 year period[65].

Many different imaging modalities can be used to perform setup imaging, but in this work we focus on the planar x-ray setup imaging system ExacTrac[39]. This type of imaging is commonly used for patients receiving high dose radiation to the cranial or spinal regions due to its high precision[48]. For ExacTrac imaging, two stereoscopic x-rays of the patient are acquired after initial positioning on the treatment couch. These x-rays are then matched to corresponding DRRs that are generated using the simulation CT to determine what translational and rotational shifts need to be done to the patient to align them properly before treatment. The ExacTrac system allows for highly precise patient alignment on the order of submillimeter accuracy[39]; however, due to a relatively small field of view and the possibility of an image similarity-based minimization process to lock on to local minima in the cost function[52], the system can align to an adjacent vertebral body instead of the one intended for treatment. This effect is particularly pronounced in the lower cervical spine and thoracic spine since adjacent vertebral bodies in these regions can be difficult to visually differentiate[124,125]. A wrong level vertebral body mismatch

can have dire consequences for the patient undergoing treatment including reduced tumor control as well as increased normal tissue damage. Some institutions attempt to mitigate this risk factor by pairing ExacTrac imaging with additional imaging to provide a redundant check on the patient's treatment position, but even with this safety measure in place errors can and do still occur.

2.1.3 Automated image review

Due to the limitations of protocol- and redundant check-based IGRT error mitigation, our group proposed that an automated image-review algorithm could be inserted into the IGRT process to act as an interlock to detect and prevent IGRT errors[93]. While deep learning algorithms have been introduced to many parts of the radiation oncology patient treatment and quality assurance workflows[126–129], to our knowledge our work is the first to tackle the application of deep learning as a failsafe to strengthen the human effort of IGRT image review. Previous work applying deep learning to computer vision problems in the spine has focused on approaches for automatic segmentation[130,131], detection[132], diagnosis[133,134], and even motion monitoring[135], with good results. Furthermore, since the mid-2010s CNNs have been shown to greatly outperform other methods in image classification tasks[136], leading us to choose this methodology for our own work.

2.1.4 Study overview

Our work develops the novel application of a deep learning-based tool for the automatic detection of patient misalignments in IGRT. Here we introduce a highly accurate deep learning-based approach to detect off-by-one vertebral body misalignments developed and validated using a multi-institutional patient dataset. By assembling a dataset consisting of images from a multi-

institutional collaboration, our work overcomes a common limitation of artificial intelligence algorithms, namely the lack of robustness due in large part to a lack of data[137,138]. This problem was called out specifically by Qu et al. who argue that "shared huge datasets are needed" in order to overcome the current limitations of deep learning-based spine image analysis[139]. Our goal in this work is to present a robust and validated deep learning-based solution to the known problem of vertebral body misalignments in the ExacTrac imaging setup system.

2.2 Materials and methods

2.2.1 Data collection and curation

Anonymized patient data was collected from all six participating institutions under the auspices of an approved Institutional Review Board (IRB) research protocol at the coordinating institution (UCLA). Each institution searched their own internal database to identify thoracic spine patients imaged using ExacTrac during the course of their radiation treatment. In all, 429 patient datasets were collected. **Table 2.1** details the datasets provided by each institution.

Table 2.1: Dataset statistics for the multi-institutional collaboration.

Institution	Number of Patients	Number of Fractions	Number of Clinical Images	Date Range
UCLA	87	330	660	6/2014 – 12/2017
Institution 2	72	246	492	10/2017 – 10/2021
Institution 3	45	228	456	4/2015 – 7/2021
Institution 4	38	116	232	7/2020 – 11/2021
Institution 5	174	642	1,284	7/2020 – 11/2021
Institution 6	13	30	60	12/2013 – 1/2021

For each of these patients, ExacTrac x-ray/DRR pairs from each treatment fraction were anonymized and collected into a central image database at UCLA. If multiple x-ray/DRR pairs were present in a single fraction (typically the case), the final x-ray/DRR pair from that treatment fraction was used. The final database consisted of 3,184 x-ray images, representing 1,592 individual treatment fractions since each fraction generates two x-ray setup images, as well as

the corresponding DRRs. A representative x-ray/DRR pair is shown in **Figure 2.1**. It was visually validated that these clinically-approved image sets did not inadvertently contain misaligned images (i.e. actual treatment errors).

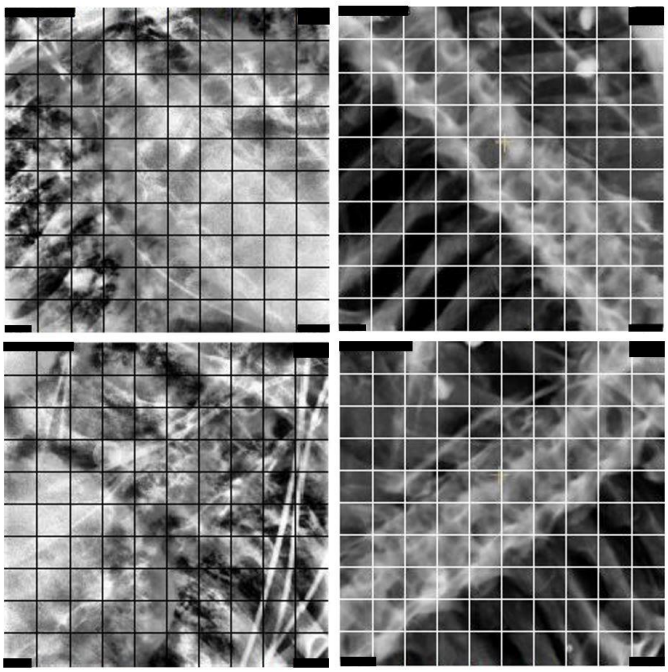

Figure 2.1: X-rays (left) and DRRs (right) from an example properly aligned thoracic spine patient.

2.2.2 Simulation of misaligned data

Simulated off-by-one vertebral body misalignments were then created for each x-ray/DRR pair using a semi-automated method based on a grid search of local maximum values of the cross-correlation coefficient. The x-ray image and DRR were each first down-sampled by a factor of four. The down-sampled DRR was then shifted pixel-wise against the stationary x-ray image, and the cross-correlation coefficient computed for each of these shifts. Cross-correlation coefficients were used as the image similarity metric since they exhibit good performance in the 2D to 3D medical image registration domain[140]. Finally, the local maximum values were labeled

on the grid view of all cross-correlation coefficients. From this map of values, the two local maximum points representing a shift in each direction along the spinal column were manually selected. These shifted DRRs were verified against the original x-ray image to ensure that the images did indeed represent a visual shift by one vertebral body. **Figure 2.2** illustrates the semi-automated method and shows the resulting shifted DRRs that were generated for a representative patient case.

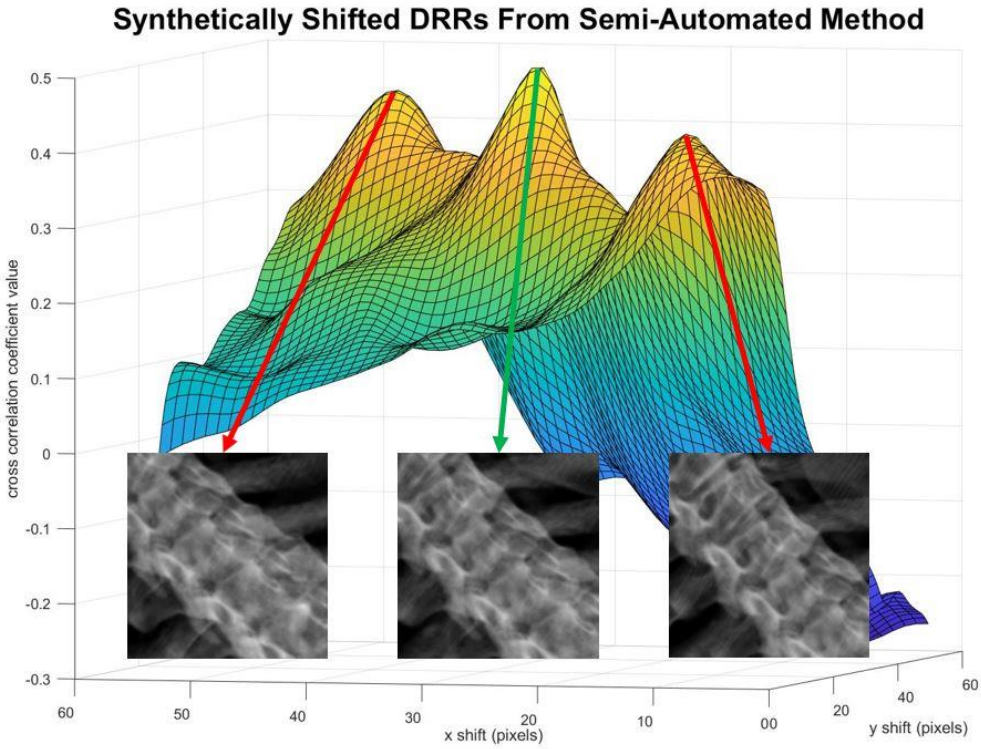

Figure 2.2: Synthetic shifted DRRs (red arrows) obtained by shifting the correctly aligned DRR (green arrow) both superiorly and inferiorly along the spinal column and realigning to a local maximum of the cross-correlation coefficient as shown by the surface plot of cross-correlation coefficient values.

In some instances, such as patients with spine fixation hardware present or a targeted vertebral body with a visible pathology in the images, the position of the local maximum did not correspond to an off-by-one vertebral body shift as judged by visual interpretation. For those images, the vertebral body shifts were performed manually without the aid of the cross-

correlation coefficient values. While these cases likely represent a relatively easy target for the error detection algorithm, we believed it was appropriate to include the entire data sample as representative of the population of treated patients.

A graphical user interface (GUI) was used to blend between the shifted DRR and unshifted x-ray to ensure that the chosen shift was visually reasonable. All images were cropped to maintain uniform image dimensions across the two classes of shifted and non-shifted images. Of note, each of the two stereoscopic images from the ExacTrac image guidance system was treated independently for this study.

Finally, 2-channel image arrays were created out of the final set of shifted and non-shifted images. The final set of DRRs contained one non-shifted, one shifted up by one vertebral body, and one shifted down by one vertebral body, for each original image pair. Each of these three DRRs was matched with the corresponding (non-shifted) x-ray. The final dataset consisted of 9,552 such 2-channel image arrays spanning the six collaborating institutions.

2.2.3 Data organization

Once the x-ray/DRR pairs were collected from each institution and synthetic alignment errors generated, the final dataset of 9,552 2-channel image arrays was organized in two distinct ways. Our primary objective was to analyze the robustness of our model using a "leave one out" approach, where the model was trained using data from all institutions but one and tested on data from the institution left out. Our secondary objective was to compare the classification accuracy from a pooled multi-institutional model to that of a single-institutional model, where for each model the training and test datasets were organized in such a way so as to ensure that no patient from any institution ended up in both datasets.

For the “leave one out” approach, all of the 2-channel images from five of the six collaborating institutions were pooled into a single training dataset. The size of the dataset ranged from 5,700 to 9,372 distinct 2-channel image arrays depending on which five of the institutions were used. This training dataset was then randomized and split into training and validation datasets using a 75/25 training/validation split. The testing dataset consisted of the images from the sixth institution, which had been left out of the training phase entirely.

For the pooled multi-institutional model, two steps were taken to create the image datasets. First, the images from each of the six individual institutions were divided into training and test datasets using an approximately 80/20 split. The split was performed on the data organized in alphabetical order by anonymized name label, with no randomization. When the exact 80/20 split occurred in the middle of a set of images from the same patient, all images from that patient were grouped into the training dataset. This was done to ensure that patients who appeared in the training dataset would not also appear in the test dataset (even if they were treated with multiple fractions) because this would potentially bias the evaluation of the final model accuracy. Second, the training datasets from all institutions were compiled into a training dataset of 7,686 distinct 2-channel image arrays, then randomized and split into final training/validation datasets using a 75/25 split. The training dataset from UCLA, consisting of 1,626 image arrays, was duplicated and saved separately prior to compiling for use in the single-institution model. The testing dataset from UCLA alone consisted of 354 image arrays while the pooled testing dataset from all six institutions consisted of 1,866 image arrays.

2.2.4 Model design – comparison of network architectures

We arrived at our final model architecture after first exploring multiple machine learning methods for our specific classification problem. We used the same spine image training dataset from UCLA to first investigate the most accurate model to use before moving to the full multi-institutional dataset. Three models were developed: a logistic regression model, a pre-trained CNN that was adapted to our application using transfer learning, and a CNN trained from scratch.

The logistic regression model was developed as a benchmark in order to compare our CNN results against more traditional machine learning techniques. This model used a combination of pixel-wise cross-correlation coefficients, cross-correlation coefficients of down-sampled images, and gradient-based cross-correlation coefficients as independent variables to classify x-ray/ DRR pairs as either shifted or non-shifted.

For transfer learning, a neural network previously trained on simple translational shifts of a fixed 1 cm was used to evaluate the potential for adaption to our specific application of detecting shifts of one vertebral body. Because relatively few thoracic spine images were available across all collaborating institutions (in comparison to the 1.2 million images originally used to train AlexNet, for example), we hypothesized that transfer learning could be an effective classification method as shown in other work[90,141]. For this transfer learning CNN, the original dataset consisted of 28,518 x-ray/DRR pairs from a variety of anatomical sites. For each of these pairs, the DRR was shifted by a fixed 1 cm in one of eight pre-determined translational directions to generate synthetic misaligned image data, leading to a final dataset size of 57,036 2-channel image arrays. Following training on this large dataset, the pre-trained network was then

applied to the smaller vertebral body training dataset described in the earlier section. The top layer of the network was trainable, whereas the weights in all preceding layers were frozen.

Finally, our purpose-built CNN was not trained on the translational images used for transfer learning, but instead was trained from scratch on the vertebral body image dataset. We investigated various network depths and parameter values as this model was trained. Unlike in our transfer learning model, weights were trainable at all levels of this model.

Receiver operating characteristic (ROC) curves were used to evaluate the performance of these three models. When the purpose-built CNN for UCLA's data was used to classify the previously unseen test image pairs, the resulting area under the curve (AUC) was 0.972. For comparison, the transfer learning-based CNN and the logistic regression models tested on the same single-institution test image dataset obtained AUCs of 0.876 and 0.801, respectively. With the specificity fixed at 99%, the purpose-built CNN achieved a sensitivity of 64.5% in correctly classifying shifts of one vertebral body as compared to a sensitivity of 32.3% for the transfer learning-based CNN and 23.7% for the logistic regression model. From these results, it was determined that a neural network trained from scratch would give the highest classification accuracy for our particular problem of detecting vertebral body misalignments.

2.2.5 Model design – parameters and hyper-parameters

Model design and hyper-parameter tuning were performed on data from UCLA alone. Network architectures of various depths were investigated and optimal classification results were achieved with a five convolutional block neural network similar to that of AlexNet[136] (shown in **Figure 2.3**). While increasing CNN depth has been shown to improve final classification accuracy[142] when trained on datasets with millions of images, it also increases the possibility of

overfitting when used with smaller datasets such as ours. Convolutional layers were followed by rectified linear activation functions and batch normalization layers. Max pooling layers were interspersed with convolutional layers. Two dense layers, each of which was followed by a 50% dropout layer to reduce overfitting[143], were used before the final classification layer. The learning rate was set to 1e-4[144]. The Adam optimizer[145] and sparse categorical cross-entropy loss function were used for training. These hyper-parameters were tuned using the data from UCLA and remained unchanged for all subsequent multi-institutional model training. Early stopping was implemented on all models, again with the aim of reducing model overfitting. The training and validation accuracy and loss as a function of training epoch are shown in **Figure 2.4** for an example model training.

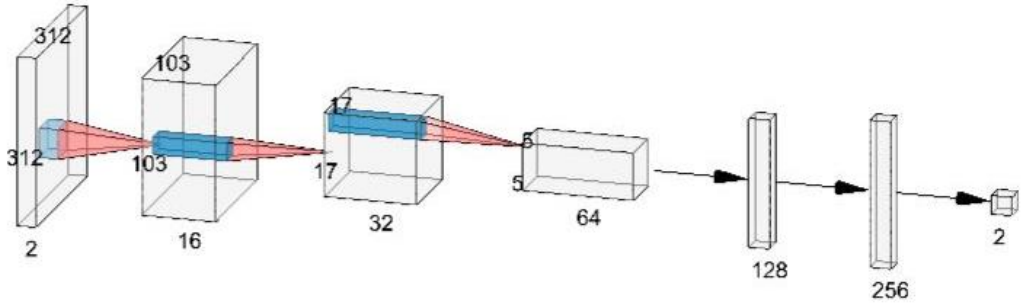

Figure 2.3: Purpose-built CNN architecture for x-ray/ DRR vertebral body image pair classification.

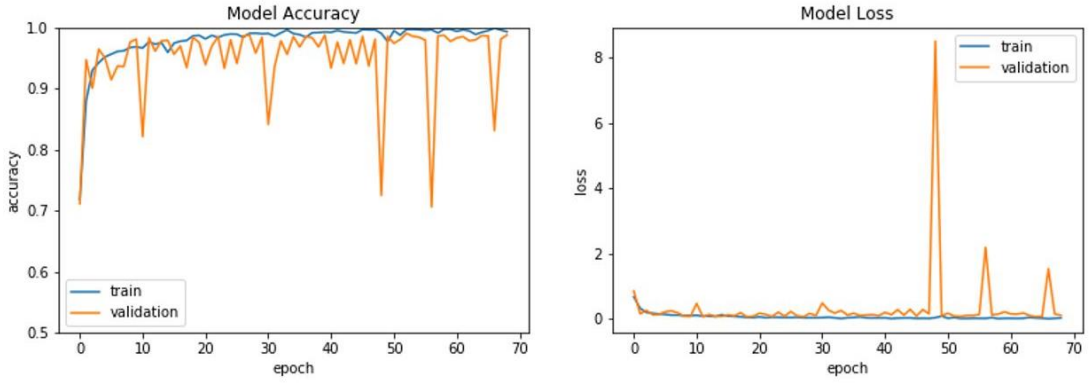

Figure 2.4: Model accuracy (left) and loss (right) as a function of training epoch for a representative model training with Institution 6 left out of the training and validation sets. Early stopping was implemented and model training stopped after 69 epochs for this particular training iteration, with the model from epoch 19 being saved and used for further analysis.

2.2.6 Model training

We designed and trained a series of CNNs to perform the binary classification task of discriminating between shifted and non-shifted images. These CNNs took as input the 2-channel arrays described in the Data collection and curation section above, where each array was labeled as either "Shifted" or "Unshifted" and consisted of either correctly aligned or mis-aligned x-ray/DRR pairs. After the CNN had been trained, it was tested on the reserved 2-channel arrays. These arrays were input into the CNN, and the model would output a score from 0 to 1 based on how likely it thought the images were to represent a patient misalignment.

Once the CNN architecture with the optimal performance on UCLA data had been identified and training hyper-parameters set, the CNN was retrained on the multi-institutional image data in two different ways:

First, we trained six models using a "leave one institution out" approach to estimate the accuracy of the model when applied to an entirely new institution's data during the testing phase. In this step, we re-trained a model using the CNN architecture and parameters described in the previous section, with the training data consisting of all images from five of the six different institutions as described above. Early stopping was implemented in all of our models with the aim of reducing overfitting, but all other hyper-parameters remained unchanged once they had been optimized using UCLA's data. At testing time, the trained model was applied to all images from the institution that had been left out of the training dataset to evaluate robustness to an outside institution's image data. This process was repeated to generate six models, each one representing each institution being left out of the training dataset.

To establish the necessity of using multi-institutional data for this error detection algorithm, we compared the CNN's performance when trained on multi-institutional versus single-institutional data. The CNN was trained using the training dataset of 1,626 x-ray/DRR pairs from UCLA. A pooled model was created by training the CNN on the pooled training dataset composed of 7,686 image arrays from all six institutions. The single-institution model was tested separately on the reserved set of patient data from UCLA, and on the pooled test set described above. The pooled model was tested only on the pooled test set.

2.2.7 ROC analysis

For both the pooled model and the six "leave one institution out" models, ROC curves were used to evaluate model performance. From these curves, the area under the curve is reported as a measure of overall model accuracy. The network outputs a continuous variable in the range of 0 to 1 representing a likelihood of image misalignment. Clinical implementation as a binary error-detection classifier would require application of a threshold, which would depend on what was deemed an acceptable sensitivity-specificity tradeoff. Sensitivity results for three different specificity values of interest, namely, 99%, 95%, and 90%, are also reported. A specificity of 99% would correlate to approximately one false positive per treatment machine per week under a typical treatment load, assuming an automated error detection algorithm was applied to every treatment—an acceptable rate of false positive disruptions compared to other safety interlocks in common use. We focus on this high specificity threshold due to the well-documented patient safety implications of alert fatigue in healthcare[146,147]. While alarm fatigue is not the only factor that can contribute to decreased staff performance in a healthcare setting[148], it is nonetheless still recognized as an important factor. Recent work has highlighted the lack of

research into alarm fatigue within radiation oncology specifically[149], and attempted to raise awareness of this problem. Specificity of 95% or 90%, corresponding to 5% and 10% false positive rates, could be appropriate in the context of stereotactic spine radiotherapy, which is a relatively infrequent and high-risk procedure. The choice of an appropriate threshold at which an alarm is triggered is an important step in reducing alarm fatigue in the clinic[147].

2.2.8 Small shift sensitivity analysis

We also analyzed the false positive rate when small shifts, on the order of a few pixels, were present in the test data. We used the final "leave one out" model where Institution 1 (UCLA) was left out of the training dataset. We first tested the final model on all images from UCLA to obtain the ROC curves and the sensitivity values at a few set specificity values, as described just above. We obtained the threshold value corresponding to the 95% specificity level for this analysis. 12 patients, representing 31 individual treatment fractions, were randomly selected from UCLA. For each fraction, both horizontal and vertical shifts of ±1 pixel, ±2 pixels, ±3 pixels, ±5 pixels, and ±8 pixels were then applied to each DRR. For the ExacTrac system, 5 pixels corresponds to a shift of approximately 1 mm, while 8 pixels corresponds to a shift of approximately 1.5 mm. Finally, the trained model was applied to all of the shifted images, and any prediction value above the threshold set by the 95% specificity level was labeled as a false positive.

2.3 Results

2.3.1 ROC analysis – "leave one institution out" models

When the six "leave one institution out" models were used to classify image pairs from the institution left out during training, the resulting AUC values ranged from 0.976 to 0.998 (**Figure 2.5**). With the specificity fixed at 99%, the corresponding sensitivities ranged from 61.9% to 99.2% with a mean of 77.6% (**Table 2.2**). When the specificity was set at 95%, corresponding sensitivities from 85.5% to 99.8% (mean: 92.9%) were observed. The median threshold value required for the 95% specificity set point was 0.682.

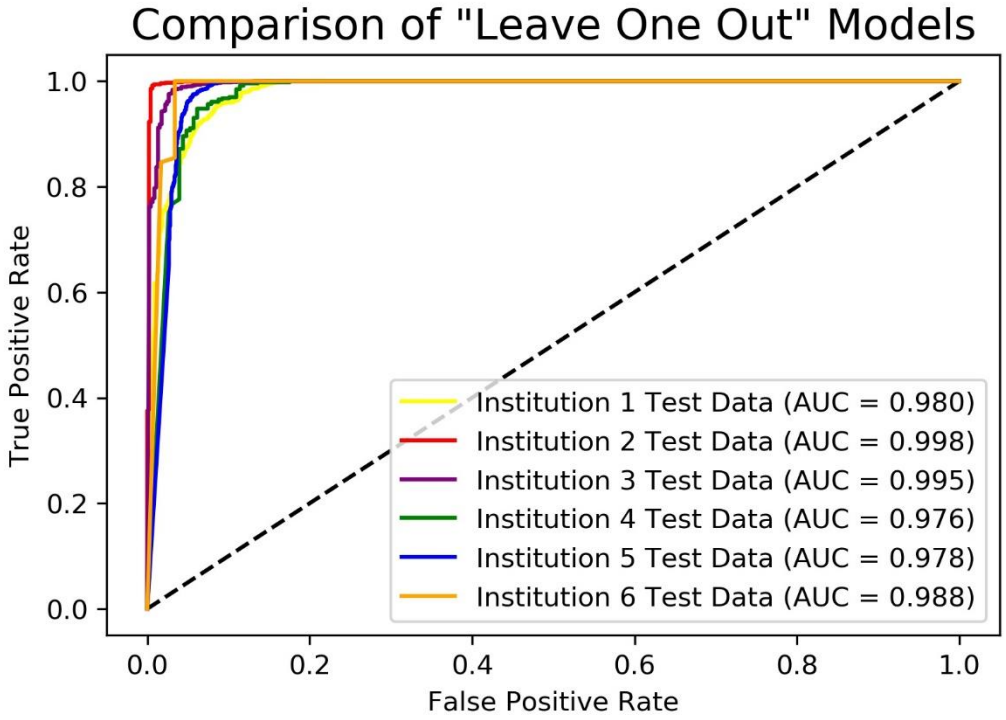

Figure 2.5: Comparison of classification performance in correctly identifying shifts of one vertebral body among the six distinct "leave one institution out" models. Each label describes the institution that was left out of training and used as the test dataset.

Table 2.2: Corresponding sensitivities for the neural network output value thresholds corresponding to 90%, 95%, and 99% specificity, respectively, for the six "leave one institution out" models.

Institution Used for Testing	% Sensitivity at 90% Specificity	% Sensitivity at 95% Specificity	% Sensitivity at 99% Specificity
Institution 1 (UCLA)	95.8	87.5	61.9
Institution 2	99.9	99.8	99.2
Institution 3	99.8	98.9	79.8
Institution 4	96.7	89.6	75.0
Institution 5	99.8	96.0	65.2
Institution 6	100.0	85.5	84.6

2.3.2 ROC analysis – pooled institution model

When the purpose-built CNN was trained and tested using only UCLA's data, it obtained an AUC of 0.975 (**Figure 2.6**). When this model trained on a single institution's data was then applied to a multi-institutional test set, the AUC dropped to 0.942. By comparison, a model with the same network architecture and hyper-parameters, but trained and tested using pooled data from all collaborators, obtained an AUC of 0.992. Sensitivity and specificity threshold values for all three of these models are reported in **Table 2.3**. It is worth noting that at the fixed specificity of 99%, the model trained using only a single institution's image data obtained a sensitivity of just 67.9% (single-institution test set) or 57.4% (multi-institution test set) as compared to 79.3% for the multi-institutional model. When the specificity was fixed at 95%, the single-institution model obtained sensitivities of 91.9% and 77.4% on the single- and multi-institutional test sets, respectively, and the multi-institution model obtained a sensitivity of 97.7%.

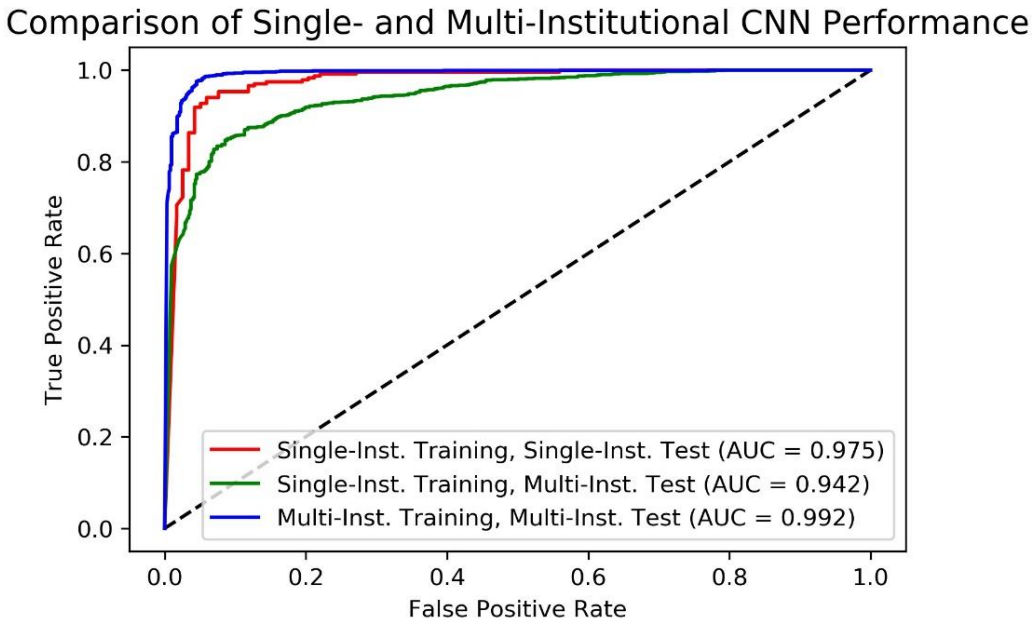

Figure 2.6: Comparison of classification performance in correctly identifying shifts of one vertebral body among three models: one trained and tested on UCLA's image data, a second trained on UCLA's data and tested on pooled data from all six collaborating institutions, and a third trained and tested on pooled data from all six collaborating institutions.

Table 2.3: Corresponding sensitivities for the neural network output value thresholds corresponding to 90%, 95%, and 99% specificity for the three models trained and tested using the datasets shown in the leftmost column. The complete ROC curves are shown in **Figure 2.6.**

Training Dataset/ Testing Dataset	% Sensitivity at 90% Specificity	% Sensitivity at 95% Specificity	% Sensitivity at 99% Specificity
Single-Institution/ Single-Institution	95.3	91.9	67.9
Single-Institution/ Multi-Institution	85.6	77.4	57.4
Multi-Institution/ Multi-Institution	99.3	97.7	79.3

2.3.3 Sensitivity to small shifts

When the model trained on data from Institutions 2-6 was applied to the 12 patients from UCLA with clinically insignificant misalignments, the false positive rate did not differ appreciatively from that seen when it was applied to the clinical (no error) images. For these 12 patients, the false positive rate was 3.2% on the unshifted images, 4.0% on the images shifted by ±1 pixel, 5.6% on the images shifted by ±2 pixels, and 6.9% on the images shifted by ±3 pixels.

The model flagged 19.4% of the images shifted by 5 pixels (approximately 1mm) and 51.2% of the images shifted by 8 pixels (approximately 1.5 mm) as misaligned. It is worth noting that in many institutions, a shift of 1 mm is considered to be the upper end of the typical tolerance used for these high-dose stereotactic spine regimens. These results are shown in **Figure 2.7**.

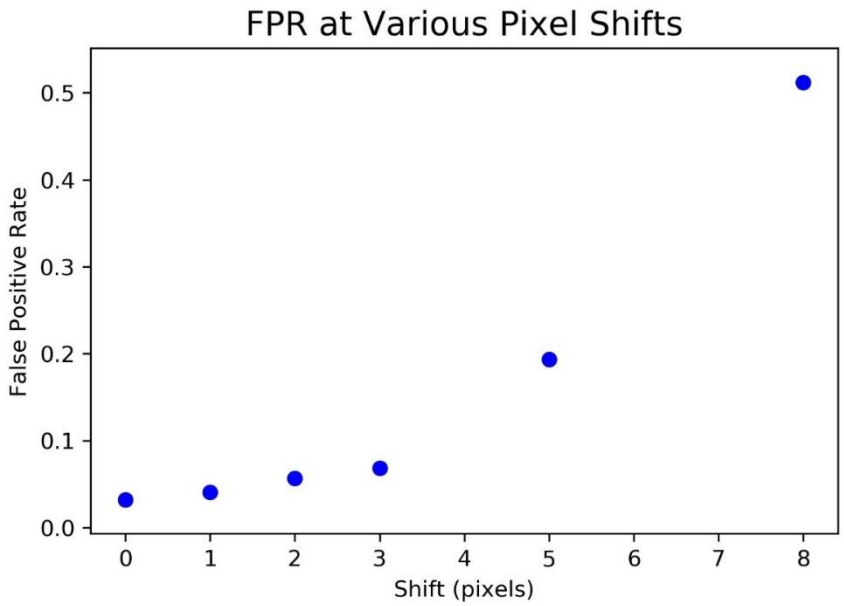

Figure 2.7: False positive rate at five different pixel shifts of interest. Of note, shifts of 5 pixels correspond to approximately 1 mm, a tolerance commonly used in clinical practice.

2.4 Discussion

Based on the results presented above, application of this error detection model to data from an unseen institution would result in an error detection sensitivity of at least 86%-100% (mean: 93%) if a false positive rate of approximately 5% was accepted, or a sensitivity of 62%-99% (mean: 78%) if only a 1% false positive rate was accepted. We believe these results demonstrate sufficient accuracy to warrant clinical implementation of this error detection model. The corresponding effort to deal with the level of false positives flagged by this model is small compared to other safety processes already in use and would be expected to reduce IGRT errors

by approximately a factor of 5. If up to 5% false positive rate was allowed, which could be acceptable given the relative infrequency of spinal radiotherapy (particularly high-risk stereotactic regimens), then errors could be reduced by a factor as large as 10.

However, further improvements in accuracy would allow this error detection algorithm to truly run in the background and only disturb the workflow in case of an actual error. Because relatively few thoracic spine patients are treated with ExacTrac at any one of our institutions in a given year, the size of our training dataset was relatively small in comparison to the truly big data such as ImageNet[150], a database containing over 10 million images which is used to train state-of-the-art image classification models. Further increasing the amount of training data, as well as the number of institutions involved, could be expected to improve upon the accuracy and robustness of our error detection model. Others have proposed the use of transfer learning to adapt models trained on ImageNet to the medical domain where high-quality training data is more limited[151]. We found in our specific implementation that transfer learning was not as effective as training from scratch, but this avenue warrants further exploration. Finally, while the stereoscopic imaging system generates a set of two oblique x-ray images, for the purposes of this work we considered each of these images independently. This limitation arose due to challenges in simulating vertebral body misalignments that preserve geometric correlations between the two stereoscopic images. **Chapter 3** focuses on a method for treating the two x-ray images and the associated DRRs as a single, interdependent image set, potentially increasing the error-detecting power of the model.

We incorporated multi-institutional data into the development of our error-detection model because we anticipated that it would lead to better accuracy on data from unseen

institutions. This expectation was validated by the results of our "leave one institution out" methodology. We can infer the presence of qualitative differences in ExacTrac image data from participating institutions by differences in algorithm sensitivity of as much as 14% at a 95% specificity threshold. All institutions utilize similar default imaging parameters for their thoracic spine patients (25 mAs and 120 or 130 kVp). A trend towards higher sensitivity for institutions with the higher default kVp value was observed, but no statistically significant correlations were observed. We were not able to identify obvious visual features in the data samples that might explain those differences. This is highlighted in **Figure 2.8** showing some representative false positive images flagged by the different "leave one institution out" models at the 95% specificity threshold for each model. We computed the distributions of the average x-ray pixel values for each institution, and found that the distribution for Institution 2 differed significantly from the other institutions (p-value < 0.05) when a Mann-Whitney U test was performed. The leave-one-out model where Institution 2 was left out of the training dataset and used as the test set achieved the highest performance, and the statistically significant difference in average x-ray pixel values suggests a real difference in the images. However, the exact nature and underlying cause of that difference is unknown and beyond the scope of the current work. Possibilities include imager performance, differences in patient populations, or a different mix of spine levels treated at each institution. At this time, all participating institutions are academic and/or high-volume medical centers that are considered expert institutions in spine radiotherapy. These data may not be representative of ExacTrac image data at all clinics in the radiotherapy community, and the model performance has not been tested on data from institutions that may image using suboptimal imaging techniques (improper mAs or kVp, for example).

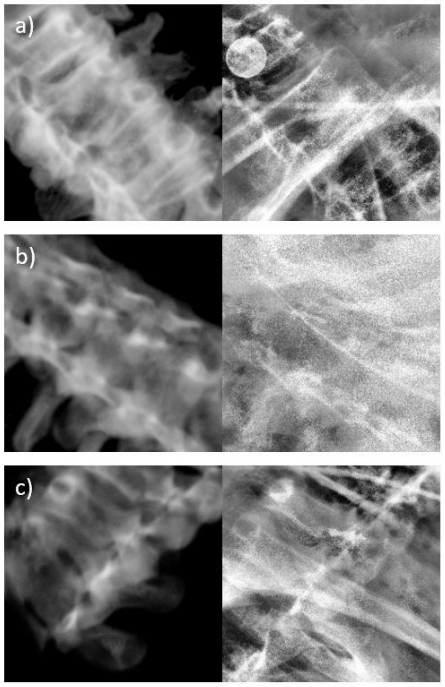

Figure 2.8: Some representative false positive x-ray/ DRR pairs from Institution 2 (a), Institution 4 (b), and Institution 5 (c). DRRs are shown on the left of each image, and the corresponding x-ray is shown on the right. Each image pair was incorrectly classified by the model trained using data from all institutions except the one from which these images came.

The power of automation to promote radiotherapy safety and efficiency is widely recognized, with most research to date focusing on automation in machine quality assurance[152], treatment planning[153] and plan evaluation[154], and auto-contouring[155]. However, little has been done to incorporate automation and machine learning into the task of image review, which is a function that requires a significant amount of effort from radiation oncologists, physicists and technologists in the IGRT workflow[14,156]. We believe our work adds significantly to this effort.

2.5 Conclusion

In this work we have developed a convolutional neural network-based model for the automatic detection of vertebral body misalignments in planar x-ray setup images. We believe such an algorithm could be integrated into the treatment workflow, either directly within an

image-guided radiotherapy system or as a standalone ancillary system, to provide a software interlock to prevent treatment of the incorrect vertebral body. Our results demonstrated that this misalignment detection model is robust when applied to previously unseen test data from an outside institution, indicating that this proposed additional safeguard against misalignment is feasible.

CHAPTER 3: INCORPORATING EXACTRAC'S STEREOSCOPIC GEOMETRY INTO A TOOL FOR DETECTING VERTEBRAL BODY MISALIGNMENTS

3.1 Introduction

3.1.1 ExacTrac's stereoscopic geometry

The ExacTrac onboard imaging system is combined with a six degree of freedom treatment couch that is able to shift the patient in the standard three translational directions, with the addition of pitch, roll, and yaw rotational shifts. Phantom studies have shown that the combination of IGRT with a six degree of freedom couch leads to improved positioning accuracy over a traditional couch utilizing only translational shifts[39]. Furthermore, the ExacTrac system can detect rotational setup errors with a high level of accuracy, enabling the patient to be reliably positioned within 1 degree in all three rotational directions[157]. When this type of couch is installed clinically, it has been observed that all six axes are used extensively for patient setup during routine radiation therapy treatments[158].

In **Chapter 2** we developed an automated error detection algorithm trained on simulated off-by-one vertebral body errors that were generated within the imaging plane of the 2D clinical DRRs. However, treating each of the two x-ray/ DRR sets independently may lead to artificially inflated model performance, as the training data does not truly replicate errors that would be seen clinically. Detailed descriptions of the unique geometry of the ExacTrac system can be found in the literature[37,38] and are summarized in **Chapter 1**, but it is worth emphasizing here that the two stereoscopic x-rays and their associated DRRs are correlated. This means that generating synthetic off-by-one vertebral body errors exclusively within the imaging plane, as was done in

Chapter 2, is not geometrically the same as incorrectly aligning the patient and re-generating the new set of oblique x-rays.

3.1.2 Study overview

In **Chapter 2** the two stereoscopic x-rays were treated independently, and the error generation was limited to 2D translational shifts within the confines of the image plane. While this is crucial groundwork, a translation in the imaging plane by one vertebral body is not geometrically the same as a true misalignment by one vertebral body in the patient. Furthermore, internal tests from Brainlab showed that monoscopic image matching results in a higher registration failure rate than that obtained by utilizing the stereoscopic view. Given these limitations, in this chapter we developed a method of generating synthetic shifts in order to more closely match the images that would result from a true off-by-one vertebral body shift. The work we present here builds upon the work done in **Chapter 2** by correcting a geometric inaccuracy present in our earlier generation of training images, and improves the error detection accuracy of our automated tool in detecting off-by-one vertebral body shifts that are more representative of true clinical cases.

3.2 Materials and methods

3.2.1 Data collection

Archived clinical patient datasets were identified from the years 2014-2017 for a single treatment machine at UCLA where ExacTrac image guidance was used for patient alignment. We searched the prescribed treatment plan names for all patients to identify only those that were treated to the thoracic spine. This was possible because of the adherence to standard plan naming

conventions present in the treatment prescription documents for all patients. 83 unique patient datasets were identified across the four years, totaling 317 treatment fractions. The final dataset consisted of 634 day-of x-ray images (two from each treatment fraction), where the first set of images acquired for each given treatment fraction was extracted for our purposes. The detailed breakdown by year is shown in **Table 3.1**. In addition, we extracted the simulation CT DICOM files for each patient, which were used to generate both aligned and simulated misaligned DRRs corresponding to each set of day-of x-rays, as described in the DRR generation section that follows.

Table 3.1: Number of patients, unique treatment fractions, and setup x-rays collected from each year.

	2014	2015	2016	2017	**Total**
Patients	18	18	27	20	**83**
Treatment fractions	85	39	90	103	**317**
X-rays	170	78	180	206	**634**

3.2.2 DRR generation

We combined the functionalities of the ExacTrac offline Replay with an open-source DRR generator to produce two geometrically-realistic off-by-one vertebral body misalignments for each unique treatment fraction (**Figure 3.1**). This methodology overcomes the primary limitation made apparent by our earlier work; namely, that treating the two stereoscopic images independently does not properly replicate real patient misalignments likely to be seen during radiation therapy treatment. By generating misaligned data in this way, we were able to preserve the interdependent nature of the stereoscopic x-rays. An additional advantage of this method for shifted DRR generation was that it eliminated the need for cropping of the final images, which likely preserves image information that may be of benefit during algorithm training.

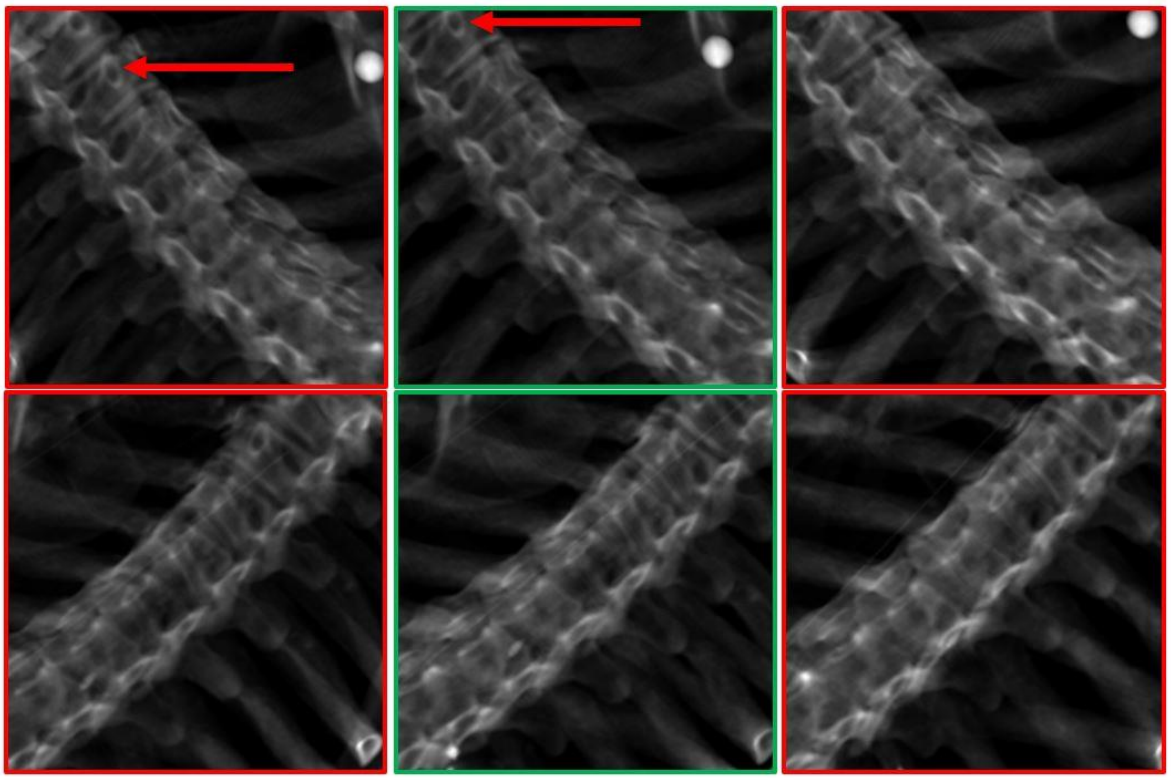

Figure 3.1: Correctly aligned DRRs (green boxes) and synthetically misaligned DRRs (red boxes) for an example treatment fraction. Close examination of the images, for example focusing on the feature indicated by the red arrows, confirms the successful misalignment by one vertebral body.

Unlike most imaging modalities used in the radiation oncology clinic, the ExacTrac workstation has a window that allows for retrospective review of patient cases and replay of the image alignment performed at the time of treatment. While this offline review does allow for the manual simulation of off-by-one vertebral body misalignments, it does not currently support the export of the corresponding DRRs. Using the Replay function of the ExacTrac workstation, a user is able to manually misalign day-of patient imaging. Once this manual misalignment has been performed, the system can be forced to find the optimized patient alignment using the automatic registration function. After this optimization has taken place, the corresponding shift coordinates are reported within the workstation (**Figure 3.2**).

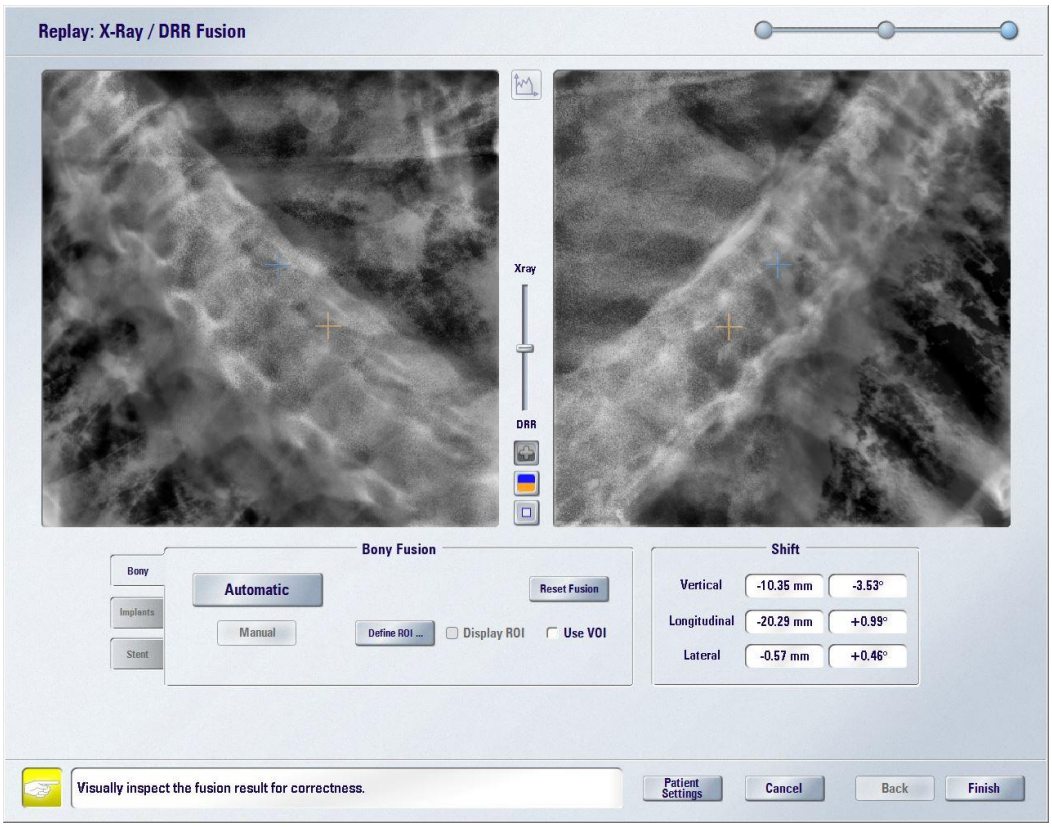

Figure 3.2: Shift coordinates obtained by manually misaligning the patient by one vertebral body and forcing the ExacTrac system to find the optimal registration between the x-rays and misaligned DRRs.

Our group has recently published an open-source tool for the offline generation of DRRs that takes as input a patient's simulation CT and any specified treatment couch shifts[159]. This algorithm uses the libraries from the Insight Segmentation and Registration Toolkit (ITK)[160] to generate geometrically-realistic DRRs based on any arbitrary shift of the patient's 3D simulation CT imaging. Importantly, the algorithm incorporates six degrees of freedom into the image generation parameters—three-dimensional translations, plus the pitch, yaw, and roll rotations. Once these shifts are applied to the 3D CT, a new interdependent set of stereoscopic DRRs can be simultaneously generated.

3.2.3 Data organization

4-channel arrays consisting of two x-rays and the two corresponding DRRs (shifted or non-shifted) were then created out of the final set of generated DRRs from 2014-2017. The final training dataset consisted of 951 such 4-channel arrays, with each treatment fraction contributing an array with non-shifted DRRs and two arrays with shifted DRRs (both inferior and superior along the spinal column). This dataset was split into training and test datasets using an approximately 80/20 training/ test split. This division occurred at the patient level, and the patients included in each dataset were identical to those included in the training and testing datasets from the single-institution model described in **Chapter 2**. We purposely included the same patients in each respective dataset here in order to facilitate an unbiased model comparison. The training dataset was randomized and then further divided into final training and validation datasets using a 75/25 split.

3.2.4 Model design and training

Treating the two stereoscopic images as one interdependent image set leads us to propose the use of a multi-input network for this work. Multi-input networks have shown excellent performance in image classification tasks, even in relatively small datasets comparable in size to our own[161]. We investigated various depths of network architectures and found that optimal classification results were achieved using the five convolutional block multi-input neural network architecture shown in **Figure 3.3**.

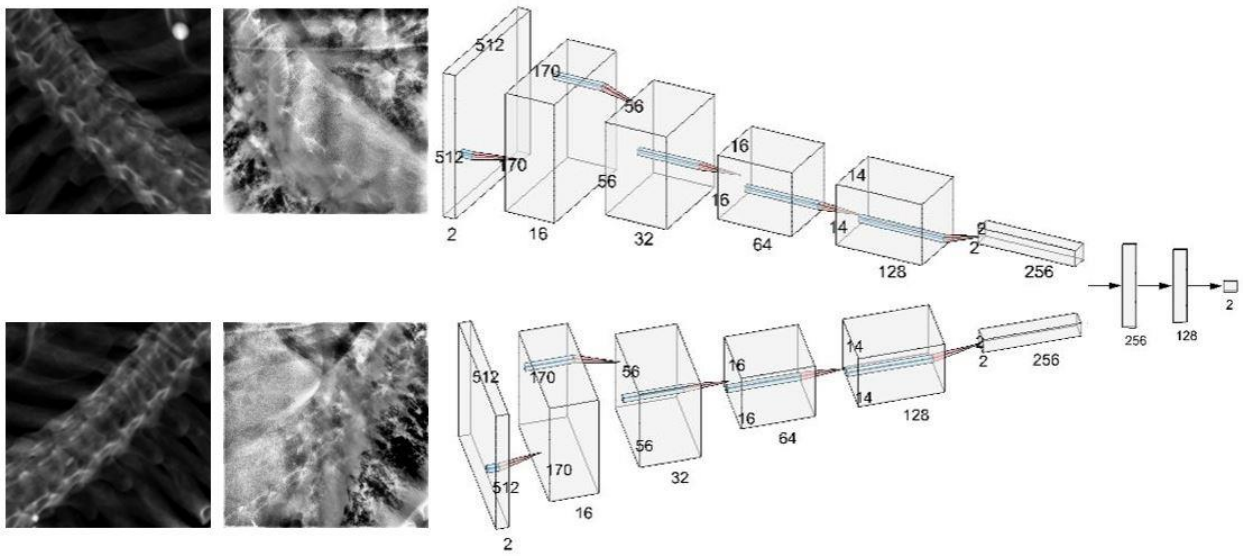

Figure 3.3: Multi-input CNN architecture for detecting off-by-one vertebral body misalignments using the full stereoscopic image set consisting of two x-rays and two corresponding (unshifted or shifted) DRRs from each treatment fraction.

The five convolutional layers were applied independently to each of the two corresponding DRR/ x-ray image sets within the larger 4-channel array. These separate convolutional layers were then merged into a single layer, with subsequent dense and dropout layers applied to the merged data. Convolutional layers in our model architecture were followed by rectified linear activation functions and batch normalization layers. Two dense layers, each of which was followed by a 50% dropout layer to reduce overfitting[143], were used before the final classification layer. The learning rate was set to 1e-4[144]. The Adam optimizer[145] and categorical cross-entropy loss function were used for training. Early stopping was implemented during model training, again with the aim of reducing overfitting. The dataset reserved for validation was used to evaluate the model during the training process and update layer weights at the end of each training epoch.

3.2.5 Model analysis

An ROC analysis was performed by applying the trained model to the unseen testing dataset and calculating the AUC as a measure of overall model accuracy. We compared the model performance on the holdout testing dataset to that obtained by the single-institution model trained and evaluated in **Chapter 2**. This was done to evaluate whether the use of interdependent image sets during model training leads to an improvement in the final model performance, specifically increased sensitivity at our specificity thresholds of interest. Sensitivity results for three specificity values of potential clinical interest (99%, 95%, and 90%) are reported here. Our model output is a continuous variable, ranging from 0 to 1, for each 4-channel image input. Clinical implementation of this model would ultimately require applying a threshold cutoff value for flagging images based on the desired specificity-sensitivity tradeoff.

3.3 Results

3.3.1 Comparison to previous model

When the final model trained to detect vertebral body misalignments in the interdependent set of stereoscopic images was used to classify the 20% of data reserved for testing, the resulting AUC was 0.988. In comparison, the final model from **Chapter 2**, tested on independent planar images from the same patients, achieved an AUC of 0.975. The ROC curves from both models are shown below in **Figure 3.4**.

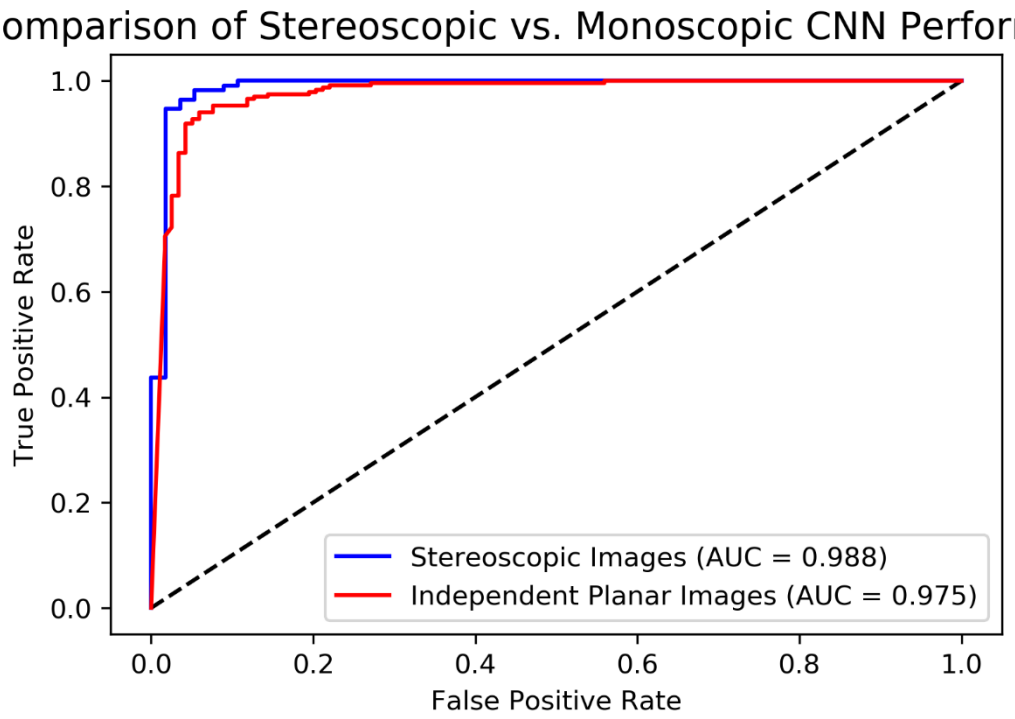

Figure 3.4: Comparison of classification performance in correctly identifying shifts of one vertebral body between the model trained and tested on independent planar image sets (monoscopic) and the model trained and tested on the interdependent stereoscopic image set.

With the specificity fixed at 99%, the CNN trained on the stereoscopic image sets achieved a sensitivity of 43.8% in correctly identifying off-by-one vertebral body shifts. When the specificity was decreased to 95%, the corresponding sensitivity increased to 96.4%. **Table 3.2** highlights the differences in sensitivity between the two models at three different specificity thresholds.

Table 3.2: Comparison of sensitivity-specificity tradeoffs between the model trained and tested on independent planar image sets in **Chapter 2** and the model trained and tested on stereoscopic image sets discussed in detail in this chapter. Shaded cells indicate the model with the higher sensitivity for each given specificity.

	% Sensitivity at 90% Specificity	% Sensitivity at 95% Specificity	% Sensitivity at 99% Specificity
Stereoscopic Image Sets	99.1	96.4	43.8
Independent Planar Image Sets	95.3	91.9	67.9

3.3.2 Detection of known errors

Two patients, one treated in 2017 and the second treated in 2023, were previously identified by members of the UCLA radiation oncology team as having been misaligned by one vertebral body during certain treatment fractions. Retrospective physician chart checks identified the patient in 2017, and a combination of physics oversight at the treatment machine with automated tools to aid in IGRT review identified the misaligned and nearly misaligned fractions from 2023. The final trained model was applied to the five total fractions having been identified as containing off-by-one vertebral body misalignments. When the threshold corresponding to 95% specificity on the testing dataset was used, the model succeeded in flagging all five of the fractions. Images from one of the treatment fractions known to contain an off-by-one vertebral body misalignment are shown in **Figure 3.5**. A close manual review identified the error based on the misaligned position of the surgical clip, indicated by the red arrows.

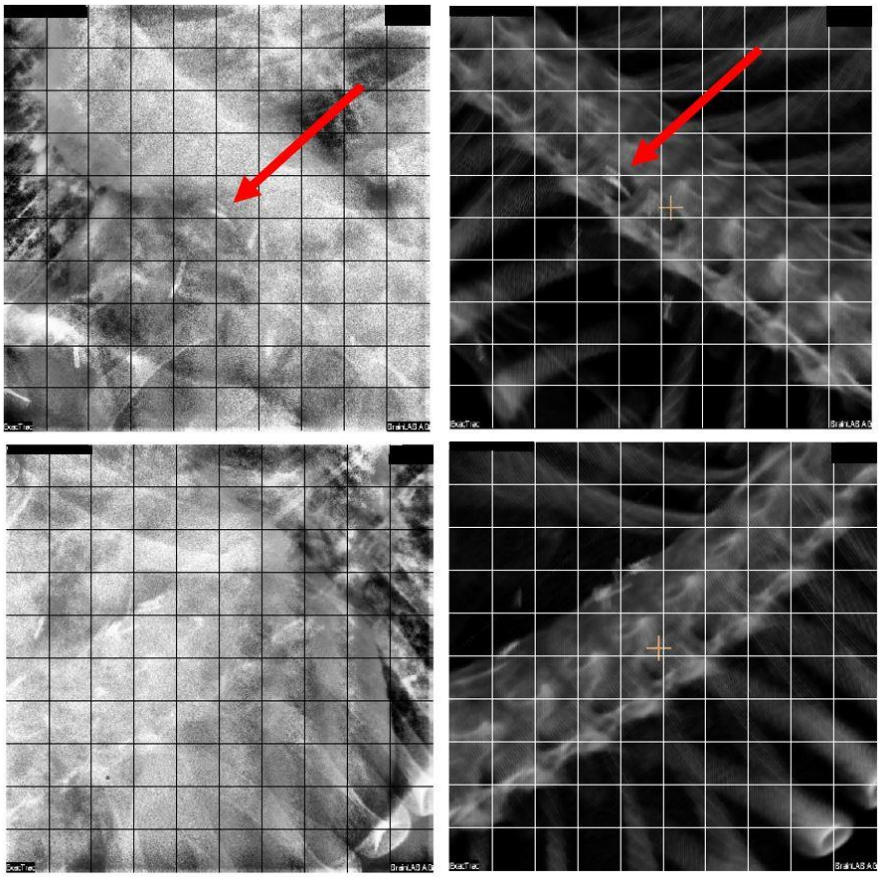

Figure 3.5: Day-of x-rays (left) and DRRs (right) misaligned by one vertebral body for a patient treated at UCLA. The presence of the surgical clip, indicated by the red arrows, ultimately led to the identification of this treatment error.

3.4 Discussion

Our results show that the application of this trained stereoscopic model to clinical data would result in an error detection sensitivity of over 96% when a specificity of 95% is used. Considering the high risk of high-dose stereotactic regimens commonly employed to treat targets in the thoracic spine region, we believe this specificity-sensitivity tradeoff is acceptable for clinical use. The level of effort required to deal with any false positives resulting from the application of our model to clinical data is low enough to be deemed reasonable, and well worth the additional time spent to prevent gross misalignments. In addition, we have demonstrated that

training our model using clinically realistic, interdependent stereoscopic image sets does not degrade the model's performance from that observed when the model was trained using independent planar image sets. This is a key finding, as application of this automated error detector clinically would ultimately require the ability to seamlessly apply the trained model to the full set of images generated during the course of a patient's treatment. Finally, the ability of our trained model to successfully flag real patient images from all of the five fractions previously known to represent true off-by-one misalignments bolsters the clinical reliability of this method. Had this automated error detector been implemented into the clinical workflow at the time of these specific patient treatments, these rare but serious errors could potentially have been avoided.

As we go from treating the stereoscopic images independently (as was done in **Chapter 2**) to treating them as one single interdependent set, we essentially cut our raw amount of training data in half. Our model's classification accuracy and robustness would both likely be improved by the incorporation of training data from more patients, and specifically patients from multiple institutions, as was shown with our earlier work[162]. In **Chapter 2**, we demonstrated that the incorporation of training data from six institutions improved the AUC to 0.992, as compared to an AUC of 0.975 when the model was trained on data from only a single institution. We would expect to see similar results for the current model trained on stereoscopic image sets. Future work should focus on the incorporation of more image data, particularly image data from a variety of medical centers, into the training dataset.

Based on a conversation with our industry contact at Brainlab, the results of the monoscopic image matching may appear to perform better for real patient cases than the

stereoscopic image matching. This could be due in part to the incorrect spatial information in the 2D independently translated DRRs generated in **Chapter 2**, making the classification job of the model much easier to perform. The value of the work proposed in this chapter primarily lies in the generation of more clinically-accurate data used to train our model. A model that could potentially be acceptable for eventual clinical use must have good discriminatory power on the interdependent stereoscopic images and be able to accurately classify errors that involve a complex and subtle mix of translations and rotations.

3.5 Conclusion

In this work we have developed an updated method for generating off-by-one vertebral body misalignments that are more representative of how such errors would appear in real clinical situations. Our results show that the use of geometrically-realistic image data for CNN training improves the final model performance on unseen testing data, indicating that this updated model has clinical potential in detecting real off-by-one vertebral body errors in patient images.

CHAPTER 4: DEVELOPMENT OF A CONVOLUTIONAL NEURAL NETWORK TO DETECT TRANSLATIONAL PATIENT SETUP ERRORS

4.1 Introduction

4.1.1 Time requirements of IGRT review

Detecting and preventing patient setup errors is of the utmost importance in the radiation oncology clinic. **Chapter 1** highlights some of the clinical implications resulting from incorrect patient setup, primarily decreased dose to the target and worsened normal tissue sparing. Image guidance has proven to be a reliable safeguard against errors stemming from incorrect patient setup, but can provide false reassurance if used incorrectly[22]. The American Society for Radiation Oncology highlighted this limitation, cautioning its members that IGRT must be deployed in a safe manner, with robust quality assurance protocols in place, in order for patients to ultimately benefit from this new technology[35]. Clinical medical physicists are vital to the safe deployment of IGRT in the clinic, and are responsible for both the acceptance testing of new equipment and the development and implementation of quality assurance protocols[163].

Even when IGRT is implemented safely and reliably, the review and interpretation of patient images requires a significant amount of effort from all members of the radiation oncology team[14,156]. Automated methods have been successful in making many aspects of the radiation therapy workflow more efficient, but the quality and safety applications[129,164,165] are of particular relevance to our own work. Here we propose the use of a convolutional neural network-based tool to efficiently analyze patient setup images for egregious misalignments that can potentially lead to significant harm. CNNs have previously shown excellent performance on various other

medical image analysis problems[90–92], more recently including automated IGRT review applications from our group[162,166–168].

4.1.2 Study overview

Due to the significant time required for a thorough review of patient setup images, our group proposed that an automated image review algorithm could be developed to detect gross patient misalignments in setup images. Here we introduce a novel approach for automatically detecting patient misalignments in ExacTrac images from all anatomical treatment sites using a CNN trained on 1 cm translational shifts. A shift of 1 cm is large enough to negatively impact a patient's treatment, but small enough that it could potentially be missed by the treating radiation therapist or overseeing physician at the treatment console. Our goal in this work is to present an automated method for detecting patient setup errors in ExacTrac images, ultimately allowing for increased efficiency in the currently time-consuming IGRT image review process.

4.2 Materials and methods

4.2.1 Training data collection

Archived clinical patient datasets were initially identified from the years 2014-2017 for a single treatment machine at UCLA where ExacTrac image guidance was used for patient alignment. We first excluded a small number of patients in each year of training data who did not have at least one nominal beam angle of 0 degrees. During our step of generating aligned and synthetically misaligned DRRs (described in the Generation of simulated DRRs section that follows), we discovered that patients without a treatment beam at 0 degrees were likely to be set up initially with the treatment couch positioned at the angle of the first beam. Since the ExacTrac

system expects the couch to be initially positioned at 0 degrees for imaging before being moved to the appropriate angle, this introduced large unexpected shifts that could not be accounted for in our DRR generation step. Following exclusion of these patients from our dataset, 2,407 total patients remained.

We next created a filter based on the prescribed treatment plan name to exclude any patients where the daily image guidance specified final patient alignment to either soft tissue or implanted fiducial markers. Since the ExacTrac system was only used for initial patient positioning for these treatment plans, we expect to see initial misalignments in the ExacTrac images that are corrected with subsequent imaging (typically cone beam CT or, less often, kilovoltage x-ray) and should therefore be excluded from our training data. The specific plans excluded under these criteria, along with the final number of patient datasets, are shown in **Table 4.1**.

Table 4.1: Number of patients excluded from each year, along with the final number of patient datasets collected for our training dataset from each year.

	2014	2015	2016	2017	**Total**
Initially identified	419	640	789	559	**2,407**
Treatment plan name					
Prostate	12	11	28	24	**75**
Lung	69	92	98	91	**350**
Rib	4	2	9	8	**23**
Abdomen	3	10	7	1	**21**
Liver	14	7	4	2	**27**
Pancreas	3	3	1	1	**8**
Adrenal	2	2	2	2	**8**
Bladder	2	0	0	0	**2**
Esophagus	0	2	0	0	**2**
Heart	0	0	0	0	**0**
Final number of patient datasets	**310**	**511**	**640**	**430**	**1,891**

Day-of x-ray images were extracted from all 11,266 treatment fractions (from the 1,891 unique patient datasets remaining after the application of our exclusion filter). In addition, we

extracted the unique simulation CT DICOM files for each patient. These files were used to generate our training dataset, with the details of the process described in the Generation of simulated DRRs section below.

4.2.2 Generation of simulated DRRs

Our colleagues have published an open-source tool for the offline generation of DRRs that takes as input a patient's simulation CT and any specified treatment couch position[159]. Details can be found in **Chapter 3**, but it is worth emphasizing again that the algorithm incorporates all three translational parameters and all three rotational parameters into the image generation, in order to accurately replicate the six degrees of freedom present in the ExacTrac system. Once these shifts are applied to the 3D CT, a new interdependent set of stereoscopic DRRs can be simultaneously generated.

We used this open-source DRR generator to generate new sets of interdependent DRRs to be used as training data for our model. The physician-approved patient shifts were used for generating the "no error" data in the training dataset of all images from 2014-2017. Random translational shifts of 1 cm from the treatment isocenter were used to generate our "error" data. One centimeter was chosen because this represents a shift large enough to potentially be clinically significant, but small enough that it is reasonable to believe such a shift might not be noticed by the treating radiation therapist or physician at the treatment console. Once these shifts in the translational directions were randomly selected, they were used in conjunction with the

DRR generator to create both a new set of stereoscopic “no error” and “error” DRRs for each treatment fraction (**Figure 4.1**).

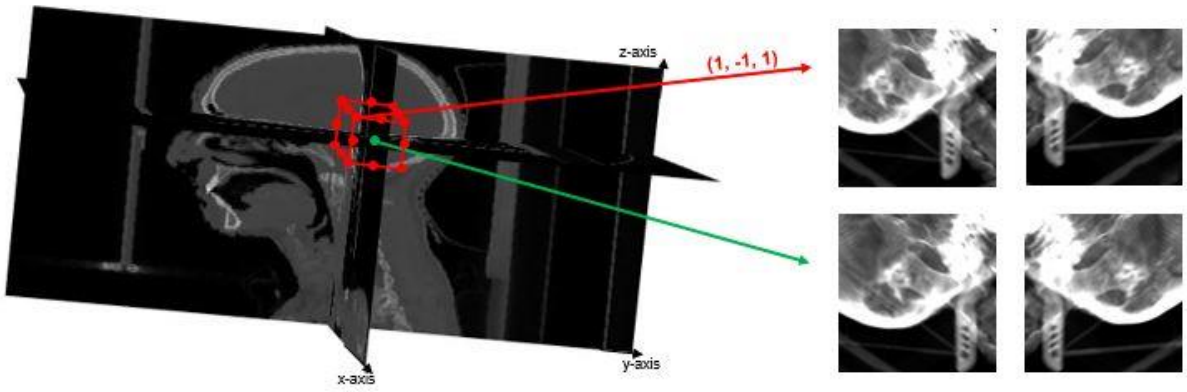

Figure 4.1: Aligned (green arrow) and simulated misaligned (red arrow) DRRs generated from a representative patient's simulation CT scan.

4.2.3 Data organization

4-channel arrays consisting of two x-rays and the two corresponding DRRs (shifted or non-shifted) were then created out of the final set of generated DRRs from 2014-2017. The final training dataset consisted of 22,532 such 4-channel arrays, since each treatment fraction contributes both an array with non-shifted DRRs and an array with shifted DRRs. This dataset was split into training, validation, and test datasets for model training and testing using an approximately 80/10/10 training/ validation/ test split. The splits occurred at the patient level, so that for patients with multiple treatment fractions, images from all fractions would only appear in one dataset.

4.2.4 Model design and training

Optimal classification results were achieved using the multi-input neural network architecture shown in **Figure 4.2**. Multi-input networks have shown excellent performance in classification tasks where multiple image inputs are present[161], and the interdependent nature of the ExacTrac images lends itself well to this architecture. Networks of various depths were investigated since increasing CNN depth can improve final classification accuracy[142] when trained on large datasets, but can also increase the possibility of overfitting when used with relatively small datasets such as our own.

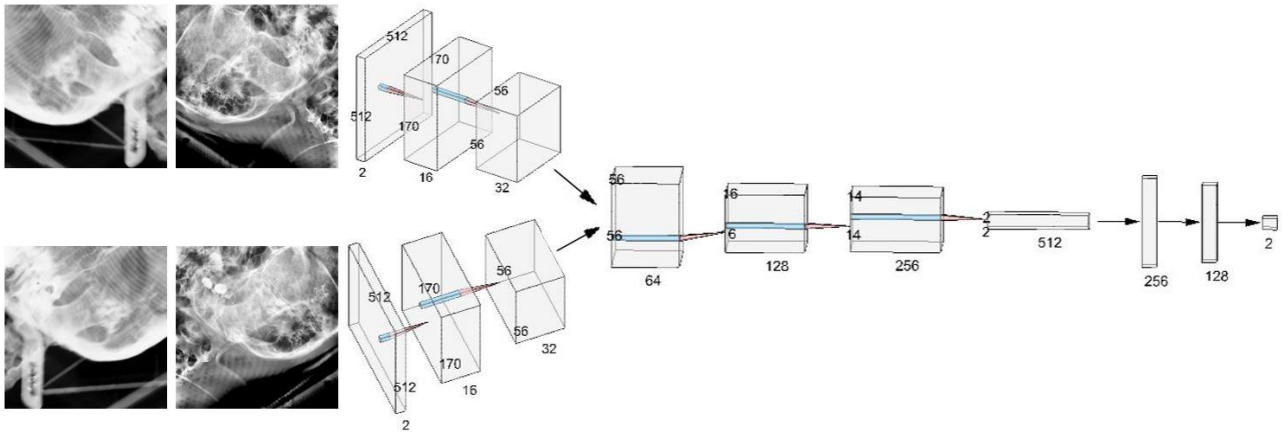

Figure 4.2: Multi-input CNN architecture for classifying 4-channel x-ray/ DRR arrays as aligned or misaligned.

The convolutional layers C1 and C2 were applied independently to each of the two corresponding DRR/ x-ray image sets within the larger 4-channel array, followed by a max pooling layer. These two separate convolutional layers were then merged into a single layer. Subsequent convolutional and max pooling layers were applied to the merged data. Convolutional layers in our model architecture were followed by rectified linear activation functions and batch normalization layers. Two dense layers, each of which was followed by a 50% dropout layer to reduce overfitting[143], were used before the final classification layer. The

learning rate was set to 1e-4[144]. The Adam optimizer[145] and categorical cross-entropy loss function were used for training. Early stopping was implemented during model training, again with the aim of reducing overfitting. The 10% of the dataset reserved for validation was used to evaluate the model during the training process and update layer weights at the end of each training epoch.

4.2.5 Data analysis

After our model was finished training, the ROC curve was used to evaluate its performance on the holdout testing dataset. The AUC was calculated and used as a measure of overall model accuracy. Like the model for detecting vertebral body misalignments, this model also outputs a continuous variable in the range from 0 to 1. Clinical implementation of this model as a real-time alert in the IGRT workflow would require a determination of the optimal specificity-sensitivity tradeoff and application of the corresponding threshold. Sensitivity results for three potential specificity values of interest (99%, 95%, and 90%) are reported here.

4.3 Results

4.3.1 ROC analysis

When the final model trained to detect 1 cm translational misalignments was used to classify the previously unseen images from the 10% of the dataset reserved for testing, the resulting AUC was 0.970 (**Figure 4.3**). With the specificity fixed at 99%, this trained CNN achieved a sensitivity of 55.6% in correctly classifying translational shifts of 1 cm (**Table 4.2**). While this high specificity is desirable in order to reduce alarm fatigue in the clinic[146,147], it must be balanced with a reasonably high sensitivity in order to successfully flag patient alignment

errors. Incorrect patient setup has been shown to be an important factor in radiation oncology treatment errors[59,60], and successfully detecting such errors is of high interest to the radiation oncology community. When the specificity was lowered to 95%, the corresponding sensitivity increased to 94.7%, indicating a drastic improvement in the model's ability to detect patient misalignments at this specificity threshold.

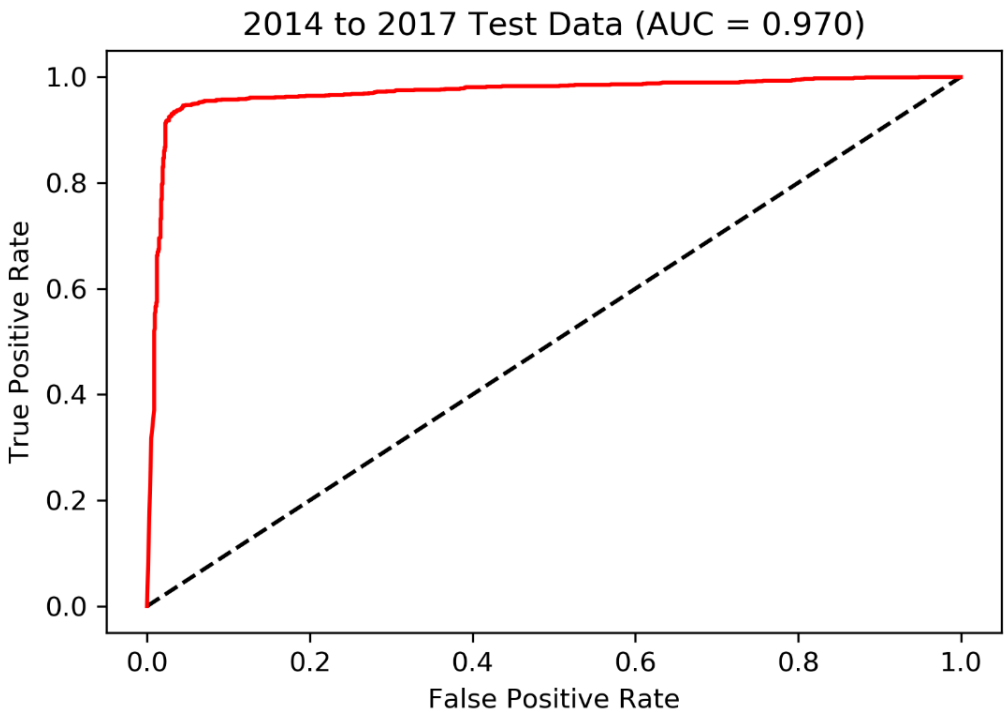

Figure 4.3: Model classification performance in correctly identifying 1 cm translational shifts in the unseen test dataset.

Table 4.2: Three specificity-sensitivity tradeoffs of potential clinical interest.

Specificity	Corresponding sensitivity
99%	55.6%
95%	94.7%
90%	95.8%

4.4 Discussion

Based on the results presented here, application of our automated error detection tool to clinical data from UCLA would result in an error detection sensitivity of approximately 95% if a 5% false positive rate was accepted. We believe these results are sufficient to warrant incorporation of this tool into the clinical workflow, either as a real-time check of clinical images or as an aid in daily offline image review. The level of effort required to deal with the false positives flagged by the model is small in comparison to the processes and procedures already in use clinically, especially when considering the types of gross misalignments preventable through implementation of this tool.

We are cognizant of the limitations of this IGRT error detection algorithm. While our dataset of over 22,000 training images is relatively large in comparison to other studies involving medical image data, it is still small in comparison to true "big data" such as ImageNet[150], which contains over 10 million images and has been used to train some of the highest performing classification models. Increasing the amount of training data would be expected to further improve our model's classification accuracy. In addition, the image data used for training came from a single institution. Studies have shown that assembling a multi-institutional collaboration of image data is key for improving model robustness[162,166]. Incorporating data from multiple institutions would be a crucial next step in improving robustness, ultimately allowing us to apply our model to image data from outside institutions and have confidence in the results. Finally, while our model demonstrated excellent performance in detecting translational shifts of 1 cm, we recognize that this is far from the only type of patient setup error that could realistically occur in the clinic. For example, other work from our group has focused on the specific problem of

detecting off-by-one vertebral body misalignments in patient setup images. Rotational shifts were not incorporated into the generation of simulated errors for the purposes of this work, even though this is another type of error that could potentially occur in the clinic. Other errors, such as wrong patient or wrong anatomical site, were likewise not investigated.

Research to date regarding the incorporation of automation into the radiation oncology clinical workflow has shown promising results, with success in automating certain tasks at the contouring, treatment planning, plan quality evaluation, and machine quality assurance stages. However, the automation of image review tasks has lagged significantly. Image review requires a great deal of time and effort to be expended from an already overworked radiation oncology team. We believe that our work here shows great promise in beginning to automate parts of this task and allowing for cognitive resources to be more efficiently directed towards interpreting the results of such a tool.

4.4 Conclusion

In this work we have developed and validated a convolutional neural network-based model that enables the automatic detection of general translational errors in ExacTrac images that have the potential to be clinically significant. We believe such an algorithm could enable a more thorough review of patient setup images, either at the time of treatment or during retrospective image review checks, by providing a fast, automated method for flagging potential misalignments. Our results demonstrated that this misalignment detection model achieves a high sensitivity at the 95% specificity, indicating that this proposed tool has clinical potential in successfully identifying true misalignments with a correspondingly low level of algorithmic false positives.

CHAPTER 5: A CONVOLUTIONAL NEURAL NETWORK-BASED RETROSPECTIVE SEARCH FOR PREVIOUSLY UNREPORTED RADIATION EVENTS

5.1 Introduction

5.1.1 Promise and limitations of large imaging databases

The rapid acceleration of technology in radiation oncology combined with the existing demands on the clinical medical physicist's time means that already limited physicist time resources must be allocated thoughtfully[83]. One specific area within the medical physicist's domain that presents both a promise and an enormous challenge is that of image review. With the vast majority of radiation treatments now using some form of image guidance[13,14], a wealth of visual data is generated every day. These archival clinical imaging databases present an opportunity to leverage the aggregate data in order to drive evidence-based improvements in clinical safety interventions[169]. Some of the challenges associated with big data initiatives such as this one are not new to the medical physicist. Articles in recent years have highlighted the logistical challenges of data storage[170], management and processing of time-sensitive health care data[171], and translation of big data into impactful clinical tools[172] as significant barriers within radiation oncology. With respect to image review, an additional challenge becomes apparent: the sheer size of most imaging databases within radiation oncology departments are simply so large that a manual review of every image by a physicist is infeasible. This is especially true in light of research demonstrating that medical physicists experience the highest workload out of all members of the radiation oncology team[173], coupled with a high level of reported burnout[174,175]. Thus, new tools must be developed in order to support the analysis of large imaging databases

and the subsequent advance of the department's incident learning system. The integration of voluntary reporting systems such as RO-ILS have been shown to lead to improved identification of the clinical areas where safety improvements are necessary[176]; however, they are limited since unknown errors have no way of being reported or documented. Analysis of large imaging databases represents an opportunity for large-scale image analysis for previously unreported and undocumented incidents.

5.1.2 Automation and incident learning systems

There is a pressing need for new tools to be developed to help support the efforts of incident learning systems, including data mining of big datasets[82]. Machine learning tools in particular have shown promise in many aspects of the radiation oncology clinical workflow, including a variety of applications in quality and safety initiatives[177]. Many of the review tasks currently performed by humans in radiation oncology have the potential to be streamlined by the development of novel automated tools[178]. Furthermore, machine learning methods are beginning to show potential in detecting when something (such as an organ contour) may be suboptimal[179]. Such tools, in combination with the big databases already present in radiation oncology, offer new opportunities for detecting and addressing preventable patient errors[180]. The potential for machine learning to assist with the continuous development of incident learning systems[181], coupled with the well-documented under-reporting of healthcare incidents[182], points directly to a need for new automated tools to improve the efficacy of learning healthcare systems.

5.1.3 Study overview

Due to the combined limitations of protocol- and redundant check-based IGRT error mitigation and the limitations of incident reporting systems, this chapter proposes a novel

automated method to retrospectively search for previously unreported incidents and near-miss events at our institution. The clinical impact of such a procedure is twofold. Using an algorithm to analyze large image databases as a "first pass" and pre-select suspicious images for a narrowly focused, manual review allows the time of the reviewer to be spent analyzing patient cases that are more likely to represent errors. The use of such tools could allow for a more comprehensive and thorough search for clinical errors and near-miss events. By analyzing previously unknown incidents and near-miss events, more focus can be turned on the situations in which these incidents occur, thus reducing the overall clinical error rate[183]. In addition, the use of our automated method could assist in obtaining a more accurate quantification of the true error rate in our department. Previous studies on the radiation oncology error rate have relied on errors that were identified and reported, whereas our method eliminates this requirement. We present here a method for identifying treatment errors and calculating the radiation therapy error rate of our department that is independent from the human-based methods traditionally employed for such tasks.

5.2 Materials and methods

5.2.1 Overview of models

Two CNN-based error detectors trained on stereoscopic x-ray/ DRR image sets from the ExacTrac IGRT system were used for the task of detecting previously unreported incidents at UCLA. The first model was trained to detect the rare but serious specific error of off-by-one vertebral body misalignments, described in detail in **Chapter 3**. At the 95% specificity, this model achieved a sensitivity of 96.4% in detecting misalignments by one vertebral body. The

second model, described in **Chapter 4**, was trained to detect general misalignment errors between the x-rays and DRRs using random simulated translational errors of 1 cm. At the 95% specificity, the sensitivity of this model in detecting translational shifts of 1 cm was 94.7%.

5.2.2 Evaluation data collection

We used our final trained models to retrospectively analyze five years of treatment images for previously unreported treatment incidents. 3,122 patients were initially identified from 2018-2022 that were treated on the same treatment machine at our institution, again using ExacTrac image guidance. We first applied a filter based on the treatment plan name to exclude cases where the final patient alignment was based on soft tissue or fiducial markers. **Table 5.1** details the number of plans excluded by year, along with the final number of patient datasets per year. We also report the subset of patients per year out of this final number who were treated to the thoracic spine. Such patients were ultimately analyzed using both the model trained to detect off-by-one vertebral body misalignments along with the model trained to detect more general misalignments.

Table 5.1: Number of patients excluded from each year, along with the final number of patient datasets collected for evaluation for each year. We also report the subset of patients from each year who were treated to the thoracic spine, as both models were applied to the images from these patients.

	2018	2019	2020	2021	2022	**Total**
Initially identified	755	793	650	531	393	**3,122**
Treatment plan name						
Prostate	85	111	89	57	3	**345**
Lung	86	111	80	50	16	**343**
Rib	8	24	21	19	3	**75**
Abdomen	3	1	6	6	2	**18**
Liver	2	14	3	0	0	**19**
Pancreas	1	9	2	3	0	**15**
Adrenal	0	7	1	2	0	**10**
Bladder	1	0	0	1	0	**2**
Esophagus	0	0	0	0	0	**0**
Heart	1	7	5	2	1	**16**
Final number of patient datasets	**568**	**509**	**443**	**391**	**368**	**2,279**
Final number of spine datasets	**41**	**43**	**29**	**48**	**27**	**188**

The final evaluation set consisted of 2,279 patients across the five years, 188 of whom were treated to the thoracic spine region. Day-of x-ray images along with the corresponding clinical DRRs were extracted from all 12,523 treatment fractions. For patients where repeat ExacTrac imaging was performed within the same fraction, only the first set of x-rays and the corresponding DRRs were used. 4-channel arrays were then created for each treatment fraction using the two clinical x-rays and the two corresponding clinical DRRs. The final dataset consisted of 12,523 arrays from all anatomical treatment sites and 521 arrays from the subset of patients treated to the thoracic spine region, with the breakdown by year shown in **Table 5.2**.

Table 5.2: Number of 4-channel arrays created per year of evaluation data using clinical x-ray/ DRR image sets for all anatomical treatment sites and for the subset of patients treated to the thoracic spine.

Evaluation Year	**2018**	**2019**	**2020**	**2021**	**2022**
Clinical 4-Channel Arrays	2,938	2,873	2,263	2,123	2,326
Spine Clinical 4-Channel Arrays	102	114	104	117	84

5.2.3 Application of models to evaluation dataset

Both trained models were applied to the previously unseen evaluation data from 2018-2022 in order to flag potential alignment errors or near-miss events that were previously unreported. For the model trained to detect general 1 cm misalignments, this image data consisted of 12,523 clinical 4-channel arrays from 2,279 patients, representing the full range of anatomical treatment sites. For the model trained to detect the specific error of off-by-one vertebral body errors, this image data consisted of 521 clinical 4-channel arrays from 188 patients treated to the thoracic spine. After each model was applied to the evaluation data, an image set was considered flagged for a potential misalignment if it was scored higher than the threshold corresponding to 95% specificity for that respective model.

5.2.4 Manual review of flagged images

All flagged images from these years were cross-referenced with our institution's record-and-verify systems, incident learning systems, and external image-guidance systems. A manual review was performed on each flagged image set to determine whether they represented treatment errors, near miss events, true imaging errors that did not ultimately translate into treatment errors, or algorithmic false positives. For those x-ray/ DRR image pairs that were verified to indeed be misaligned, further investigation was done to ascertain whether the misalignment on ExacTrac eventually translated into a treatment error, or whether downstream processes remedied the alignment error prior to patient treatment.

5.3 Results

5.3.1 Classification of flagged images

When the CNN trained to detect general 1 cm misalignments was used to classify clinical 4-channel arrays from 2018-2022, a total of 337 fractions (out of 12,523 total fractions) were flagged as misaligned. When the second CNN, trained to detect off-by-one vertebral body misalignments, was applied to the subset of patients treated to the thoracic spine from the same years, 48 fractions (out of 521 total fractions) were flagged as misaligned.

Cross-referencing of the 337 fractions flagged by the model trained to detect 1 cm translational shifts identified eight treatment errors, seven suboptimal alignments, and two near-miss events that had not been previously reported to UCLA's incident learning system. Ninety-two additional fractions were correctly flagged as misaligned, but manual verification confirmed that further imaging (ExacTrac or alternative imaging modality) was performed and the patients were correctly aligned prior to treatment. For 45 of the fractions, it was impossible to tell if the patient was correctly aligned based on ExacTrac imaging alone. We verified that repeat imaging using a different modality was performed and the patient was properly aligned prior to treatment for each of these cases. In a single instance, the patient images were correctly flagged as misaligned but manual review of the treatment schedule indicated that the patient was ultimately not treated on this day. Imaging acquisition errors and algorithmic false positives accounted for the remaining 182 of the flagged fractions. The full results, broken down by year, are shown in **Table 5.3**.

Table 5.3: Full classification results for the fractions flagged by the model originally trained to detect 1 cm translational shifts.

	2018	2019	2020	2021	2022	**Total**
Total fractions	2,938	2,873	2,263	2,123	2,326	**12,523**
Flagged fractions	71	70	51	75	70	**337**
Treatment error	0	0	0	8	0	**8**
Near-miss	2	0	0	0	0	**2**
Suboptimal alignment	0	0	3	3	1	**7**
Misaligned, corrected with repeat ExacTrac	18	18	6	18	11	**71**
Misaligned, corrected with different imaging modality	4	4	4	3	6	**21**
Misaligned, patient not ultimately treated	0	1	0	0	0	**1**
Can't tell if aligned based on ExacTrac imaging alone	10	5	1	10	19	**45**
Imaging acquisition errors	9	2	4	3	4	**22**
Algorithmic false positive	28	40	33	30	29	**160**

A full categorization of the fractions flagged by the model trained to detect off-by-one vertebral body misalignments is shown in **Table 5.4**. Manual review of the 48 flagged fractions identified no instances of previously unreported treatment errors or near-miss events. Twenty-one fractions were correctly flagged by the algorithm as misaligned, but manual review confirmed that subsequent imaging either in the form of repeat ExacTrac or cone beam CT (CBCT) was performed and that the patient was accurately aligned prior to treatment. For five fractions, determining if the patient was aligned based on the ExacTrac images alone was not possible. For these cases, we confirmed that repeat imaging in the form of a CBCT was performed in order to correctly align the patient. Ten of the flagged fractions were found to be image acquisition errors and 12 were algorithmic false positives.

Table 5.4: Full classification results for fractions flagged by the model originally trained to detect off-by-one vertebral body misalignments.

	2018	2019	2020	2021	2022	**Total**
Total fractions	102	114	104	117	84	**521**
Flagged fractions	7	12	9	10	10	**48**
Misaligned, corrected with repeat ExacTrac	2	2	0	1	1	**6**
Misaligned, corrected with CBCT	0	5	1	6	3	**15**
Can't tell if aligned based on ExacTrac imaging alone	0	3	1	0	1	**5**
Imaging acquisition errors	5	1	3	1	0	**10**
Algorithmic false positive	0	1	4	2	5	**12**

5.3.2 Previously unreported treatment errors

Eight treatment fractions were ultimately determined to represent previously unreported treatment errors. Seven of the fractions were from a single patient treated to 54 Gy in 30 fractions to the brainstem. Our manual review of the ExacTrac images for the seven flagged fractions revealed that the therapists were likely too aggressive with "masking" of the DRRs during the image registration process. Masking involves blocking out parts of the DRR, forcing the ExacTrac image registration algorithm to ignore that anatomy as it attempts to register the DRRs to the x-rays. For these fractions, the area around the orbits in particular appear to be well-aligned. Closer inspection of the alignment as a whole however shows that the base of the skull was misaligned by approximately 1 cm in all of the seven flagged fractions. The ExacTrac images from one representative treatment fraction are shown in **Figure 5.1**.

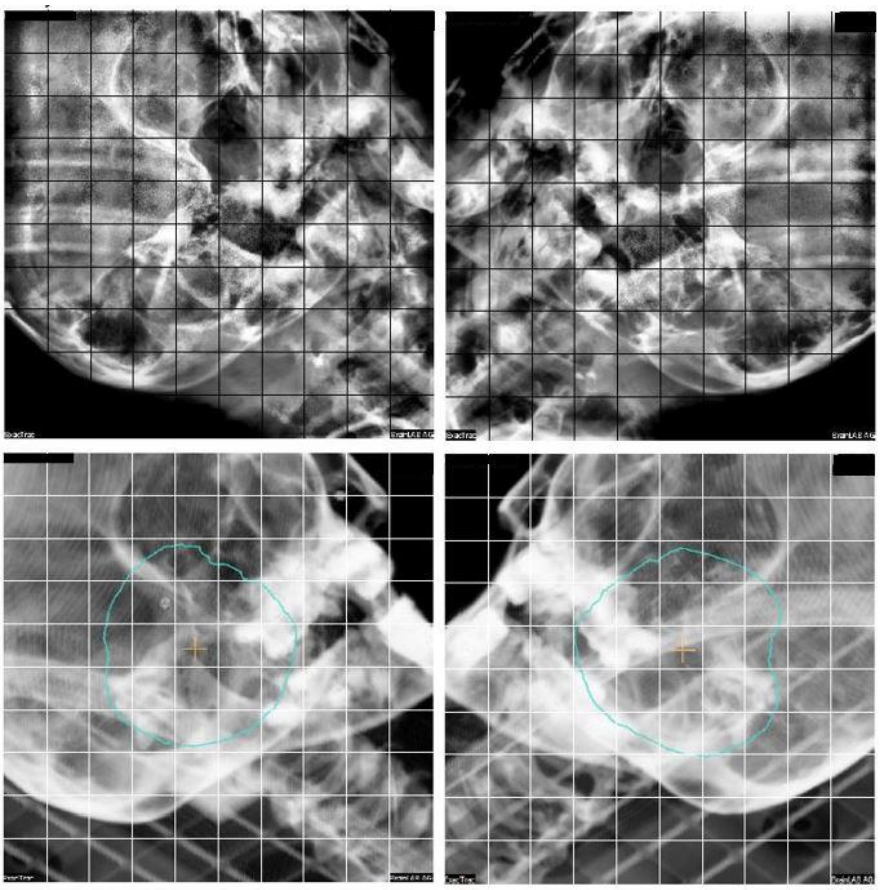

Figure 5.1: ExacTrac images for one out of the seven flagged fractions from a single patient ultimately determined to be previously unreported treatment errors. Each image is 10 cm x 10 cm, and the gridlines represent a distance of 1 cm.

5.3.3 Previously unreported near-miss events

Two flagged fractions were determined to be previously unreported near-miss events, both involving mix-ups of patient names. For one of these flagged fractions (shown in **Figure 5.2**), we immediately noticed that an attempt was made to align one patient's brain treatment plan with a second patient's pelvic x-rays. After performing a thorough manual review of the clinic schedule for this day, we realized that the treatment machine went down early in the day, interrupting the brain patient's treatment. Both this brain patient and a pelvic patient scheduled for a later treatment time shared an uncommon first name. When the treatment machine came

back online later that day, it appears that the therapists mistakenly brought the pelvic patient back instead of the brain patient whose treatment needed to be resumed. It is likely that the therapists used only the patient's first name in the waiting area, unaware that they had two patients at that time with the same name. This mistake was only caught after the pelvic patient was positioned on the treatment couch and imaged, and it appears that the therapists then immediately realized that they had the incorrect patient in the treatment vault. While this particular incident was caught immediately due to the patients being treated in vastly different anatomical regions, such a near-miss could be disastrous if the patients' treatments involved similar anatomies.

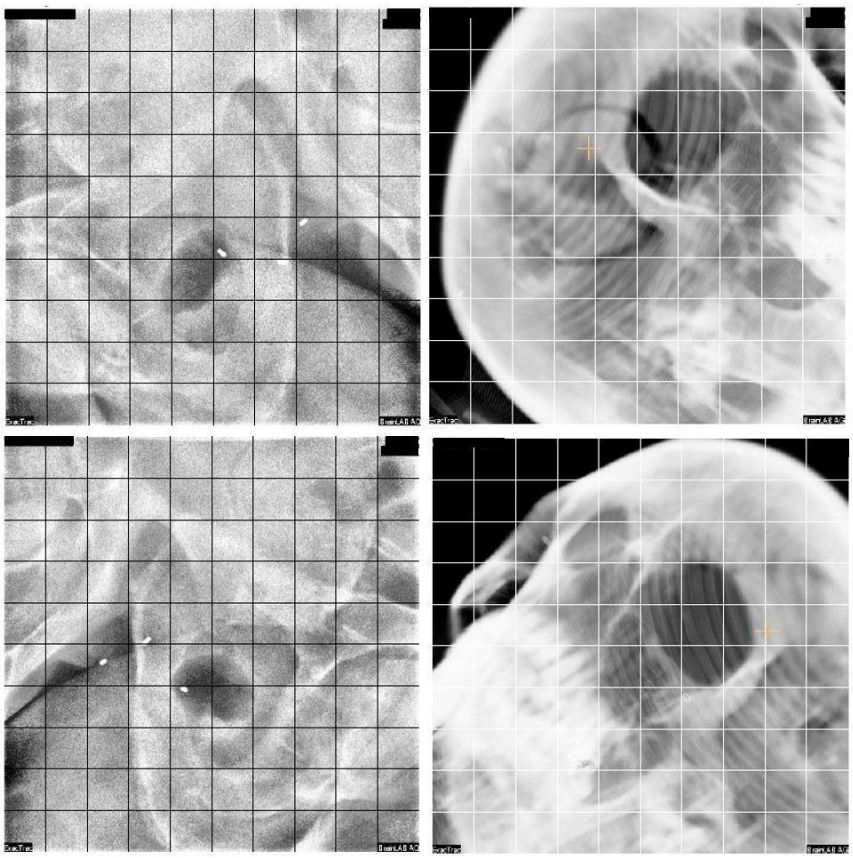

Figure 5.2: X-rays (left) and DRRs (right) for one of the previously unreported near-miss events. Two patients with the same uncommon first name, one being treated to the brain and the other to the pelvis, were mixed up by the therapists.

5.3.4 Suboptimal patient alignments

Seven of the flagged fractions were reviewed and determined to represent suboptimal patient alignment, where the patient was misaligned by a few millimeters at most. Such instances represent learning opportunities and openings for further improving the quality of patient care delivered by all members of the radiation oncology department. For all of these cases, we determined that the suboptimal alignment would not have led to any treatment error. However, the patient could have been positioned more accurately, potentially resulting in a more optimized delivery of the intended dose.

5.4 Discussion

Application of the error detection method we present here to five years of unseen clinical data at our institution resulted in the discovery of eight treatment errors and two near-miss events that were previously unreported. These numbers correspond to an error rate per fraction of 0.06% for our institution, which is in line with error rates reported in the literature of well under 1% of total treatment fractions. We calculated a near-miss rate per fraction of 0.02% for our institution, but to our knowledge no reports on the rates of near-miss events exist in the literature with which to compare this figure. The two error detection models described in this work have shown great promise in identifying previously unreported incidents, and the relatively low false positive rate allows for a comprehensive manual review to be performed on all flagged patients.

The discovery of the previously unreported treatment errors highlights opportunities for training on best practices to improve both the safety and the quality of patient treatments. Seven of the flagged fractions determined to be treatment errors likely occurred due to improper use of the masking function during the process of image registration at the ExacTrac console. This error

in particular highlights the importance of continued education on the best practices for each imaging modality. The ExacTrac system will generally do an excellent job at aligning patients with treatment targets located in the head, due to the plethora of bony anatomical landmarks. For this particular case, such landmarks were purposefully obscured and as a result the patient was ultimately treated in a misaligned position. Repeated training and education on the best practices to use with ExacTrac, or with any clinical system, is critical in a department dedicated to a culture of safety.

Likewise, our discovery of the previously unreported near-miss events involving mix-ups of patient names represent another opportunity to revisit policies and procedures in order to ensure patient safety. For the vast majority of clinical days, it is highly unlikely that two patients in the waiting room would share a name, much less an uncommon first name. Thus, calling the patient back by first name and verifying at the treatment console is considered sufficient. The near-miss described in this chapter highlights the importance of the time out procedure at the therapist console, as this procedure was either not performed or not performed adequately for this particular treatment fraction. The mix-up was not discovered until the patient was on the treatment table and imaging was performed, and then only because the anatomical region was completely different from what was expected. Had the two patients been undergoing treatment to the same general region, this mix-up may not have even been noticed.

29.4% (113 of the 385 total flagged fractions between the two models) were determined to be real patient misalignments. However, further investigation of these fractions found that the initial misalignment was corrected prior to treatment, either by subsequent ExacTrac imaging or by another imaging modality such as cone beam CT. These initial misalignments are common in

the course of a normal radiotherapy treatment, as the therapists acquire repeat imaging to determine the shifts needed to correctly align the patient, and to verify the patient's position immediately prior to treatment. While these misalignments were all caught and corrected during the regular workflow, our algorithm still adds value in flagging these images for a manual review to ensure that the appropriate corrections were indeed applied.

We found that the majority of the false positives (113/172) tended to be situations where the imaging quality was less than ideal. For example, the use of suboptimal mAs and subsequent overexposure in the resulting x-rays was a recurring theme in the list of false positives. Another subset of the patients ultimately classified as false positives had the presence of significant hardware in the imaging field of view. It is possible that deviations from expected imaging quality such as these ones may have led our algorithm to incorrectly classify such images as misaligned. However, we do not necessarily view this classification as problematic. Suboptimal imaging or deviations from routine patient imaging represent situations where mistakes could more easily be missed, and a careful review of such images at the time of treatment is necessary to ensure accurate radiation delivery.

The rapid acceleration of image guided radiotherapy has resulted in huge databases of visual images that can be used to continually improve the quality and safety of the radiation oncology clinic. To date however, little has been done to incorporate automation and machine learning into the task of image review—a task which requires a significant amount of effort from radiation oncologists, physicists and technologists[156]. We believe our work here adds significantly to the effort both of automating image review and, subsequently, of more fully utilizing patient imaging databases to better identify treatment incidents.

5.5 Conclusion

In this work we have developed a convolutional neural network-based approach for the automatic detection of potential IGRT misalignments in planar x-ray setup images that warrant further investigation via manual review. We believe such an algorithm could be a valuable asset in analyzing the large databases of patient setup images generated every day in the modern radiotherapy clinic for previously unknown errors and near-miss events. Our results demonstrated the potential feasibility of this application, with our misalignment detection algorithms identifying eight previously unknown incidents over five years of archived clinical image data.

CHAPTER 6: DOSIMETRISTS' REPORTED BARRIERS AND FACILITATORS TO CLINICAL IMPLEMENTATION OF TREATMENT PLANNING AUTOMATION

6.1 Introduction

6.1.1 Automated tools in radiation oncology

In recent years there has been a substantial increase in research and development involving automation of the radiation therapy treatment planning workflow. Treatment planning automation can be used to reduce the occurrence of sub-optimal treatment plan quality[184], facilitate adaptive radiotherapy[185], reduce treatment latency, and allow human cognitive resources to be directed to their most high value uses. Almost every step of the radiation treatment workflow, from normal tissue contouring[155,186], to IMRT treatment planning[153,187], to online adaptive replanning[188], to the physics plan review process[154,189,190], has been subject to automation research. Automated normal tissue contouring tools have demonstrated accuracy comparable to manual contours for many target organs[191], and demonstrated significant time savings in clinical workflow studies[192,193]. A wealth of research has demonstrated that automated radiotherapy planning techniques are capable of producing clinical-quality plans[194,195]. Prospective studies have shown that in carefully controlled situations, automated treatment planning generates plans of comparable or better clinical quality, at significant time savings[196–198]. A variety of commercially available automated tools exist, including but not limited to atlas-based[199] and deep-learning based[193] auto-contouring, knowledge-based planning[200], rule-based automated planning[201], and automated field-in-field planning[202].

6.1.2 Sparsity of implementation research

However, little is known about the scale of clinical implementation of automated treatment planning techniques in the United States. We hypothesized that clinical adoption of treatment planning automation may be less than fully realized and furthermore that barriers to implementation may exist. These hypotheses were based on the reported observation of barriers to automation in similar areas of healthcare such as diagnostic radiology[203,204] and pharmacy[205]. Additionally, the authors have previously collected anecdotal data that medical dosimetrists, who in many clinics perform the majority of treatment planning, often express hesitation at using treatment planning automation. In order to ensure that advances in research translate into advances in clinical care, a focused effort is required in order to understand the barriers to implementation[101]. While an individual clinic may be committed to the adoption of evidence-based best practices in principle, the actual implementation of these practices requires a thorough understanding of all the breakpoints where such implementation can fail[102]. The diversity of health care settings in the United States represents a major challenge to the widespread dissemination of evidence-based best practices[206].

6.1.3 Study overview

In this chapter, we examine the barriers and facilitators to adoption of commercially-available automatic treatment planning tools into the clinical workflow using a survey of medical dosimetrists. We focus on how implementation of treatment planning automation is viewed by medical dosimetrists within the radiation oncology clinic. Here we define treatment planning automation as the automation of parts of the treatment planning workflow, such as auto-contouring and automated dose optimization. To our knowledge, complete end-to-end

automation of the treatment planning workflow has very limited if any clinical implementation, but our survey left open the possibility for respondents to address complete automation as well. To date, no published research has examined whether or why medical dosimetrists may view these tools favorably or unfavorably. Our primary goal is to identify the barriers to implementation from the perspective of the medical dosimetrist. Our secondary goal is to offer potential facilitators to increase the adoption of evidence-based best practices with respect to automated treatment planning in the context of the radiation oncology clinic.

6.2 Materials and methods

6.2.1 Survey best practices

Several of the best practices compiled by Krosnick in his 1999 review paper[207] are worth briefly mentioning as they are relevant to our own survey methodology. First, in order to draw general conclusions about a population based on survey responses from a sample, it is imperative to ensure a representative sample has been obtained. Our survey measured familiarity with and attitudes towards automation, but we were careful not to restrict potential respondents based on their prior use of these tools. Instead, the only criteria we imposed was a sampling of dosimetrists currently employed in California, regardless of their prior experience with automation in their workplace. Second, we employed close-ended questions throughout our survey. Close-ended questions can be used effectively when the choices given constitute a comprehensive list of all possible options[207]. In all of our survey questions, great care was taken to ensure that respondents were presented with a comprehensive list of all possible options. Third, all points on rating scale questions were fully labeled and intended to divide the response

continuum into approximately equal intervals in order to maximize validity[208]. It is well-documented that respondents have a tendency to place themselves towards the middle when answering rating scale questions[207]; however this phenomenon was not observed at large in our data. Finally, research has shown a tendency for respondents to agree with statements more frequently than they disagree[207]. While we did employ the frequent use of "agree/ disagree" question formats, care was taken to ensure that we maintained a balance between positive and negative descriptions of automation.

6.2.2 Survey design

Survey questions broadly probed the following areas: frequency of use of treatment planning automation (auto-contouring and automated dose optimization), positive and negative perceptions about automation performance, potential implementation changes that would affect accessibility and usability, and demographics and institutional descriptive statistics. Positive and negative questions were balanced to reduce bias[209]. Level of agreement questions were heavily utilized because they facilitate balanced positive and negative statements.

The final survey questions can be broken down into five general subsections: Prior Use, Auto-Contouring (AC), Automated Treatment Planning (ATP), General Level of Agreement, and Demographics. The Prior Use section consisted of a single question that asked respondents to mark any specific automation tool they had used at any point during their career. **Table 6.1** lists the categories of treatment planning automation delineated in this question, along with the commercial product names used as examples of each category. Responses to this question determined the branching logic that would follow for the remainder of the survey. In the Auto-Contouring section, respondents were asked questions to gauge their level of experience with AC

tools, what types of commercially-available tools they have used, what anatomic sites they have used AC for, and reasons why they view the tools they have used favorably or unfavorably. The Automated Treatment Planning section asked respondents to answer questions regarding their level of experience with ATP and how often they use it, what types of ATP algorithms they have used, what anatomic sites they have used ATP for, and reasons why they like or dislike the ATP tools that they have experience using. Branching logic was used to only show survey questions relevant to the specific AC and/or ATP tools that the participant had prior experience using. The General Level of Agreement section consisted of a list of statements designed to elicit responses on how participants view automation in ways that may not be specific to individual automation tools. The Demographics section consisted of questions about the participant's age, gender, length of time employed in the field, education and relevant certifications, and current clinical environment. The survey was deployed using the secure survey platform Qualtrics. The complete list of survey questions can be found in **Appendix 1**.

Table 6.1: Example products for each category of auto-contouring (AC) and automated treatment planning (ATP) surveyed.

AC/ ATP Category	Example Products
Deep learning-based auto-contouring	Mirada DLCExpert, MIM ContourProtege AI
Atlas-based and/or model-based auto-contouring	MIM Atlas Segment, Varian Velocity, RayStation MABS/MBS, Pinnacle SPICE, Elekta ABAS
Knowledge-based plan quality assessment	Sun Nuclear PlanIQ
Automated planning using knowledge-based planning algorithms	Varian RapidPlan
Automated field-in-field planning	Radformation EZFluence
Automated planning using rule-based or template-based algorithms	Philips Pinnacle Auto-Planning, Raysearch Raystation Auto-Planning

6.2.3 Recruitment of subjects

Subjects were recruited through two different channels in accordance with the approved IRB protocol. We used LinkedIn to solicit responses from medical dosimetrists currently

employed in a clinical capacity within the state of California. All medical dosimetrists with LinkedIn accounts showing employment as clinical dosimetrists in California were contacted directly via LinkedIn. Additionally, professional contacts employed outside of our department were recruited. Finally, all chief physicists of non-academic medical centers in California listed in the AAPM member directory were contacted with requests for references to medical dosimetrists. This was done in an attempt to balance the responses in a way that more accurately reflected the distribution of academically and non-academically employed dosimetrists in California. No participant was recruited with whom there was a supervisorial relationship with any of the investigators. During the recruitment process, we did not require familiarity with or regular use of automated treatment planning as a prerequisite for survey participation. In the event of a non-response from a potential participant, we reached out a second time but did not aggressively pursue a higher response rate beyond this second contact. It has been shown that a low response rate does not inherently indicate the presence of non-response error in the final data[210,211].

6.2.4 Statistical analysis

Fisher's exact test was used to test for statistical significance of correlations between measures of use of automation with reported demographic variables. Clustering analysis was performed in the R programming environment[212] in order to identify latent groups in the responses.

6.3 Results

6.3.1 Respondent demographics

In total 171 medical dosimetrists were contacted either on LinkedIn or by email, of which 57 responded and were sent a survey link. Of the dosimetrists who were sent a survey link, 34 completed the survey. Survey results broadly sampled level of education, gender, and place of employment (academic vs. non-academic hospital vs. community clinic), but appeared to be weighted towards relatively young dosimetrists, with 61.8% of respondents reporting ages less than 39. Complete demographics are contained in **Table 6.2**.

Table 6.2: Survey respondent demographics.

	Responses (n=34)
1. Age	
20-29	4
30-39	17
40-49	5
50-59	7
60+	1
2. Years of Experience	
< 5	11
5-9	11
10-19	9
20+	3
3. Gender	
Male	18
Female	16
4. Level of Education	
Associate's degree	4
Bachelor's degree	16
Master's degree	13
Doctorate	1
5. Place of Employment	
Academic medical center	16
Non-academic hospital	12
Community practice	6
6. Number of radiotherapy machines	
1	3
2-4	14
5-8	10
9+	7

6.3.2 Familiarity with AC and ATP

Clinical use of AC remains limited, with 70.6% of respondents (24/34) reporting that they used auto-contouring less than weekly. Use of ATP was more frequent, with 41.2% reporting that they used it at least weekly. Respondents reported approximately equal familiarity with AC and ATP, with average familiarity scores of 2.82 and 2.59 out of 5 for AC and ATP respectively. Despite recent research demonstrating that deep learning-based AC is more accurate than atlas-based AC, most respondents reported that they used atlas-based AC. Use of ATP algorithms was more heterogeneous, and the most commonly used algorithm was automated field-in-field planning (see **Figure 6.1**).

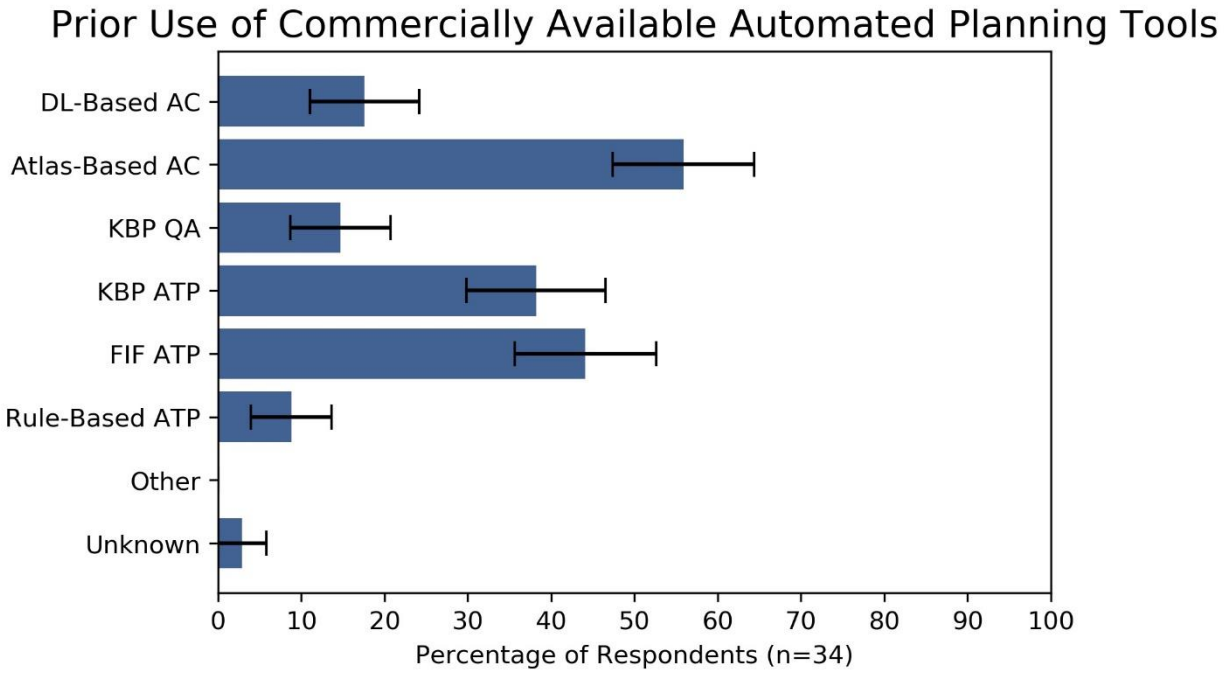

Figure 6.1: Reported frequency of use of commercially-available automated tools. Error bars represent one standard error.

Respondents were more likely (18/34) to have heard the most about auto-contouring from scientific talks and vendor booths at professional meetings, and more likely (22/34) to have heard about it least frequently from peers at their own clinic or elsewhere.

6.3.3 Barriers and facilitators to use of automated tools

A number of potential barriers and facilitators to use of automation were reported frequently (**Figure 6.2**). The most commonly identified barrier to clinical use of AC was contour inaccuracy, with 30 out of 34 survey respondents reporting that increased accuracy would make them more likely to use AC tools. The most commonly identified potential facilitator to use of ATP was if the ATP algorithms would produce plans that were easier to modify in order to get an optimal plan. 21 out of 34 participants responded this way. However, respondents did see value, or potential value, in both AC and ATP. A significant majority of respondents (23/34) reported liking ATP because it allowed them to work through a higher caseload (**Table 6.3**).

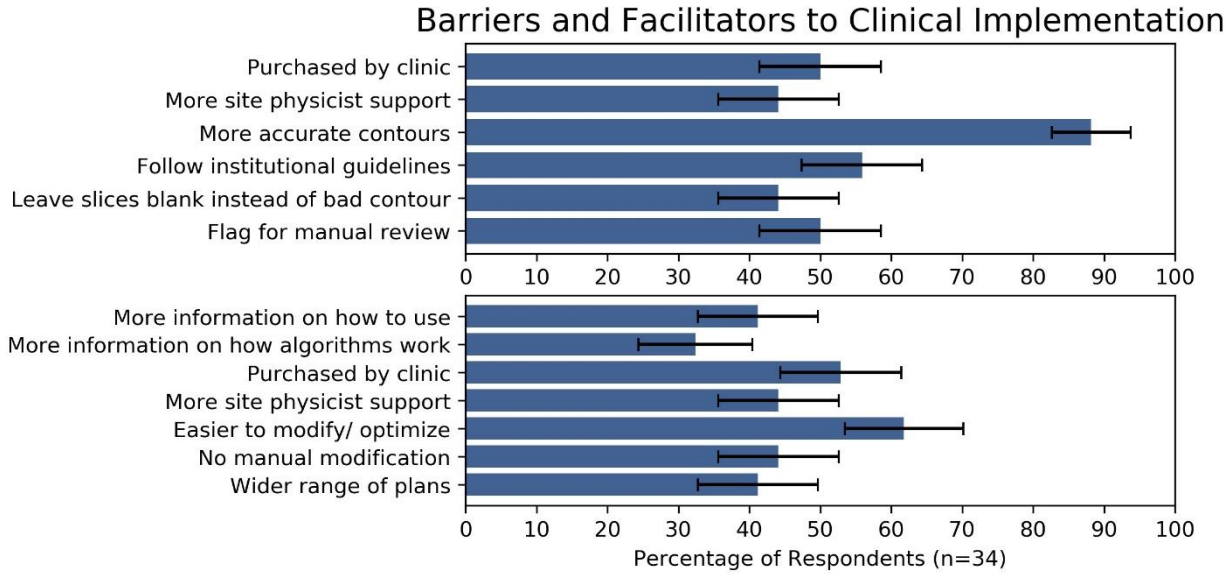

Figure 6.2: Reported barriers and facilitators to use of auto-contouring (top) and automated treatment planning (bottom) tools. Error bars represent one standard error.

Table 6.3: Reported reasons for liking/ disliking auto-contouring (AC) and automated treatment planning (ATP).

	Responses (%)
1. Dislike of AC	
Would rather contour from scratch	70.6
Concerned about algorithm making an error	64.7
2. Dislike of ATP	
Do not believe plans are of the same quality	41.2
Takes more time than generating plan from scratch	44.1
Enjoy optimization, do not want to lose that part of job	58.8
3. Like of ATP	
Work through higher patient caseload	67.6
Higher degree of confidence in the plans	29.4

An area of concern for dosimetrists was that the use of automation could increase the likelihood of errors. Amongst users of deep learning-based and atlas-based AC, 50% and 52.6% respectively were concerned that it could lead to treatment errors. Among users of knowledge-based planning (KBP) quality assessment and KBP automated planning, 40% and 38.5% respectively were concerned that it would lead to treatment errors. Only 21.4% of users of automated field-in-field planning were concerned about errors.

Dosimetrist perceptions of the impact of AC and ATP on job satisfaction and job security appeared to be important. A majority (20/34) reported that they "disliked ATP because they enjoy plan optimization and don't want to give up that part of their job", and 63.6% agreed or somewhat agreed with the statement that they "value their planning skills highly and would be disappointed to see them devalued". Likewise, a majority (21/34) agreed or somewhat agreed with the statement "I worry that automated treatment planning will hurt the job market for dosimetrists." A slight majority (19/34) agreed or somewhat agreed that routinely using ATP could lead to atrophy of planning skills.

6.3.4 Fisher's exact test

No statistically significant correlations were found between auto-contouring level of experience (rated on a scale of 1 to 5) or frequency of use versus level of education, place of employment, and number of machines in the clinic. Similarly, no significant correlation was found between automated treatment planning level of experience (rated on the same 1 to 5 scale) or frequency of use versus these same demographic variables. A statistically significant correlation was found between auto-contouring level of experience and view of planning goals at their clinic as standardized ($p = 0.046$), and between automated treatment planning level of experience and the same metric of planning goal standardization ($p = 0.014$).

6.3.5 Latent class analysis

Clustering analysis using latent class analysis identified a partition that related to the dosimetrist's clinical environment as a barrier to use for both automated treatment planning and auto-contouring. A cluster of dosimetrists was identified (comprising 25.5% of respondents overall) that were mainly employed at hospital-based medical centers (HBMC) or community medical centers (CMC). 100% of participants in this cluster reported being more likely to use ATP if it was both purchased by their clinic and if they received more support from their supervisor or site physicist **(Figure 6.3)**. This suggests that for dosimetrists employed at non-academic institutions, lack of access to the technology itself may be an important barrier to use. In the complementary group comprising 74.5% of respondents, the majority reported employment at academic medical centers (AMC). Within this cluster, only 36.8% reported that they would be more likely to use ATP if it was purchased by their clinic, and only 25% reported that they would be more likely to use ATP if they received more support from their site physicist

or supervisor. For dosimetrists within this cluster, it appears that access to both ATP tools and support for those tools is a much less significant barrier than for their colleagues at non-academic institutions.

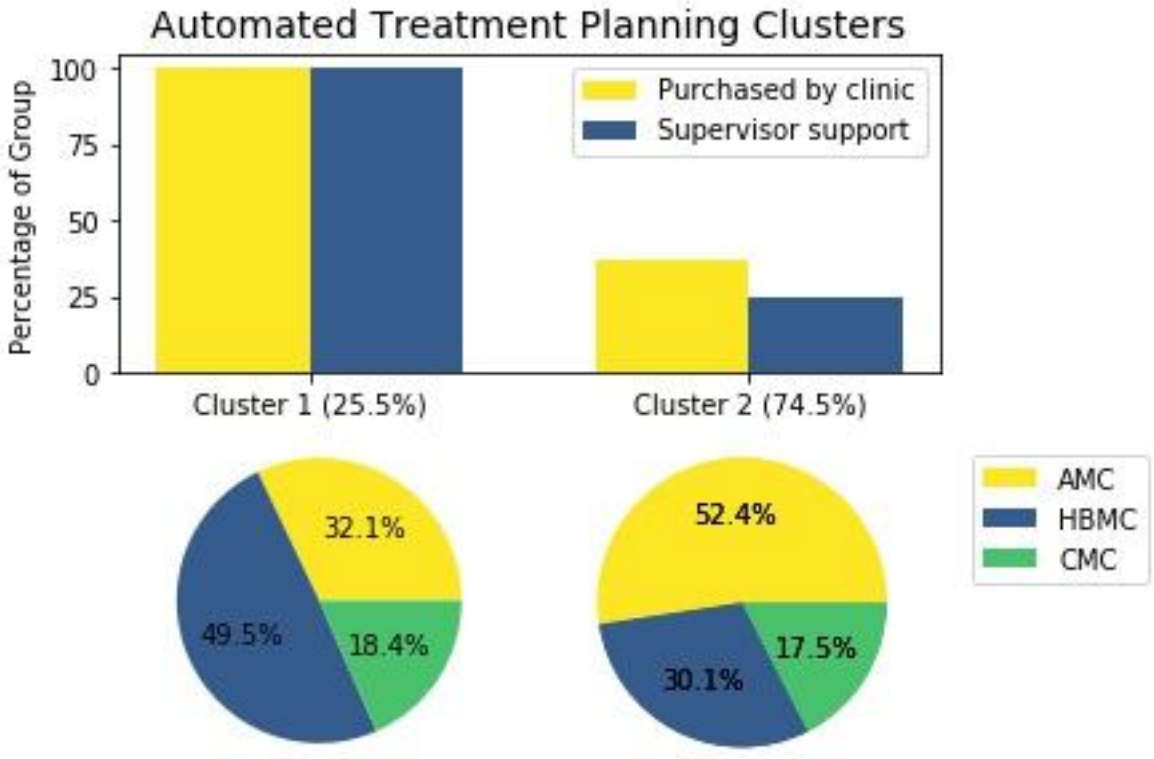

Figure 6.3: Percentage of dosimetrists reporting certain factors as potential facilitators to use of automated treatment planning by cluster (top); employment breakdown of Cluster 1 (bottom left); employment breakdown of Cluster 2 (bottom right).

An analogous clustering was identified in relation to use of AC with an almost identical group membership. Latent class analysis identified one group (comprising 21% of total respondents) where 20.1% of the cluster reported employment at an academic medical center, 42.7% reported employment at a hospital-based medical center, and 37.2% reported employment at a community medical center. For this cluster, 100% of group members reported being more likely to use AC if it was purchased by their clinic while 99.4% reported being more likely to use AC if they received more support from their supervisor or site physicist **(Figure 6.4)**. In the

remaining 79% of respondents, 54.2% reported employment at an academic medical center, 33.3% at a hospital-based medical center, and 12.5% at a community medical center. Only 36.7% of this group reported being more likely to use AC if it was purchased by their clinic, and only 29.4% reported being more likely to use AC if they received more support from their supervisor or site physicist. These results again point to lack of access to tools and support for use of these tools as an important barrier for dosimetrists employed in a non-academic medical center setting.

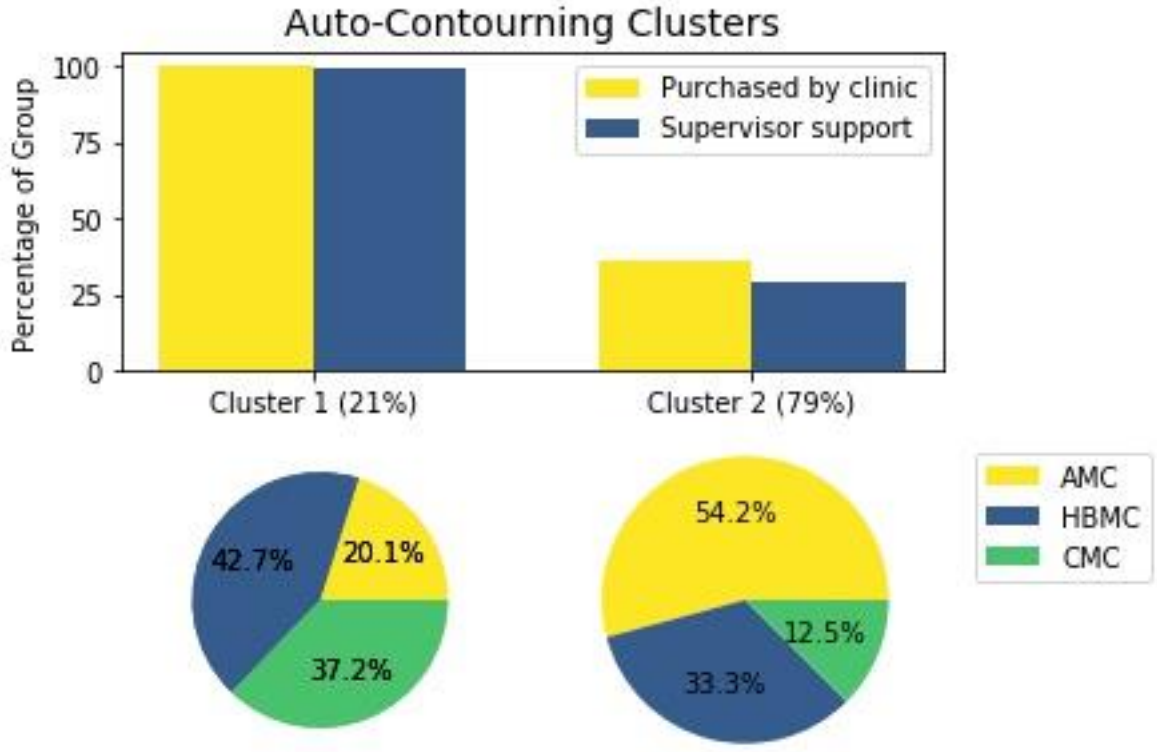

Figure 6.4: Percentage of dosimetrists reporting certain factors as potential facilitators to use of auto-contouring by cluster (top); employment breakdown of Cluster 1 (bottom left); employment breakdown of Cluster 2 (bottom right).

6.4 Discussion

This survey has identified three broad barriers to use of automation in treatment planning. The first barrier relates to the limited accuracy and usability, or perception thereof, of the algorithms. A remarkable 30 of 34 respondents thought that auto-contouring inaccuracy limited its use. It is noteworthy that a minority of respondents (6/34) reported use of deep-learning based auto-contouring, which has been shown in the literature to be significantly more accurate than other methods[191,213,214]. Therefore, a broader availability of deep-learning based tools could facilitate a broader use of auto-contouring in the clinic. A strong majority of respondents thought that it was difficult to modify the output of an automated planning algorithm and this limited the algorithms' usefulness. This points to human factors engineering[111] as an important component of treatment planning automation—and of the subsequent clinical implementation—that needs more attention. It should also be noted that the limits of usable accuracy may be lower than what is typically perceived by treatment planners[215]. Survey results showed that dosimetrists heard about automation most frequently from scientific talks and vendors. Vendors and academic proponents of automated tools may be perceived as biased in their descriptions of algorithm performance. Peer-to-peer teaching and continuing medical education focused on automation could address this potential perception gap. Finally, statistically significant correlations were observed between level of experience with automated treatment planning and dosimetrists' perceptions of the degree to which planning goals were standardized at their clinic. This highlights the importance of standardization of clinical goals for the uptake of automation and supports the potential role of automated planning in the context of clinical trials[216].

The second barrier relates to the perception of the dosimetrist that using automation increases the probability of an error reaching the patient. This directly relates to the well-documented automation bias[217,218]. In principle, all the results of treatment planning automation should be reviewed by one or typically more than one human observer. However, little is known about the effectiveness of this review and there is reason to believe it is less than 100% effective[219]. This points to the need for more research into the effects of automation bias in treatment planning, and if significant, approaches towards minimizing it.

Third, dosimetrists are concerned that treatment planning automation will make their jobs both less satisfying and less secure. A large majority of dosimetrists reported that they enjoyed plan optimization, wouldn't want to lose that part of their job or see it devalued, and expressed explicit job security fears. Contrastingly, in one of the most one-sided results, 25/34 respondents agreed or somewhat agreed that they would want to use ATP if it worked well. This points to the need for more attention given to developing a picture of what the dosimetrist role looks like as a clinic transitions more fully towards automated technologies[220,221]. Ultimately, dosimetrists viewed increasing levels of contouring automation as inevitable, with 82.4% agreeing or strongly agreeing that by the end of their career, most or all normal tissue contours will be auto-generated. A smaller fraction (17/34) believed that by the end of their career, ATP will replace most manual plan optimization.

The results of this survey can be interpreted in light of a technology adoption model such as Venkatesh et al.'s Unified Theory of Acceptance and Use of Technology (UTAUT)[105]. Venkatesh et al. outlined four broad factors that determine how well new technology is taken up within the workplace: performance expectancy, effort expectancy, social influence, and

facilitating conditions. The perception of performance of auto-contouring tools was negative for most respondents. For automated treatment planning, most respondents felt that effort to use was higher than it should be (effort expectancy). Thus, efforts to improve these factors (or the perception of these factors) would be likely to improve adoption of treatment planning automation.

While every effort was made to ensure we obtained a representative sample, we are cognizant of the limitations we faced. Since much of our subject recruitment was conducted via LinkedIn, our sample was weighted to those dosimetrists actively utilizing LinkedIn. We noticed that our sample demographic tended to skew younger than what a truly representative sample would likely show. However, research in this area has shown that statistically correcting for potential demographic biases is not likely to impact the overall conclusions drawn from the data[222]. Furthermore, the over-representation of younger dosimetrists in our sample has the potential advantage of offering insight into the factors that will be most relevant to the continuing clinical implementation of automation in radiation oncology, since the majority of the respondents likely expect to continue their employment in this field for decades to come. Finally, there may be a self-selection process at play because our sample was weighted to responses collected via LinkedIn. These dosimetrists may be more sensitive to or aware of new and emerging technologies in their field and have different perceptions of automated treatment planning than their colleagues not using LinkedIn. Every effort was made to emphasize that personal use of automated treatment planning was not a prerequisite for our survey, and in our own data we observed responses from dosimetrists ranging from no experience to highly experienced.

In regard to sample size, we acknowledge that our sample of 34 respondents is not large. We believe some factors mitigate the low absolute number of respondents in our survey. Our survey was detailed, requiring an estimated 15 minutes to complete, and each response provided a high density of information. Our 34 respondents came from 23 unique institutions (six academic and 17 non-academic) in California, and survey responses from individual dosimetrists likely represent practice patterns of other dosimetrists at those institutions and correspond to a considerable patient population served. We attempted to reach all medical dosimetrists employed in California by contacting them directly on LinkedIn, and systematically contacting all the chief physicists of non-academic medical centers in California who were listed in the AAPM member directory. Of note, no publicly available email directory exists for the American Association of Medical Dosimetrists, so we were unable to contact registered dosimetrists in a systematic way via this organization and any internal email address list it may maintain. Although limited in size, our sample required significant effort to collect and we believe will not be easily surpassed.

Our data may also be limited due to effects from the social desirability bias[207]. Automation is a new and exciting field, and scores of research has emphasized the benefits of such technological advances[223,224]. Respondents may have felt pressure to conform to this social bias and offer responses in line with the prevailing opinion of automation as "good." In our own data, almost all respondents expressed some positive and some negative views of automation. Our survey questions were designed to present both positive and negative views of automation in order to reduce this particular form of bias in the responses.

Response order bias may arise in surveys when respondents process questions in a satisficing instead of optimizing way[225]. Satisficing respondents are more likely to choose the

first reasonable option they are presented with in a list of possible options. However, this effect was not observed in our data. This may be due to our use of "select all that apply" questions to evaluate the underlying attitudes towards automation of our survey sample. The majority of our respondents took the opportunity to select multiple options on these question types, suggesting that they were evaluating each option thoroughly.

Another potential source of bias in our data is due to our use of "agree/ disagree" questions to measure respondents' views of and attitudes towards automation. These questions can pose challenges through a tendency for respondents to initially agree with the assertion being made in the statement and spend more time looking for reasons to agree with the statement than looking for reasons to disagree[207]. In order to minimize this bias, we balanced level of agreement statements with positive and negative views of automation. We observed approximately equal numbers of agreement and disagreement, indicating a low degree of agreement bias in our results.

6.5 Conclusion

To our knowledge this is the first systematic investigation into the views of automation by medical dosimetrists, who perform the majority of treatment planning at many if not most radiotherapy facilities. We have explicitly identified potential barriers and facilitators to use of automated technologies in the radiation therapy treatment planning workflow. This investigation highlights several concrete approaches that could potentially increase the translation of treatment planning automation into the clinic, as well as areas of needed research.

CHAPTER 7: CLINICAL PHYSICISTS' PERCEPTIONS OF WEEKLY CHART CHECKS AND THE POTENTIAL ROLE FOR AUTOMATED IMAGE REVIEW

7.1 Introduction

7.1.1 Weekly chart checks

Physics weekly chart checks are an essential part of the professional responsibility of medical physicists, and are a key guardrail for identifying errors and improving quality[226]. They are imperfect however—the effectiveness of weekly chart checks is limited, with the sensitivity of detecting errors during physics chart checks reported in the literature ranging from 43% to 63%[62,226]. In order to examine potential approaches for improving the effectiveness of weekly chart checks, it is of the utmost importance to hear directly from medical physicists. From the field of implementation science, we know that hearing directly from the end user is crucial to improving any clinical process.

Automation of specific physics checks has been suggested as one potential avenue for reducing errors[219]. The earliest report of an automatic error detector dates back to 2007, with the publication of a clustering algorithm to detect plan outliers[227]. Since then, research into automating many parts of the physics check practice can be found in the literature[154,228–231]. These publications focus primarily on the verification of technical details or data transfer—quantitative values which are good initial candidates for automation. Such studies make the case that automating repetitive tasks frees up time for the medical physicist to investigate events or complex issues, or to spend greater time on plan quality evaluation—in sum, to devote greater cognitive effort to complicated patient cases rather than the routine checking of treatment

parameters. However, limited research has been done into the applications of automation to the specific task of IGRT image review. Though there is wide practice variation in whether physicists perform image review, for those clinical medical physicists who routinely review images as part of their chart checks, it can be a time-consuming process. Thus, IGRT image review may be a good candidate for automation research.

7.1.2 Recent AAPM guidelines

Our work stands in light of the recent efforts of the American Association of Physicists in Medicine Task Group 275 (AAPM TG-275) who sought to "provide practical, evidence-based recommendations on physics plan and chart review for radiation therapy." TG-275's recommendations were given following a Failure Mode and Effects Analysis (FMEA)[232] based partly on a survey distributed to clinical medical physicists working in radiation oncology. The results of the survey deployed as part of this Task Group highlight a wide range of different chart checking practices currently being used by clinical medical physicists[233]. While this survey covered a broad range of demographic and chart checking topics, it used multiple choice questions to capture data, which limits the potential for a deeper qualitative analysis into why such variations among institutions and professionals exist. The work we present here complements the efforts of TG-275, in that we utilized semi-structured interviews to understand in greater detail the current shortcomings of the weekly chart check process and examine potential ways for improving the process.

Building off the findings in TG-275, the authors of Medical Physics Practice Guideline 11.a (MPPG-11.a) strove to develop a professional guideline of the minimum standard of patient chart checks that should be performed in order to ensure patient safety[234]. Both AAPM TG-275

and MPPG-11.a acknowledge that as new technologies, particularly new automated technologies, enter the market, the workflow of chart checks will change. Deficiencies present in the error-detection potential of current chart checks[62,226] further justify the need for continuous improvement of the chart check process, especially as patient treatments grow ever more complex. Indeed, the authors of TG-275 opined: "Taken together, the results of these studies indicate a need to improve plan/ chart review processes. *Improvements are needed not only in the content of what is checked but also in the implementation of these checks to improve performance through various methods including standardization and automation"* (emphasis added).

7.1.3 Study overview

The primary goal of this work is to understand clinical medical physicists' perspectives on the current weekly chart check process and identify the shortcomings in the practice. We use a novel thematic analysis approach to identify common themes among semi-structured interviews of clinical medical physicists who are currently involved in their institution's chart check process. The secondary goal is to collect feedback from clinical medical physicists to explore avenues for future work on development of automated tools to aid in the time-consuming IGRT image review portion of weekly chart checks—an area which has been understudied in the research to date. Specifically, our group aims to understand what features of automated tools designed to assist with IGRT image review tasks would be most useful to the end user—in this case, clinical medical physicists. A critical component of this secondary analysis involves investigating any potential barriers to implementation that may arise as a clinic moves forward with adopting such technologies into their clinical workflow. Identification of barriers and

facilitators is essential to maximizing the adoption and utility of a new tool, technique or innovation[101,102]. To meet these goals, we conducted a qualitative study, using semi-structured interviews with practicing medical physicists, and analyzed them using thematic analysis.

7.2 Materials and methods

7.2.1 Recruitment of subjects

Our sampling frame included clinical medical physicists who participate in their institution's weekly chart check process in both academic and non-academic (including governmental and community clinic) centers, across a multi-state sample including all regions of the United States. **Table 7.1** shows how many physicists we interviewed belonging to each group. Interviewees were recruited in accordance with the approved IRB protocol. Participants were identified through professional contacts of the authors, and no one was recruited for an interview with whom there was any supervisorial relationship. Nineteen semi-structured interviews with physicists at 16 different clinics were conducted via Zoom with an approximate length of 30 minutes each. All interviews were recorded and saved for later analysis.

Table 7.1: Employment demographics of our interviewees.

Place of employment	Number of interviewees
Academic medical center	10
Non-academic medical center	9

7.2.2 Quantitative survey questions

Our interview script included several quantitative survey questions. Respondents were first asked to describe their current weekly chart check workflow. They were asked how many weekly chart checks they perform per week, how often they perform IGRT image review as part of their regular weekly chart check, and how long they typically spend on this image review.

They were then asked about any tools they currently use to automate any part of their weekly chart check. Physicists were asked to rate on a scale from 1 to 10, where 1 was least important and 10 was most important, the importance of 1) reducing the time spent on image review, and 2) increasing the effectiveness of image review.

7.2.3 Semi-structured interview design

Semi-structured interviews are a well-established technique for data collection in various healthcare fields[235–239]. Interviews follow a general script, while also allowing room for deviation from the script, potential probing questions, and for conversational flow. Our interview script included several topics. Respondents were asked open-ended questions regarding what they view as the shortcomings of the current chart check process, focusing specifically on the IGRT image review component. Interviewees were then asked about what features of an automated tool to assist with the IGRT image review portion of chart checks they would find useful. A beta version of an automated IGRT image review tool developed by our group[240] was shown to interviewees and feedback was collected on desired features of such an automated tool. The software interface shown to interview participants displayed an image alignment score along with previous scores for the same patient and cumulative data for the clinic overall. Finally, respondents were asked to describe the potential barriers and facilitators to use of automation in the weekly chart check workflow that they could anticipate arising in their own clinic. The full interview script, including both the survey questions and the semi-structured topic questions, can be found in **Appendix 2**.

7.2.4 Transcription of interviews

Zoom audio recordings of the interviews were transcribed using NVivo[241] transcription software (Lumivero, Denver, CO). Manual review and correction were performed to ensure transcriptions were accurate.

7.2.5 Thematic analysis

Thematic analysis is, broadly speaking, "a method for identifying, analyzing, and reporting patterns (themes) within data"[242]. It has previously been identified as a research method that has broad applicability to a range of qualitative health research questions[243]. Thematic analysis involves the collection of data, often via semi-structured interviews or focus groups, and the subsequent analysis of common themes across that dataset. This technique is well-established in qualitative research, with applications ranging from analyzing the perceptions of corruption in the construction industry[244] and identifying health themes in the realm of smart home technology[245], to the more narrowly medical field-focused applications of generating themes describing the views of postnatal health care[246] and, most recently, identifying themes related to the experiences of both frontline healthcare providers[247] and patients[248] during the COVID-19 pandemic. Thematic analysis offers an accessible approach to qualitative research in general[249], as it does not require the use of a pre-existing theoretical framework (unlike various other approaches to qualitative research). It is well-suited for our task of analyzing how clinical medical physicists currently conduct weekly chart checks and their feedback regarding the introduction of automation to the process.

Clarke, Braun, and Hayfield, in their book chapter on the topic[249], describe the six steps necessary to conduct a high quality thematic analysis: data familiarization, coding, generating

themes, reviewing themes, defining and naming themes, and writing the report. We conducted data familiarization through a preliminary read-through of our collected data, to gain a general familiarity with the type of data involved. During the coding process, we analyzed the data in finer detail and defined excerpts of text into various codes. **Figure 7.1** depicts a frequency analysis of the ten most commonly used codes in this step. This step remained fluid, and we utilized a mix of semantic and latent coding[250]. Semantic codes focused on the things explicitly stated by participants, while latent codes focused more on the underlying meanings of what was said[249]. For this study we took an inductive approach to coding, where the codes were generated based directly on the data itself, rather than a deductive approach that brought in preconceived notions about what the data might show. This approach is favored for cases in which there are no previous studies on the topic that may inform the researcher about what to expect in the data[251]. During theme generation, we combined multiple codes into larger overarching themes that told a story about the data. In this step it was vital to not confuse themes with topics—a common issue in thematic analysis research identified by Braun and Clarke[252]. Themes are patterns of shared meaning, characterized by a central concept. Shared topics, such as all the responses to the same interview question, do not by default fit into this narrow definition. The themes were then reviewed, and particular attention paid to thinking critically about whether the story told through the candidate themes answered our original research questions[253]. The steps to this point were iterative, and we continually checked our codes and themes against the data to ensure that we both accurately portrayed the data through our themes and addressed the outlined goals of this work. Following multiple iterations, we defined and named our final themes, which are reported

in the Results section that follows. This chapter represents our work in writing up our findings to tell a cohesive story about the interview data through the distinct but related final themes.

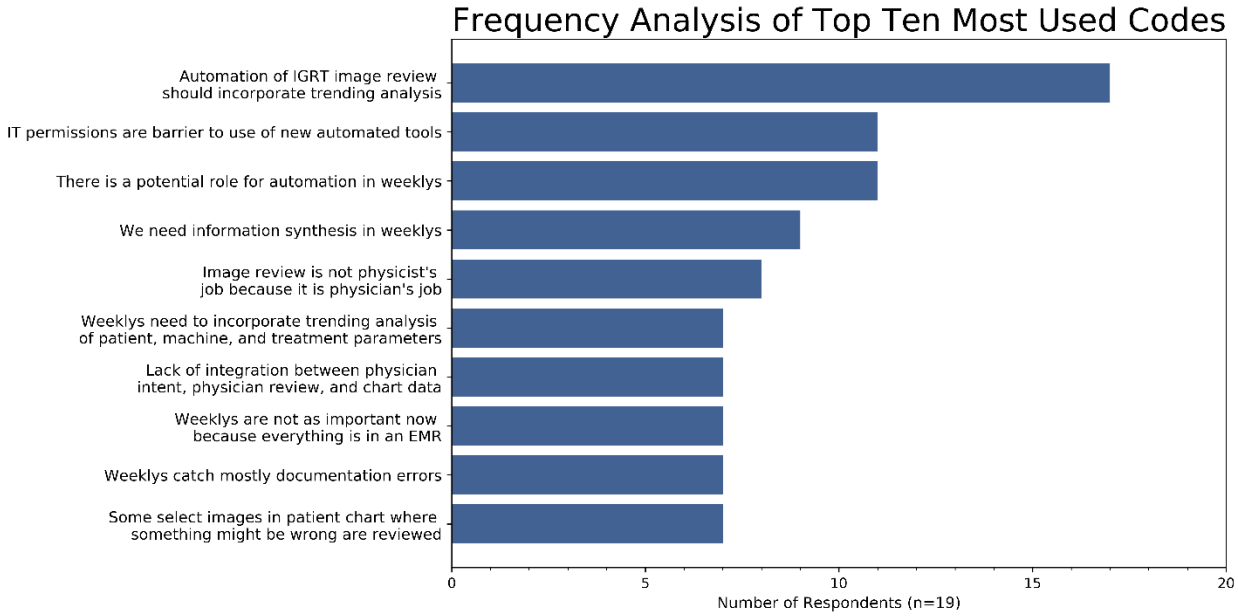

Figure 7.1: A frequency analysis of the ten most commonly used codes in the thematic analysis.

The entire research group met and discussed the interviews in order to gain a general familiarity with the topics brought up by the respondents. Two members of the group coded interviews independently, meeting regularly to discuss areas of agreement and of deviation in the coded interviews. Once nine interviews had been coded and reviewed, we determined that interviews were being coded the same by the two researchers and a single researcher coded the remaining ten interviews independently. New codes were still brought to the larger research team for discussion as they were developed. All members of the research team provided regular input as the codes (and later, themes) were developed and modified. The interdisciplinary nature of our team allowed us to maintain methodological rigor. Clinical investigators' backgrounds included medical physics graduate students, clinical medical physicists, and a physician with qualitative research training and experience.

7.3 Results

7.3.1 Thematic Saturation

We interviewed medical physicists from 16 unique institutions, located in 15 different states. Interviewees represented all geographic regions of the United States, including the Northeast, Southeast, Midwest, Southwest, and West. The saturation curve shown below in **Figure 7.2** shows the integral number of new codes encountered as a function of interview number. Guest et al. used a similar methodology to show that thematic saturation in their dataset occurred within 12 interviews[254]. Based on this curve, we determined that conducting further interviews would likely not yield a significant number of new codes and thus would likely not change our final themes.

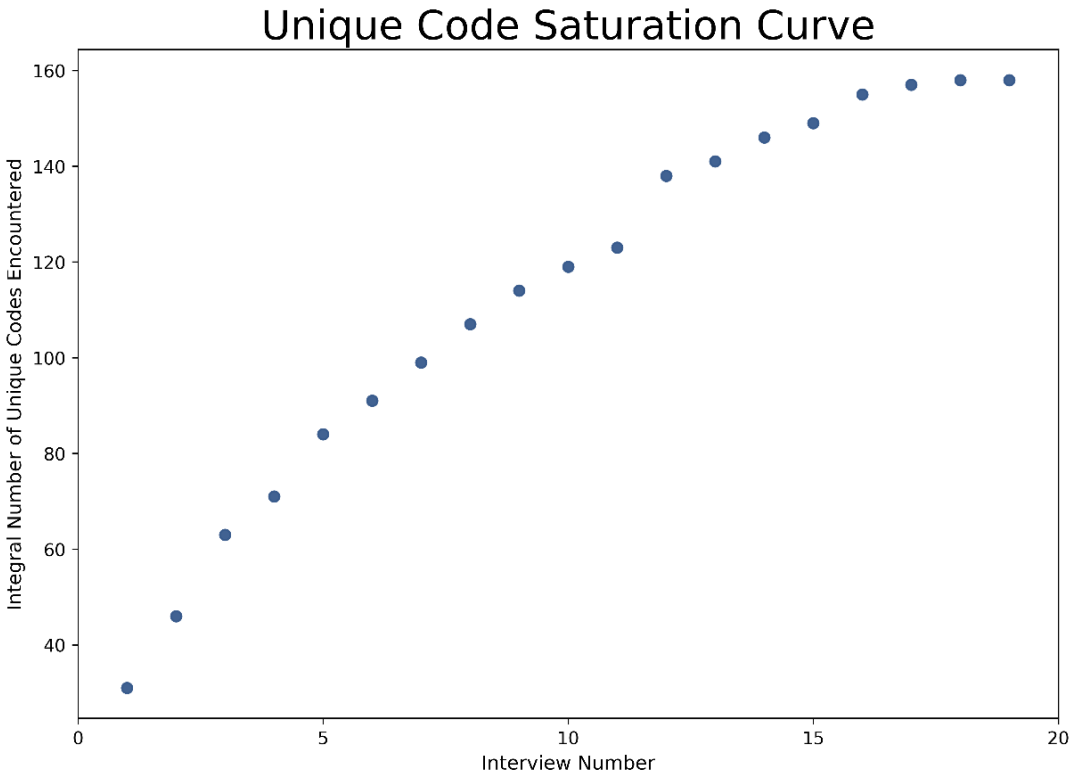

Figure 7.2: The integral number of unique codes encountered during our coding process (with 158 total unique codes identified) as a function of the interview number. From this curve, we concluded that the addition of further interviews would likely not significantly change our final themes.

7.3.2 Quantitative results

The medical physicists we interviewed reported reviewing an average of 21 ± 7 patient charts per week (range: 8 to 40). Thirteen physicists reported that they do not routinely review patient setup images during their weekly check, with only 5/19 reporting that they review all images as part of their physics check. The most commonly cited reason for not reviewing images was the physicist reporting that they consider this aspect of chart checks to be the physician's responsibility, with nine respondents explicitly voicing this sentiment. While medical physicists who receive their graduate education at CAMPEP-accredited institutions are required to receive basic training in anatomy and physiology, the guidelines offered in AAPM Report No. 365[255] are not comparable to the intense anatomy training required of physicians. When asked how much time they spend reviewing an image in a patient's chart, the responses ranged from 30 seconds to ten minutes (median: 2.5 minutes). The responses to this question included physicists who reported reviewing all images as part of their weekly chart checks along with those who reported reviewing only select images. For those who reported reviewing only select patient images, the majority voiced that the images they do review tend to be those that appear to have an issue or that are complicated cases, and thus significantly more time must be devoted to image review. The time difference in IGRT image review between these two groups is reflected in the large time range reported above. Thirteen physicists reported currently using some form of automation in their weekly chart check workflow, including both commercially-available tools and in-house software. Physicists did rank highly the value of using automation to reduce the time spent on weekly chart checks (average 6.3 on a scale from 1 to 10), but they placed significantly more

value on increasing the effectiveness of weekly chart checks (that is, catching more errors), reporting an average of 9.2 on this same scale.

7.3.3 Four major themes

Thematic analysis identified four major themes across the semi-structured interviews that we conducted.

A. Weekly chart checks need to adapt to an electronic record-and-verify chart environment

While physicists we interviewed shared that they personally had caught errors during the course of their weekly chart checks, these errors tended to be documentation errors. Physicists expressed frustration with the current weekly chart check workflow, specifically calling out the long list of documentation checks as an inefficient use of time.

"I would say the vast majority are small things like documentation, like something was left out or like a document wasn't put in or, you know, maybe something wasn't entered right." (Non-academic medical center)

"Here's what I'm going to hear that I've made a suboptimal chart check, it's because something wasn't billed correctly. Some document that is there wasn't approved. Or some document that was there was mislabeled…which has nothing to do with actually checking if the treatment went well or not." (Academic medical center)

Physicists called attention to the fact that many of the things that are still being checked during the physics weekly chart checks are holdovers from an earlier era of paper charts. Specifically, physicists mentioned many failure modes are now extremely improbable with the

electronic record-and-verify systems in place in most clinics, and yet these failure modes are still a major part of the weekly checks.

"I think nowadays it's just, there's so many things that are like more robust, more automated. There's just so much less room. You're not going to have the wrong MU, like the machine is not going to let you have the wrong MU. So things like that that could have been a big problem before are not really issues anymore, and we're still treating them like they are." (Non-academic medical center)

"We used to check SSDs and some machines we have, like with Siemens, like surface mapping. So guess what? The SSDs are always right." (Academic medical center)

"We're looking at the paper, we're calculating the number of fractions, the dose and total dose. We're doing all this manually in the past, as you remember. Now like in the Varian system, everything is calculated for you." (Non-academic medical center)

Instead, physicists said that a greater focus and priority should be given to directing cognitive resources and expertise towards investigating anomalies in patient charts, and away from menial chart check tasks.

"There should be some way to filter out things so that you pay attention to the [chart] that's important. Most of them go like clockwork. But there are those odd ones where you have to stop and think, and it would be better to devote time to those." (Academic medical center)

"And so effectively, your weekly chart checks would just be like reviewing anything that looks unusual." (Non-academic medical center)

Overall, physicists felt that current weekly chart check workflows were inefficient and of low value:

"Spending five or ten minutes on the patient just for the sake of doing that, which seems to be kind of more the approach right now, I don't think is really useful." (Academic medical center)

"It seems like a lot of time that goes into something that doesn't add a whole lot of value. And I personally, I question if that's where we should be spending our time." (Non-academic medical center)

B. Physicists have the potential to add value to patient care by analyzing images without duplicating the work done by physicians

The majority of the physicists we interviewed reported that they don't routinely review patient setup images. The most commonly given reason for this was that the physicist views IGRT image review as a physician responsibility.

"I kind of leave that to the physician and look for their feedback, if there's a problem with the images." (Academic medical center)

"We just take a glance at them then make sure they are reviewed but we're not necessarily the ones review[ing] them. It's the physicians' charge code, so it's physicians' responsibility." (Academic medical center)

An additional factor limiting physicist review of images is a perception that they are not appropriately trained to do so:

"Some of us don't even know that much of anatomy to say, you know, whether the image is good or not good. That's how I see it from my point of view, right? I'm not clinically trained in looking at anatomy and telling what's right, what's wrong." (Non-academic medical center)

"I think a physician probably cares more about details of the anatomy alignment. I'm looking at the rough error there." (Non-academic medical center)

Several interviewees reported that their interpretation of setup images could be hindered by the fact that the physician alignment priorities were not documented in a way that was easy to integrate into the weekly chart check process. Thus, for those physicists who did report reviewing images regularly or for the patient cases where a physicist image review was necessary, a significant amount of time was spent on understanding setup instructions rather than on a review of the images themselves.

"If there's an issue with the setup, I have questions like, why does this setup look this way? There's not really a good communication between what was done at the machine versus what someone can look up later to understand the decisions they made at the time of treatment." (Academic medical center)

"I think probably the thing that takes me the longest is trying to figure out what the physician has ordered." (Academic medical center)

However, many physicists still expressed that IGRT image review was one area where their technical expertise could potentially have the greatest impact on patient care and treatment quality.

"That's probably the most useful thing you can do is like, look at, verify images." (Non-academic medical center)

"So I think that the place we have the most room to gain … or to add the most value is a better review of our images. But the way we do it right now, a human looking at it and then

having to remember what they looked at last week or from two days, like the Monday versus the Tuesday image, is really ineffective." (Academic medical center)

When IGRT image review is being done by a physicist, it's currently done in isolation. Interviewees expressed that opportunities likely exist to utilize automation for a different look at image review. For example, automation could allow physicists to analyze large volumes of image data quantitatively or to look at setup images as a continuation of trending image alignment metrics. The introduction of a new algorithmic approach to image review and the interpretation of its results would be well within the purview of the clinical medical physicist, and could add a new layer of information not currently accessible, ultimately improving patient safety.

"If you have to do [image review], it can take a ton of time, and you're still left guessing at the end whether it was right or wrong. So something that's objective, that can look through a large volume of data very quickly, that would be helpful." (Academic medical center)

"One of the biggest issues is there's no way to assess qualitatively; you look at each one in a vacuum." (Academic medical center)

C. Greater support for trending analysis would increase the value of weekly checks

A lack of trending information was identified as a major shortcoming in the current weekly chart check process. Interviewees specifically reported that the large amount of data was difficult for a human to interpret without looking at it as part of a larger trend, but that such trending information is difficult to access in the current weekly chart check workflow.

"I believe that there is more value to trending different parameters from the Linac and from the imaging as you do it and from the delivery basically all together than to the actual review of the chart because everything is computerized." (Non-academic medical center)

"We have no way to trend right now in ARIA. You know, there's no automated report to say, show me all the head and neck patients that have large target volumes and their daily shifts or their daily image alignment score, right?" (Academic medical center)

"That's something that computers and the software can do really well and humans can't. With all this AI stuff—like humans, we can't trend well. But computers can, and they can get that information in a way that's useful for us." (Academic medical center)

The importance of trending was highlighted even further when interviewees talked about image review. Physicists expressed the need for a greater focus to be given to trending analysis in IGRT image review, citing bladder and rectal filling in prostate cases, weight loss and tumor shrinkage in head and neck cases, and adaptive re-planning as examples of patient cases where tracking and trending setup images over time could help them be more efficient and proactive in their chart checks.

"I, as a human, have a hard time integrating the information from one daily image to another. And I think there's a lot of information to be gained there." (Academic medical center)

"I do not know exactly how much the structure of the organs have changed. I don't have like any trending of, as I said, bladder filling or rectal filling for prostates. I don't have any tracking in terms of volume changes of the patient's GTV in a head and neck case." (Non-academic medical center)

"One issue that I have is really if you're considering re-simulating a patient due to anatomic changes, that's a very, you know, qualitative decision at this point. It would be nice to have some sort of algorithmic approach to this with some threshold." (Non-academic medical center)

D. Increased automation has the potential to make weekly checks a higher value activity

Physicists expressed that many of the things currently being checked on their weekly chart checks are documentation or numerical checks, which are prime candidates for a transition to automated checks. Opportunities likely exist to utilize automation for many of these checks and rethink what should be manually reviewed by the physicist as part of their weekly chart check process.

"I mean, we lack a lot of automation. I think there's a lot of things that can be automated, as you say, a pre-check, weekly check and all kinds of stuff can be automated." (Non-academic medical center)

"Because we lack automation to really just kind of turn over every stone [chart checks] are, you know, a medium level of activity." (Academic medical center)

"More of that kind of stuff is better and then, you know, it's safer and it takes away all the bean counting of our job, which is great." (Non-academic medical center)

One benefit of automation is that humans have a limited attention span when faced with repetitive tasks, and so might not catch all the errors present in the data. Physicists expressed this concern directly, citing that human error likely makes weekly chart checks less effective than they should be.

"I think it's better to catch those things than a human eye. And because you are doing that as a routine, you get fatigued and you skip things." (Non-academic medical center)

"Most of the time, everything works like clockwork, so you become accustomed to that. But then there are those that for some reason don't fit and you have to have either a sharp eye or just be very lucky to spot it." (Academic medical center)

Many of the physicists we interviewed expressed that a thorough weekly chart check is very effective at catching egregious errors, but it is not scalable because of the time it takes. In their view, the current weekly chart check process contains many inefficiencies and ultimately takes too much time and mental energy for every single patient check to be done thoroughly.

"I feel like the problem with the weekly chart checks is number over quality." (Academic medical center)

"In some ways I'm impressed and some other ways I'm like, well, how much time does it take spent on a weekly check to catch that? And how much is actually luck and how much is there in parallel that we don't catch. I mean, to me, it always seems like it's the visible tip of the iceberg, and if we caught that there were probably quite a few other things." (Academic medical center)

Interviewees pointed out that a key opportunity for automation was the flagging of anomalies for further investigation. Such tools could allow for physicists' time and cognitive resources to be directed in a more focused manner on investigating the complex patient cases, ultimately improving patient safety in the clinic overall.

"An ideal setup for me for chart check would be something that is really good at flagging suspicious things." (Academic medical center)

"With the machine, automation can kick in to replace that part and we spend time to investigate the real problem, that will be nice." (Non-academic medical center)

"We'll catch errors more consistently than all of us doing slightly different weekly chart check. Second we'll be efficient. And so we free up time for us to do something more constructive than repetitive work." (Academic medical center)

7.3.4 Barriers to use

Three broad categories of potential barriers to use of automation in the weekly chart check workflow were reported in response to our semi-structured interview questions on the topic: existing clinic environment, workload and time concerns, and tool-specific technical factors. The most commonly reported barriers to use in each of these categories are shown in **Figure 7.3**. The single most commonly identified barrier to clinical use of an automated tool was overcoming their clinic's IT permissions, with 11 out of the 19 respondents reporting this barrier could hinder their adoption of such tools. Other existing clinical environment factors that could be barriers to implementation included the cost of a new tool and software fatigue (meaning a wide variety of software is already in use in their clinic, and there is low motivation to introduce more). A second area of concern for physicists was that their clinic's adoption of automated tools could, perversely, lead to increased workload on the physics staff. Physicists reported concern that the time needed to trust algorithm results, the time needed to implement a new tool, and the time needed to operate the software and interpret results could all lead to increased strain on an already busy physics staff. Third, physicists pointed to several tool-specific technical factors that could present barriers to clinical adoption. Algorithm reliability (or lack thereof), vendor-

agnostic integration, and a lack of trust in “black box” machine learning algorithms were all mentioned as potential barriers to clinical use.

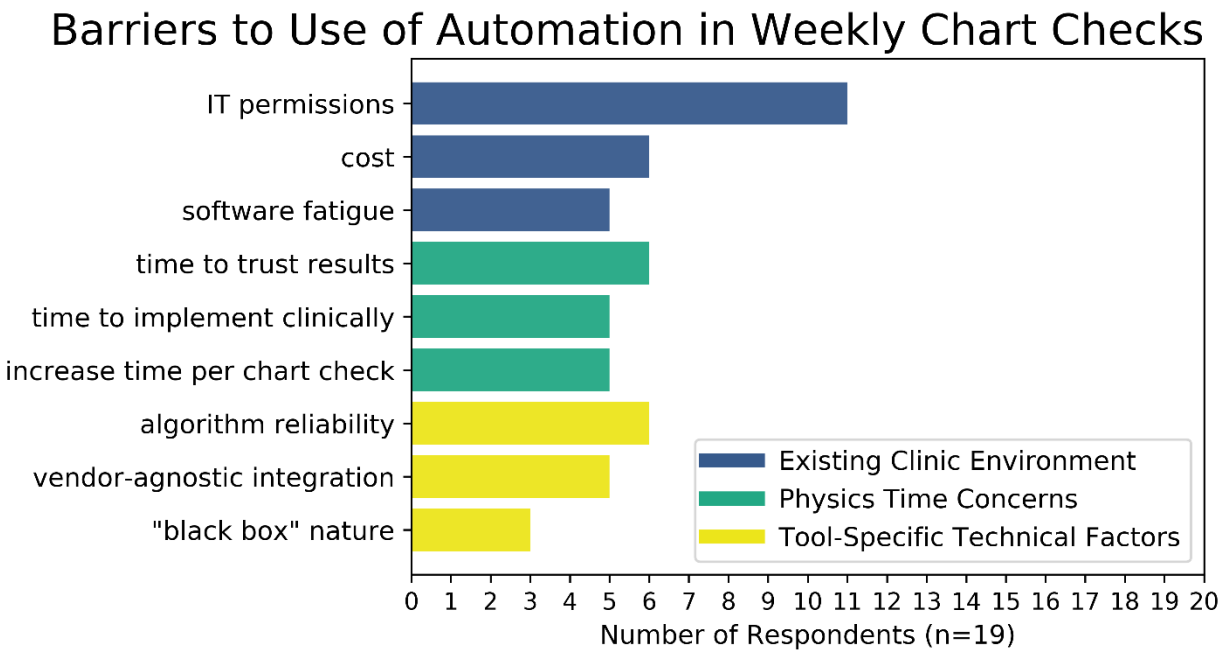

Figure 7.3: The most commonly reported barriers to the use of automation in the weekly chart check workflow.

7.4 Discussion

Our results show that most physicists believe that change of their weekly chart check procedures is urgently needed. As technology change in the radiation oncology clinic has continued and indeed accelerated over the past few decades, weekly checks have not kept pace. Physicists pointed out many checks currently done as a matter of routine practice in their weekly chart checks that no longer add to patient safety the way they did when they were first introduced. This reflects a key recommendation of AAPM TG-275, which points to the need for a TG-100 FMEA approach to be taken in regards to the physics weekly chart checks. It is noteworthy that none of the physicists we interviewed directly mentioned the need for FMEA to

be considered in weekly chart checks, nor was TG-275 mentioned even indirectly. TG-275 and the associated practice guidelines are relatively recent; perhaps it will simply take time for this recommendation to be widely adopted.

Physicists rated the value of increasing the effectiveness of weekly chart checks higher than decreasing the time spent. This sentiment could be attributed to physicists' dedication to safety or, due to the perfunctory nature of many current chart check tasks, physicists may be spending minimal time on weekly checks to begin with. Physicists we interviewed spoke to the human error and mental fatigue aspects of weekly chart checks as limitations in their value. Clearly, physicists feel that chart checks have imperfect sensitivity to detect errors, and evidence exists in the literature to support this[62,219,226].

Our data suggest that the field of medical physics could examine the potential for physicists to be more involved in image review. Not all physicists perform image review as part of their weekly chart check tasks, as we have shown in this chapter. It is beyond the scope of this work to set forth guidelines for the clinical medical physicist regarding image review, but the results presented here may help inform the debate of whether and how medical physicists should be involved in the process. Currently, physicists are under ever increasing time pressures[256,257], which means that more and greater complexity chart checks are being done with less resources. Physicists have the potential to add value in a different, perhaps more algorithmic, approach to image review. The physicists we interviewed expressed the sentiment that if physicist resources could effectively be freed up for image review, this could be a valuable additional layer of safety in the clinical workflow. Weekly chart checks currently take a good deal of physicists' time to

complete, but if the routine checks could be automated or otherwise made more efficient then physicist efforts could be more realistically dedicated to analyzing new image review metrics.

Physicists we interviewed felt that there was a potential role for automation in the weekly chart check workflow to supplement the human checks, and did not express concern about automation replacing them in the clinic. Specifically, physicists reported the desire for more automated tools that flag anomalies, allowing them to analyze complex cases and investigate complex problems. AAPM TG-275 suggests that software vendors should continue to work towards the development of automated tools that can assist with chart review tasks, while also being cognizant of the limits of automation. Automation bias is a real and well-documented concern[217,218], and the development of automated tools should be coupled with research into minimizing the associated automation bias. It is noteworthy that only one physicist cited job security as a barrier to use of automated tools in their weekly chart checks. This contrasts strongly with job security concerns related to automated treatment planning, which is a considerable source of worry for dosimetrists with our work in **Chapter 6** finding that 21/34 dosimetrists explicitly voiced this concern[258]. While computers and automation can assist in the process and make the weekly checks more efficient and more effective, physicists still seem to feel that there will be a place for them in the clinic. This sentiment fits neatly within one of the stated trajectories of Medical Physics 3.0 (MP3.0)[259], namely Sustainability, which argues for a redistribution of the medical physicist's responsibilities in order to better pursue value-based goals. Physicists can preserve a place for themselves in the clinic by advocating for new automated tools that complement their role and allow them to shift their focus to complex tasks, thereby increasing their value to the clinic overall.

When interviewees were asked about what features of a new automated weekly chart check tool would be important, every single one of them highlighted the need for trending analysis. Many of the checks currently being performed in the weekly chart check workflow are performed in isolation, and the physicists we interviewed spoke of the value that trending analysis could add. As more and more automated tools are put forward by industry, our data suggest that examining what sort of trending analysis support can be offered by such tools could be useful to clinical medical physicists. Physics checks encompass a huge volume of data, which is potentially an opening for AI-based dimensionality reduction techniques to assist with a more effective chart review. While automation can perhaps support a greater emphasis on trending analysis, physicists pointed out that their expertise would still be needed in the clinic to interpret such results.

Even as the applications of automated technologies continue to expand, there is still a gap between research findings and the clinical implementation of those findings. As new automated technologies are developed, a conscious effort should be made to study the barriers to use and ensure that advances in research translate to advances in clinical care. We found that the most commonly reported barriers to use of automated weekly chart check tools can be broken down into the following categories: existing clinic environment, workload and time concerns, and tool-specific technical factors. As researchers and developers continue to automate ever more of the physics chart check tasks, a focused effort should be given to understanding these barriers and the implications for clinical adoption of their automated tools.

Our study has a number of limitations. While we attempted to reach a diverse group of medical physicists through our recruitment efforts, we are cognizant of the limitations we faced.

Only two of our 19 respondents reported employment at community clinics, although we did reach seven physicists employed at governmental health clinics. Taking these two groups together, we interviewed nine physicists employed at non-academic medical centers in comparison with the ten physicists employed at academic medical centers. Our respondents represented 16 unique institutions and 15 states from all regions of the country (Northeast, Southeast, Midwest, Southwest, and West), although our cohort was limited to clinical physicists currently practicing in the United States. The conclusions we reach in this study may not translate well to other countries, where the chart check requirements could be vastly different for clinical medical physicists. By soliciting participants based on professional contacts, we may have introduced a selection bias in the physicists we considered for this study, but this does not necessarily translate to a true self-selection bias in the respondents. Of the 21 clinical physicists who were contacted about this study, only two declined to participate or did not respond. Related to selection bias, we may have interviewed physicists with strong opinions on weekly chart checks or automation research, and not reached those who are ambivalent on the topics. Finally, we acknowledge that our sample size of 19 respondents is not large. This is not atypical for a qualitative research study; studies published in recent years investigating Diversity, Equity, and Inclusion in radiation oncology[260] and resilience among medical physics residents[261] have conducted semi-structured interviews with cohort sizes of 26 and 32, respectively. In addition, a rigorous thematic analysis includes depth of data collection and achieving thematic saturation, more than the specific number of people interviewed. Our structured interview script was detailed, requiring 30 minutes to complete, and each interview provided a wealth of information. As noted above, during the course of our thematic analysis of these 19 interviews, we concluded

that thematic saturation had been achieved and that further interviews would likely not yield additional themes.

7.5 Conclusion

In this work, we use a novel thematic analysis to identify the shortcomings of the weekly chart check process from the perspective of the clinic medical physicist. We describe four major themes, which each complement the findings and recommendations of AAPM TG-275. Clinical medical physicists described both a need for greater automation in the weekly check process and a sentiment that the process itself must adapt in light of increasingly automated systems. As automated technologies continue to become increasingly prevalent in the clinic, the FMEA approach advocated by TG-275 and the value-based care focus advocated by MP3.0 suggest that the current way of doing weekly chart checks needs to be re-evaluated. This would allow for more effective physics chart checks that emphasize follow-up, trending analysis, FMEA, and other higher value tasks that improve patient safety.

CHAPTER 8: CONCLUSIONS AND FUTURE WORK

8.1 Summary of work

The goal of Specific Aim 1 was to develop novel tools for the automatic detection of patient misalignments in daily setup images. In **Chapter 2** we developed a CNN-based model to automatically detect off-by-one vertebral body misalignments in patients treated to the thoracic spine. We established the necessity of using image data from a multi-institutional collaboration in order to improve model performance, increasing the area under the ROC curve from 0.942 to 0.992 with the incorporation of training data from all institutions. At the 95% specificity, the leave-one-institution-out models achieved a mean sensitivity of 92.9% in detecting off-by-one vertebral body misalignments. The model sensitivities ranged from 85.5% to 99.8%, suggesting that there are real quantifiable differences in the images from different institutions. An updated method for generating off-by-one vertebral misalignments that are consistent with the stereoscopic geometry of the ExacTrac system was developed in **Chapter 3**. A multi-input CNN trained on this geometrically-realistic data obtained an AUC of 0.988 as compared to 0.975 for a model trained on independent planar image sets from the same patients. We observed that the sensitivity at 99% specificity decreased from 67.9% to 43.8% by using stereoscopic data in our model training, but that the sensitivity increased from 91.9% to 96.4% at the 95% specificity level. Overall, the model performance did not degrade with the transition to clinically-realistic image data used for model training and testing. In **Chapter 4**, we evaluated the performance of a model developed to detect more generic patient misalignments. A multi-input model trained and tested on random translational shifts in all anatomic regions achieved an area under the ROC

curve of 0.970 in detecting shifts of 1 cm from treatment isocenter. At the 95% specificity, the model achieved a sensitivity of 94.7% in detecting these translational misalignments.

A novel method for quantifying the IGRT error rate at UCLA was developed to address Specific Aim 2. In **Chapter 5**, we applied the models developed in **Chapter 3** and **Chapter 4** to retrospective image data collected from patients' daily setup imaging in order to search for previously unreported treatment errors and near miss events. A treatment error resulting from overly aggressive masking during x-ray to DRR image registration was discovered as a result of applying the generic multi-input model. Manual review determined that this patient was misaligned for 7 of 30 fractions, and that the magnitude of the misalignment was approximately 1 cm at the base of the skull. We calculated an error rate per fraction of 0.06% for UCLA, which is in line with the well under 1% figure commonly cited in the literature. We also identified two previously unreported near miss events, both involving mix-ups of patient names. Based on these incidents, we calculated a near-miss rate per fraction of 0.02%, although no data currently exists in the literature with which to compare this figure.

In Specific Aim 3, we evaluated the barriers and facilitators to implementing new automated technologies clinically. **Chapter 6** addressed the barriers to use of auto-contouring and automated treatment planning tools as reported by medical dosimetrists. The most commonly reported barriers to use were contour inaccuracy and the inability to easily modify automated plans. Dosimetrists also expressed explicit job security concerns, with 21/34 worrying that automated tools would hurt the job market. Cluster analysis showed that dosimetrists employed at community-based clinics were more likely to report that lack of access to automated tools was an important barrier to use than their peers employed at academic medical centers. The

current weekly physics chart check workflow was evaluated in **Chapter 7** to understand how a new tool designed to assist with IGRT image review could best be clinically integrated. Thematic analysis was used to generate four distinct themes from semi-structured interviews of clinical medical physicists: 1. weekly chart checks need to adapt to an electronic record-and-verify chart environment, 2. physicists have the potential to add value to patient care by analyzing images without duplicating the work done by physicians, 3. greater support for trending analysis would increase the value of weekly checks, and 4. increased automation has the potential to make weekly checks a higher value activity. Physicists ranked highly the value of using automation to reduce the time spent on weekly chart checks (average 6.3 on a scale from 1 to 10), but they placed significantly more value on increasing the effectiveness of weekly chart checks (that is, catching more errors), reporting an average of 9.2 on this same scale.

8.2 Future directions

When training our CNN-based model to detect vertebral body misalignments, we originally treated each set of x-ray/ DRR image pairs independently. Image data was only considered in-plane, ignoring the stereoscopic geometry of the ExacTrac system. **Chapter 3** describes the preliminary stages of incorporating the stereoscopic geometry into the generation of synthetically shifted training data. Extending this methodology to a multi-institutional collaboration of institutions and larger dataset would likely show even further improvements in the model performance. Our multi-input model to detect more general setup errors was trained on the limited set of generated 1 cm translational errors. Introducing different, perhaps subtler, types of errors in the data generated for training purposes could allow for the detection of a wider range of patient misalignments that could occur in the clinic.

Both of the models we developed for automatic detection of patient misalignments were applied to retrospective image data from patients who had already completed their treatments. Five years of treatments were analyzed for treatment errors, but future work could either expand the time frame investigated at UCLA or investigate retrospective data from outside institutions to better quantify the true IGRT error rate. In addition, it would perhaps be more powerful or clinically useful to integrate both into the clinical workflow for real-time alerts of patient misalignments. Future work on integrating the models to provide feedback at the time of treatment is worth exploring.

Our survey study of medical dosimetrists only investigated reported barriers and facilitators to use of automated tools that are commercially available. Expanding the scope of the questions to include automated tools currently in development could provide valuable feedback and directions to researchers regarding features that would be helpful to the end user. Future work using qualitative research methods to explore the motivations behind the reported barriers to use in greater depth could also prove illuminating. The semi-structured interviews we conducted with clinical medical physicists allowed us to understand the current weekly chart check workflow and how a proposed automated tool could fit into this pipeline. Follow up studies investigating the results of implementing such a tool and quantifying the clinical impact should be explored.

APPENDIX

A.1 Survey distributed to medical dosimetrists

Prior Use:

1. I have used the following types of automation tools for radiation oncology (select all that apply). **Please select all options that you have used at any point in your career**, even if you do not currently use that tool or if you have used it infrequently.

- Deep learning-based auto-contouring (for example, Mirada DLCExpert, MIM ContourProtege-AI)
- Atlas-based and/or model-based auto-contouring algorithms (for example, MIM Atlas Segment, Varian Velocity, RayStation MABS/MBS, Pinnacle SPICE, Elekta ABAS)
- Knowledge-based plan quality assessment (for example, Sun Nuclear PlanIQ)
- Automated planning using knowledge-based planning (KBP) algorithms (for example, Varian RapidPlan)
- Automated field-in-field planning (for example, Radformation EZFluence)
- Automated planning using rule-based or template-based algorithms (for example, Phillips Pinnacle Auto-Planning, Raysearch Raystation Auto-Planning)
- Automated planning using other algorithm not listed above
- Automated planning, specific algorithm unknown

1b. If you selected “other algorithm”, please specify the algorithm used in the field below:

Auto-Contouring:

2. Please rate your level of experience with auto-contouring from 1 (not familiar at all) to 5 (extremely familiar)

- 1
- 2
- 3
- 4
- 5

3. I have used deep learning-based auto-contouring (e.g. Mirada DLCExpert, MIM ContourProtege-AI) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast
- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

4. I have used atlas-based and/or model-based auto-contouring (e.g. MIM Atlas Segment, Varian Velocity, RayStation MABS/MBS, Pinnacle SPICE, Elekta ABAS) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast

- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

5. How often do you use auto-contouring?

- Daily
- Weekly
- Once or twice a month
- Less than once a month
- Never

6. Please rank the following options to indicate where you have heard about auto-contouring. Enter 1 for where you've heard about it the most and 4 for where you've heard about it the least.

- Scientific talks at professional meetings
- Vendor booths at professional meetings
- Peers at other clinics
- Colleagues at my own workplace

7. I would be more likely to use auto-contouring algorithms if (select all that apply):

- My clinic purchased an auto-contouring product
- My site physicist and/or supervisor provided more support of auto-contouring
- Auto-contouring algorithms were more accurate.

- Auto-contouring algorithms could produce contours that followed my institution's contouring guidelines.
- Auto-contouring algorithms would leave image slices blank rather than producing a contour in those slices that then needed to be heavily edited
- Auto-contouring algorithms could tell me specific parts of the contour that needed my attention
- Other

7b. Please specify other things that would make you more likely to use auto-contouring algorithms in the field below:

8. I dislike auto-contouring because (select all that apply):

- I would rather start contouring from scratch than have to modify auto-contours, even if the time taken is about the same
- I am concerned that the algorithm will make a contouring error that I won't catch
- Other

8b. Please specify other reasons you dislike auto-contouring in the field below:

9. Please rate your level of agreement with the following statements based on your personal experience using **deep learning-based auto-contouring** (e.g. Mirada DLCExpert, MIM ContourProtege-AI). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- Modifying auto-contours takes longer than creating the contours from scratch

- The contours produced by the deep learning-based algorithm are often so incorrect that auto-contouring adds nothing of value
- Auto-contouring saves enough time for me to think that it's worth using
- I am concerned that use of auto-contouring could lead to treatment errors; for example if the algorithm contours an organ incorrectly and leads to an overdose of that organ
- I believe that deep learning-based auto-contouring is ready for routine clinical use
- When I use deep learning-based auto-contouring, I spend a lot of time checking the contours produced by the algorithm

10. Please rate your level of agreement with the following statements based on your personal experience using **atlas-based and/or model-based auto-contouring** (e.g. MIM Atlas Segment, Varian Velocity, RayStation MABS/MBS, Pinnacle SPICE, Elekta ABAS). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- Modifying auto-contours takes longer than creating the contours from scratch
- The contours produced by the atlas-based/ model-based algorithm(s) are often so incorrect that auto-contouring adds nothing of value
- Auto-contouring saves enough time for me to think that it's worth using
- I am concerned that use of auto-contouring could lead to treatment errors; for example if the algorithm contours an organ incorrectly and leads to an overdose of that organ
- I believe that atlas-based/ model-based auto-contouring is ready for routine clinical use

- When I use atlas-based/ model-based auto-contouring, I spend a lot of time checking the contours produced by the algorithm

Automated Treatment Planning:

11. Please rate your level of experience with automated treatment planning **in general** from 1 (not familiar at all) to 5 (extremely familiar)

- 1
- 2
- 3
- 4
- 5

12. I have used knowledge-based plan quality assessment (e.g. Sun Nuclear PlanIQ) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast
- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

13. I have used automated planning using KBP algorithms (e.g. Varian RapidPlan) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast
- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

14. I have used automated field-in-field planning (e.g. Radformation EZFluence) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast
- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

15. I have used automated planning using rule-based or template-based algorithms (e.g. Phillips Pinnacle Auto-Planning, Raysearch Raystation Auto-Planning) for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast

- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

16. I have used automated treatment planning for the following body sites (select all that apply):

- Head and neck
- Thorax
- Breast
- Pelvis (prostate, bladder, rectal, etc.)
- Extremities
- Intra-cranial
- None of the above

17. How often do you use automated treatment planning?

- Daily
- Weekly
- Once or twice a month
- Less than once a month
- Never

18. I would be more likely to use automated treatment planning if (select all that apply):

- I was provided with more information about how to use the tools

- I was provided with more information about how the algorithms work behind the scenes
- My clinic purchased an automated treatment planning product
- My site physicist and/or supervisor provided more support and/or more training on automated treatment planning
- The automated treatment planning algorithm produced a plan that was easier to modify or "tweak" to get the optimal plan
- The automated treatment planning algorithm produced better plans that didn't need any modifications by the dosimetrist
- Automated treatment planning was available for a wider range of types of treatment plans
- Other

18b. Please specify other things that would make you more likely to use automated treatment planning in the field below:

19. I like automated treatment planning because (select all that apply):

- It allows me to work through a higher patient caseload
- I have a higher degree of confidence in the quality of the plans that I am submitting to the prescribing physician
- Other

19b. Please specify other reasons that you like automated treatment planning in the field below:

20. I dislike automated treatment planning because (select all that apply):

- I do not believe that the plans are of the same quality as those generated by experienced dosimetrists

- Modifying automated plans takes more time than generating a comparable quality plan from scratch
- I enjoy plan optimization and I don't want to give up that part of my job
- Other

20b. Please specify other reasons that you dislike automated treatment planning in the field below:

21. Please rate your level of agreement with the following statements based on your personal experience using **knowledge-based plan quality assessment** (e.g. Sun Nuclear PlanIQ). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- Knowledge-based plan quality assessment leads to higher quality treatment plans
- Knowledge-based plan quality assessment saves time by helping me know when my plan is good enough so that I can stop optimizing
- I am concerned that use of knowledge-based plan quality assessment could lead to treatment errors; for example if the algorithm makes an error and I don't catch it.
- Knowledge-based plan quality assessment decreases the amount of time I spend on any single patient case.
- I believe that knowledge-based plan quality assessment is ready for routine clinical use.

22. Please rate your level of agreement with the following statements based on your personal experience using **KBP automated treatment planning** (e.g. RapidPlan). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- KBP automated planning leads to higher quality treatment plans
- KBP automated planning saves time by helping me know when my plan is good enough so that I can stop optimizing
- I am concerned that use of KBP automated planning could lead to treatment errors; for example if the algorithm makes an error and I don't catch it.
- KBP automated planning decreases the amount of time I spend on any single patient case.
- I believe that KBP automated planning is ready for routine clinical use.

23. Please rate your level of agreement with the following statements based on your personal experience using **automated field-in-field planning** (e.g. EZFluence). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- Automated field-in-field planning leads to higher quality treatment plans
- I am concerned that use of automated field-in-field planning could lead to treatment errors; for example if the algorithm makes an error and I don't catch it.
- Automated field-in-field planning decreases the amount of time I spend on any single patient case.

- I believe that automated field-in-field planning is ready for routine clinical use.

24. Please rate your level of agreement with the following statements based on your personal experience using **automated treatment planning using rule-based or template-based algorithms** (e.g. Pinnacle Auto-Planning, RayStation Auto-Planning). If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- The automated treatment planning tool I have used leads to higher quality treatment plans
- I am concerned that use of automated treatment planning could lead to treatment errors; for example if the algorithm makes an error and I don't catch it.
- Automated treatment planning decreases the amount of time I spend on any single patient case.
- I believe that the automated treatment planning tool I have used is ready for routine clinical use.

25. Please rate your level of agreement with the following statements based on your personal experience using automated treatment planning. If you feel that the statement does not apply to your personal experience, select "not relevant to me."

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me

- The automated treatment planning tool I have used leads to higher quality treatment plans
- I am concerned that use of automated treatment planning could lead to treatment errors; for example if the algorithm makes an error and I don't catch it.

- Automated treatment planning decreases the amount of time I spend on any single patient case.
- I believe that the automated treatment planning tool I have used is ready for routine clinical use.

General level of Agreement:

26. Please rate your level of agreement with the following statements:

Options: disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, not relevant to me.

- I believe auto-contouring will continue to get better and by the end of my career, most or all normal tissue contours will be automatically generated.
- I worry that automated treatment planning will hurt the job market for dosimetrists.
- If automation reduces the time to make a plan, I will just get more plans and be even busier.
- I worked hard to gain my treatment planning skills and I value them highly. To see them devalued would be a disappointment.
- I am concerned that routinely using automated treatment planning will cause me to get out of practice on planning difficult patient cases.
- I believe automated treatment planning algorithms will continue to get better and by the end of my career will replace most manual treatment plan optimization.
- Planning goals at my clinic aren't very standardized.
- I would want to use automated treatment planning tools if I knew they worked well.

Demographics:

27. What is your age?

- 20-29
- 30-39
- 40-49
- 50-59
- 60+
- Prefer not to answer

28. How many years have you been employed as a medical dosimetrist?

- Less than 5
- 5-9
- 10-19
- 20+

29. Are you certified by the Medical Dosimetrist Certification Board?

- Yes
- No

30. What is your gender?

- Male
- Female
- Other/ non-binary
- Prefer not to answer

31. What is your highest level of education?

- Associate's degree
- Bachelor's degree
- Master's degree
- Doctorate
- Prefer not to answer

32. Which of the following best describes your current clinical environment?

- Non hospital-based community practice
- Hospital-based non-academic medical center
- Academic medical center

33. How many radiotherapy treatment machines are in use at your clinic (including any satellite clinics)?

- 1
- 2-4
- 5-8
- 9+

A.2 Semi-structured interview script

First I will be asking you a few questions about your current workflow:

1. How many chart checks do you perform per week?
2. How many setup/treatment images do you review per week?
    a. About how many images are generated per patient per week?
    b. Do you review kV and/or MV planar images? Do you review CBCT images?

c. Do you review them in Offline Review, or a similar "live" system that allows review of the fusion, or do you review them in PDF form only?

3. Please talk me through how long you think you might spend reviewing the images per patient for a weekly check. Consider all time costs, including the time it takes to open the patient in the review software, load each image, and close the patient.

    Prompts:

    a. How fast does it go with images that are easy to check? What fraction of images is this?

    b. How fast does it go with images that are more subtle to check?

4. Do you use any products that automate weekly chart checking? Automate pre-treatment checks?

    a. Prompt with Radformation ChartCheck, ClearCheck, Varian Chart QA.

5. What are the shortcomings of the current weekly chart check image review process in your view?

Our group is developing software to work with the ARIA Oncology Information System from Varian. It interrogates the treatment database at the end of the day, identifies new images, and uses an AI tool to assess the quality of the alignment and flag any anomalies. A summary report then is automatically generated for all patients.

(Show Powerpoint slides here)

6. We think that the two most useful potential aspect of the tool are 1) to reduce the amount of time spent on weekly checks without increasing error rate; and 2) to increase the

effectiveness of weekly checks (catch more errors). Please rate the relative importance of the following:

a. Reduce the amount of time spent on weekly checks without increasing error rate (1-10 1 being the lowest importance and 10 being the highest)
b. Increase the effectiveness of weekly checks (catch more errors) 1-10

Prompts/probes about why?

a. Tell us about how you feel about the value of weekly chart checks? Do weekly chart checks frequently catch errors?
b. Can you tell us an error you've caught with weekly chart checks? Was it correctible? Did it lead to a process change?

As we develop the tool, we are looking to understand what would be most helpful for people who are going to use it.

7. How would you prefer to interact with this tool?

a. Program with a GUI that you start up when you want to use it?
b. Receive an emailed report daily?
c. Receive an emailed report at a configurable time interval?
d. Any other method?

8. There are a number of features we are considering including in this tool. I am going to review a list of them to see whether you would find them useful or not and why.

a. Would you like to see an alignment score for every image? Why or why not?

b. Would you prefer that the tool flags images with alignment scores falling below a pre-determined threshold, or would you prefer for it to flag images with the lowest percentile alignment scores (so that it always flags a fixed percentage of images)?
c. Would you like a graphical display of alignment scores? Why or why not?
d. Would you like to see per-patient trendlines? Why or why not?
e. Would you like the tool to include the images themselves, or would you revert to Offline Review to review flagged images? Why or why not?
f. Are there any other features you would suggest that I have not already mentioned?

9. If you were using this tool, what might make you more comfortable decreasing the amount of time reviewing images not flagged by the tool?

We are also interested in understanding more about potential barriers to using the tool.

10. What barriers to your use of the tool might you anticipate?
    a. One potential barrier is IT permissions. If the software tool could be easily loaded onto your Varian network, how difficult do you anticipate it would be to obtain permission to install the software? (probe if difficult) What would make it difficult? (Probe if not difficult) What would make it not too difficult?
    b. Existence of appropriate computational infrastructure?
    c. Consider a recent example of technology adoption in radiation therapy that was (or was not) successful. What factors determined that technology's success? What were the barriers to adoption that were or were not overcome?
11. Would you be interested in using this tool if it was available?
    a. (If yes) why? (If not), why not?

b. Probe the issue of whether they review every single image or not – does this help improve safety or reduce time or both?

12. Do you have any further comments, suggestions, concerns, or ideas regarding this tool?

REFERENCES

1. Hall EJ, Giaccia A. (2006). *Radiobiology for the Radiologist*. Lippincott Williams & Wilkins.
2. Barton MB, Jacob S, Shafiq J, et al. Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012. *Radiother Oncol*. 2014;112(1):140-144. doi:10.1016/j.radonc.2014.03.024
3. Delaney G, Jacob S, Featherstone C, Barton M. The role of radiotherapy in cancer treatment. *Cancer*. 2005;104(6):1129-1137. doi:10.1002/cncr.21324
4. Yap ML, Zubizarreta E, Bray F, Ferlay J, Barton M. Global access to radiotherapy services: Have we made progress during the past decade? *J Glob Oncol*. 2016;2(4):207-215. doi:10.1200/jgo.2015.001545
5. Nutting C, Dearnaley DP, Webb S. Intensity modulated radiation therapy: A clinical review. *Brit J Radiol*. 2014;73(869):459-469. doi:10.1259/bjr.73.869.10884741
6. Baskar R, Lee KA, Yeo R, Yeoh KW. Cancer and radiation therapy: Current advances and future directions. *Int J Med Sci*. 2012;9(3):193. doi:10.7150/ijms.3635
7. Xing L, Thorndyke B, Schreibmann E, et al. Overview of image-guided radiation therapy. *Med Dosim*. 2006;31(2):91-112. doi:10.1016/j.meddos.2005.12.004
8. Sterzing F, Engenhart-Cabillic R, Flentje M, Debus J. Image-guided radiotherapy: A new dimension in radiation oncology. *Dtsch Arztebl Int*. 2011;108(16):274. doi:10.3238/arztebl.2011.0274
9. Jaffray D, Kupelian P, Djemil T, Macklis RM. Review of image-guided radiation therapy. *Expert Rev Anticancer Ther*. 2007;7(1):89-103. doi:10.1586/14737140.7.1.89

10. Chen GTY, Sharp GC, Mori S. A review of image-guided radiotherapy. *Radiol Phys Technol*. 2009;2(1):1-12. doi:10.1007/s12194-008-0045-y

11. De Los Santos J, Popple R, Agazaryan N, et al. Image guided radiation therapy (IGRT) technologies for radiation therapy localization and delivery. *Int J Radiat Oncol*. 2013;87(1):33-45. doi:10.1016/j.ijrobp.2013.02.021

12. Ibbott GS. The need for, and implementation of, image guidance in radiation therapy. *Ann ICRP*. 2018;47(3-4):160-176. doi:10.1177/0146645318764092

13. Simpson DR, Lawson JD, Nath SK, Rose BS, Mundt AJ, Mell LK. A survey of image-guided radiation therapy use in the United States. *Cancer*. 2010;116(16):3953-3960. doi:10.1002/cncr.25129

14. Nabavizadeh N, Elliott DA, Chen Y, et al. Image guided radiation therapy (IGRT) practice patterns and IGRT's impact on workflow and treatment planning: Results from a national survey of american society for radiation oncology members. *Int J Radiat Oncol Biol Phys*. 2016;94(4):850-857. doi:10.1016/j.ijrobp.2015.09.035

15. van Herk M. Different styles of image-guided radiotherapy. *Semin Radiat Oncol*. 2007;17(4):258-267. doi:10.1016/j.semradonc.2007.07.003

16. Jaffray DA. Image-guided radiotherapy: From current concept to future perspectives. *Nat Rev Clin Oncol 2012 912*. 2012;9(12):688-699. doi:10.1038/nrclinonc.2012.194

17. Lamba M, Breneman JC, Warnick RE. Evaluation of image-guided positioning for frameless intracranial radiosurgery. *Int J Radiat Oncol*. 2009;74(3):913-919. doi:10.1016/j.ijrobp.2009.01.008

18. Verellen D, Ridder M De, Linthout N, Tournel K, Soete G, Storme G. Innovations in

image-guided radiotherapy. *Nat Rev Cancer 2007 712*. 2007;7(12):949-960. doi:10.1038/nrc2288

19. Dawson LA, Eccles C, Bissonnette JP, Brock KK. Accuracy of daily image guidance for hypofractionated liver radiotherapy with active breathing control. *Int J Radiat Oncol*. 2005;62(4):1247-1252. doi:10.1016/j.ijrobp.2005.03.072
20. Balter JM, Brock KK, Litzenberg DW, et al. Daily targeting of intrahepatic tumors for radiotherapy. *Int J Radiat Oncol*. 2002;52(1):266-271. doi:10.1016/S0360-3016(01)01815-6
21. Hong TS, Tomé WA, Chappell RJ, Chinnaiyan P, Mehta MP, Harari PM. The impact of daily setup variations on head-and-neck intensity-modulated radiation therapy. *Int J Radiat Oncol*. 2005;61(3):779-788. doi:10.1016/j.ijrobp.2004.07.696
22. Dawson LA, Sharpe MB. Image-guided radiotherapy: Rationale, benefits, and limitations. *Lancet Oncol*. 2006;7(10):848-858. doi:10.1016/S1470-2045(06)70904-4
23. Webster A, Appelt AL, Eminowicz G. Image-guided radiotherapy for pelvic cancers: A review of current evidence and clinical utilisation. *Clin Oncol*. 2020;32(12):805-816. doi:10.1016/j.clon.2020.09.010
24. Wang S, Tang W, Luo H, Jin F, Wang Y. The role of image-guided radiotherapy in prostate cancer: A systematic review and meta-analysis. *Clin Transl Radiat Oncol*. 2023;38:81-89. doi:10.1016/j.ctro.2022.11.001
25. Gwynne S, Webster R, Adams R, Mukherjee S, Coles B, Staffurth J. Image-guided radiotherapy for rectal cancer — A systematic review. *Clin Oncol*. 2012;24(4):250-260. doi:10.1016/j.clon.2011.07.012

26. Ren X-C, Liu Y-E, Li J, Lin Q. Progress in image-guided radiotherapy for the treatment of non-small cell lung cancer. *World J Radiol*. 2019;11(3):46. doi:10.4329/wjr.V11.I3.46

27. Kearney M, Coffey M, Leong A. A review of image guided radiation therapy in head and neck cancer from 2009–2019 – Best practice recommendations for RTTs in the clinic. *Tech Innov Patient Support Radiat Oncol*. 2020;14:43-50. doi:10.1016/j.tipsro.2020.02.002

28. Dhont J, Harden S V., Chee LYS, Aitken K, Hanna GG, Bertholet J. Image-guided radiotherapy to manage respiratory motion: Lung and liver. *Clin Oncol*. 2020;32(12):792-804. doi:10.1016/j.clon.2020.09.008

29. Fuss M, Boda-Heggemann J, Papanikolau N, Salter BJ. Image-guidance for stereotactic body radiation therapy. *Med Dosim*. 2007;32(2):102-110. doi:10.1016/j.meddoc.2007.01.007

30. Zelefsky MJ, Kollmeier M, Cox B, et al. Improved clinical outcomes with high-dose image guided radiotherapy compared with non-IGRT for the treatment of clinically localized prostate cancer. *Int J Radiat Oncol*. 2012;84(1):125-129. doi:10.1016/j.ijrobp.2011.11.047

31. Bujold A, Craig T, Jaffray D, Dawson LA. Image-guided radiotherapy: Has it influenced patient outcomes? *Semin Radiat Oncol*. 2012;22(1):50-61. doi:10.1016/j.semradonc.2011.09.001

32. Bissonnette JP, Medlam G. Trend analysis of radiation therapy incidents over seven years. *Radiother Oncol*. 2010;96(1):139-144. doi:10.1016/j.radonc.2010.05.002

33. Grégoire V, Guckenberger M, Haustermans K, et al. Image guidance in radiation therapy

for better cure of cancer. *Mol Oncol*. 2020;14(7):1470-1491. doi:10.1002/1878-0261.12751

34. Hall EJ, Wuu CS. Radiation-induced second cancers: The impact of 3D-CRT and IMRT. *Int J Radiat Oncol*. 2003;56(1):83-88. doi:10.1016/S0360-3016(03)00073-7

35. Qi XS, Albuquerque K, Bailey S, et al. Quality and safety considerations in image guided radiation therapy: An ASTRO safety white paper update. *Pract Radiat Oncol*. 2023;13(2):97-111. doi:10.1016/j.prro.2022.09.004

36. Jaffray DA, Langen KM, Mageras G, et al. Safety considerations for IGRT: Executive summary. *Pract Radiat Oncol*. 2013;3(3):167-170. doi:10.1016/j.prro.2013.01.004

37. Jin JY, Ryu S, Faber K, et al. 2D/3D Image fusion for accurate target localization and evaluation of a mask based stereotactic system in fractionated stereotactic radiotherapy of cranial lesions. *Med Phys*. 2006;33(12):4557-4566. doi:10.1118/1.2392605

38. Yin FF, Ryu S, Ajlouni M, et al. Image-guided procedures for intensity-modulated spinal radiosurgery: Technical note. *J Neurosurg*. 2004;101(Supplement3):419-424. doi:10.3171/sup.2004.101.supplement3.0419

39. Jin J-Y, Yin F-F, Tenn SE, Medin PM, Solberg TD. Use of the BrainLAB ExacTrac x-ray 6D system in image-guided radiotherapy. *Med Dosim*. 2008;33(2):124-134. doi:10.1016/j.meddos.2008.02.005

40. Yan H, Yin FF, Kim JH. A phantom study on the positioning accuracy of the Novalis Body system. *Med Phys*. 2003;30(12):3052-3060. doi:10.1118/1.1626122

41. Ackerly T, Lancaster CM, Geso M, Roxby KJ. Clinical accuracy of ExacTrac intracranial frameless stereotactic system. *Med Phys*. 2011;38(9):5040-5048. doi:10.1118/1.3611044

42. Jin JY, Ryu S, Rock J, et al. Evaluation of residual patient position variation for spinal radiosurgery using the Novalis image guided system. *Med Phys*. 2008;35(3):1087-1093. doi:10.1118/1.2839097

43. Montgomery C, Collins M. An evaluation of the BrainLAB 6D ExacTrac/Novalis Tx System for image-guided intracranial radiotherapy. *J Radiother Pract*. 2017;16(3):326-333. doi:10.1017/S1460396917000139

44. Lee SW, Jin JY, Guan H, Martin F, Kim JH, Yin FF. Clinical assessment and characterization of a dual-tube kilovoltage X-ray localization system in the radiotherapy treatment room. *J Appl Clin Med Phys*. 2008;9(1):1-15. doi:10.1120/jacmp.v9i1.2318

45. Gevaert T, Verellen D, Engels B, et al. Clinical evaluation of a robotic 6-degree of freedom treatment couch for frameless radiosurgery. *Int J Radiat Oncol*. 2012;83(1):467-474. doi:10.1016/j.ijrobp.2011.05.048

46. Rahimian J, Chen JCT, Girvigian MR, Miller MJ, Rahimian R. (2011) Frame-based and frameless accuracy of Novalis® Radiosurgery. In AA de Salles (Ed.). *Shaped Beam Radiosurgery: State of the Art* (1st ed., 37-46). Springer, Berlin.

47. Verbakel WFAR, Lagerwaard FJ, Verduin AJE, Heukelom S, Slotman BJ, Cuijpers JP. The accuracy of frameless stereotactic intracranial radiosurgery. *Radiother Oncol*. 2010;97(3):390-394. doi:10.1016/j.radonc.2010.06.012

48. Gevaert T, Verellen D, Tournel K, et al. Setup accuracy of the novalis ExacTrac 6DOF system for frameless radiosurgery. *Int J Radiat Oncol Biol Phys*. 2012;82(5):1627-1635. doi:10.1016/j.ijrobp.2011.01.052

49. Teh BS, Paulino AC, Lu HH, et al. Versatility of the Novalis System to deliver image-

guided stereotactic body radiation therapy (SBRT) for various anatomical sites. *Technol Cancer Res Treat*. 2007;6(4):347-354. doi:10.1177/153303460700600412

50. Watchman CJ, Hamilton RJ, Stea B, Mignault AJ. Patient positioning using implanted gold markers with the Novalis Body system in the thoracic spine. *Neurosurgery*. 2008;62(5 Suppl). doi:10.1227/01.neu.0000325938.08605.eb

51. Agazaryan N, Tenn SE, F Desalles AA, Selch MT. Image-guided radiosurgery for spinal tumors: Methods, accuracy and patient intrafraction motion. *Phys Med Biol*. 2008;53(6):1715. doi:10.1088/0031-9155/53/6/015

52. Russakoff DB, Rohlfing T, Ho A, et al. Evaluation of intensity-based 2D-3D spine image registration using clinical gold-standard data. *Lect Notes Comput Sci*. 2003;2717:151-160. doi:10.1007/978-3-540-39701-4_16

53. Kohn LT, Corrigan JM, Donaldson MS, eds. (2000). To err is human: Building a safer health system. Washington, DC: National Academies Press.

54. Reason J. Human error: Models and management. *BMJ*. 2000;320(7237):768-770. doi:10.1136/bmj.320.7237.768

55. Reason J. The contribution of latent human failures to the breakdown of complex systems. *Philos Trans R Soc London B, Biol Sci*. 1990;327(1241):475-484. doi:10.1098/rstb.1990.0090

56. Perrow C. (1984). *Normal accidents: Living with high-risk technologies*. Princeton, NJ: Princeton University Press.

57. Reason J. (1997). *Managing the risks of organizational accidents*. Singapore: Ashgate.

58. Ford EC, Gaudette R, Myers L, et al. Evaluation of safety in a radiation oncology setting

using failure mode and effects analysis. *Int J Radiat Oncol*. 2009;74(3):852-858. doi:10.1016/j.ijrobp.2008.10.038

59. Boadu M, Rehani MM. Unintended exposure in radiotherapy: Identification of prominent causes. *Radiother Oncol*. 2009;93(3):609-617. doi:10.1016/j.radonc.2009.08.044

60. Clark BG, Brown RJ, Ploquin JL, Kind AL, Grimard L. The management of radiation treatment error through incident learning. *Radiother Oncol*. 2010;95(3):344-349. doi:10.1016/j.radonc.2010.03.022

61. Bogdanich W. *Radiation offers new cures, and ways to do harm*. New York: The New York Times, 2010.

62. Ezzell G, Chera B, Dicker A, et al. Common error pathways seen in the RO-ILS data that demonstrate opportunities for improving treatment safety. *Pract Radiat Oncol*. 2018;8(2):123-132. doi:10.1016/j.prro.2017.10.007

63. Clarity PSO. RO-ILS Quarterly Report for Q3 2016. 2017; Available from: https://www.astro.org/uploadedFiles/_MAIN_SITE/Patient_Care/Patient_Safety/ROILS/Content_Pieces/Q32016Report.pdf

64. Huq MS, Fraass BA, Dunscombe PB, et al. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. *Med Phys*. 2016;43(7):4209-4262. doi:10.1118/1.4947547

65. Yeung TK, Bortolotto K, Cosby S, Hoar M, Lederer E. Quality assurance in radiotherapy: Evaluation of errors and incidents recorded over a 10 year period. *Radiother Oncol*. 2005;74(3):283-291. doi:10.1016/j.radonc.2004.12.003

66. Klein EE, Drzymala RE, Purdy JA, Michalski J. Errors in radiation oncology: A study in

pathways and dosimetric impact. *J Appl Clin Med Phys*. 2005;6(3):81-94. doi:10.1120/jacmp.v6i3.2105

67. Kutcher GJ, Mageras GS, Liebel SA. Control, correction, and modeling of setup errors and organ motion. *Semin Radiat Oncol*. 1995;5(2):134-145. doi:10.1054/srao00500134

68. Hunt MA, Kutcher GJ, Burman C, et al. The effect of setup uncertainties on the treatment of nasopharynx cancer. *Int J Radiat Oncol Biol Phys*. 1993;27(2):437-447. doi:10.1016/0360-3016(93)90257-V

69. Yamashita H, Haga A, Hayakawa Y, et al. Patient setup error and day-to-day esophageal motion error analyzed by cone-beam computed tomography in radiation therapy. *Acta Oncol (Madr)*. 2010;49(4):485-490. doi:10.3109/02841861003652574

70. Algan O, Jamgade A, Ali I, et al. The dosimetric impact of daily setup error on target volumes and surrounding normal tissue in the treatment of prostate cancer with intensity-modulated radiation therapy. *Med Dosim*. 2012;37(4):406-411. doi:10.1016/j.meddos.2012.03.003

71. Kaur I, Rawat S, Ahlawat P, et al. Dosimetric impact of setup errors in head and neck cancer patients treated by image-guided radiotherapy. *J Med Phys*. 2016;41(2):144-148. doi:10.4103/0971-6203.181640

72. Bogdanich W. *Case Studies: When Medical Radiation Goes Awry*. New York: The New York Times, 2010.

73. Bogdanich W. *As Technology Surges, Radiation Safeguards Lag*. New York: The New York Times, 2010.

74. Williams MV. Radiotherapy near misses, incidents and errors: Radiotherapy incident at

Glasgow. *Clin Oncol.* 2007;19(1):1-3. doi:10.1016/j.clon.2006.12.004

75. Leveson NG, Turner CS. An investigation of the Therac-25 accidents. *IEEE Comput.* 1993;26(7):18-41. doi:10.1109/mc.1993.274940

76. Huang G, Medlam G, Lee J, et al. Error in the delivery of radiation therapy: Results of a quality assurance review. *Int J Radiat Oncol Biol Phys*. 2005;61(5):1590-1595. doi:10.1016/j.ijrobp.2004.10.017

77. Ford EC, Fong de Los Santos L, Pawlicki T, Sutlief S, Dunscombe P. Consensus recommendations for incident learning database structures in radiation oncology. *Med Phys*. 2012;39(12):7272-7290. doi:10.1118/1.4764914

78. Mutic S, Brame RS, Oddiraju S, et al. Event (error and near-miss) reporting and learning system for process improvement in radiation oncology. *Med Phys*. 2010;37(9):5027-5036. doi:10.1118/1.3471377

79. Nyflot MJ, Zeng J, Kusano AS, et al. Metrics of success: Measuring impact of a departmental near-miss incident learning system. *Pract Radiat Oncol*. 2015;5(5):e409-e416. doi:10.1016/j.prro.2015.05.009

80. Arnold A, Ward I, Gandhidasan S. Incident review in radiation oncology. *J Med Imaging Radiat Oncol*. 2022;66(2):291-298. doi:10.1111/1754-9485.13358

81. Pawlicki T, Coffey M, Milosevic M. Incident earning systems for radiation oncology: Development and value at the local, national and international level. *Clin Oncol*. 2017;29(9):562-567. doi:10.1016/j.clon.2017.07.009

82. Ford EC, Evans SB. Incident learning in radiation oncology: A review. *Med Phys*. 2018;45(5):e100-e119. doi:10.1002/mp.12800

83. Yorke E, Gelblum D, Ford E. Patient safety in external beam radiation therapy. *Am J Roentgenol*. 2011;196(4):768-772. doi:10.2214/ajr.10.6006

84. Pham JC, Girard T, Pronovost PJ. What to do with healthcare incident reporting systems. *J Public health Res*. 2013;2(3):jphr.2013.e27. doi:10.4081/jphr.2013.e27

85. Marks LB, Jackson M, Xie L, et al. The challenge of maximizing safety in radiation oncology. *Pract Radiat Oncol*. 2011;1(1):2-14. doi:10.1016/j.prro.2010.10.001

86. Hendee WR, Herman MG. Improving patient safety in radiation oncology. *Med Phys*. 2011;38(1):78-82. doi:10.1118/1.3522875

87. Covington EL, Popple RA, Cardan RA. Technical note: Use of automation to eliminate shift errors. *J Appl Clin Med Phys*. 2020;21(3):192-195. doi:10.1002/acm2.12830

88. Jensen N, Boye K, Damkjær S, Wahlstedt I. Impact of automation in external beam radiation therapy treatment plan quality control on error rates and productivity. *Int J Radiat Oncol*. 2018;102(3):S149-S150. doi:10.1016/j.ijrobp.2018.06.362

89. Santhanam A, Dou H, Kurihara A, et al. Three-dimensional feature recognition-based automated patient treatment mismatch verification system for radiation therapy. *Int J Radiat Oncol*. 2012;84(3):S742. doi:10.1016/j.ijrobp.2012.07.1984

90. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans Med Imaging*. 2016;35(5):1299-1312. doi:10.1109/tmi.2016.2535302

91. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88. doi:10.1016/j.media.2017.07.005

92. Abdou MA. Literature review: Efficient deep neural networks techniques for medical

image analysis. *Neural Comput Appl*. 34. doi:10.1007/s00521-022-06960-9

93. Lamb JM, Agazaryan N, Low DA. Automated patient identification and localization error detection using 2-dimensional to 3-dimensional registration of kilovoltage X-ray setup images. *Int J Radiat Oncol Biol Phys*. 2013;87(2):390-393. doi:10.1016/j.ijrobp.2013.05.021

94. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans Med Imaging*. 2018;37(12):2663-2674. doi:10.1109/tmi.2018.2845918

95. Zhao W, Shen L, Islam MT, et al. Artificial intelligence in image-guided radiotherapy: A review of treatment target localization. *Quant Imaging Med Surg*. 2021;11(12):4881. doi:10.21037/qims-21-199

96. Zhao W, Han B, Yang Y, et al. Incorporating imaging information from deep neural network layers into image guided radiation therapy (IGRT). *Radiother Oncol*. 2019;140:167. doi:10.1016/j.radonc.2019.06.027

97. Davis D, Evans M, Jadad A, et al. The case for knowledge translation: Shortening the journey from evidence to effect. *BMJ*. 2003;327(7405):33-35. doi:10.1136/bmj.327.7405.33

98. Lang ES, Wyer PC, Haynes RB. Knowledge translation: Closing the evidence-to-practice gap. *Ann Emerg Med*. 2007;49(3):355-363. doi:10.1016/j.annemergmed.2006.08.022

99. Morris ZS, wooding S, Grant J. The answer is 17 years, what is the question: Understanding time lags in translational research. *J R Soc Med*. 2011;104(12):510. doi:10.1258/jrsm.2011.110180

100. Woolf SH. The meaning of translational research and why it matters. *JAMA*. 2008;299(2):211-213. doi:10.1001/jama.2007.26

101. Bauer MS, Damschroder L, Hagedorn H, Smith J, Kilbourne AM. An introduction to implementation science for the non-specialist. *BMC Psychol*. 2015;3(1):1-12. doi:10.1186/S40359-015-0089-9

102. Tansella M, Thornicroft G. Implementation science: Understanding the translation of evidence into practice. *Br J Psychiatry*. 2009;195(4):283-285. doi:10.1192/bjp.bp.109.065565

103. Tucker S, McNett M, Mazurek Melnyk B, et al. Implementation science: Application of evidence-based practice models to improve healthcare quality. *Worldviews Evidence-Based Nurs*. 2021;18(2):76-84. doi:10.1111/wvn.12495

104. Bauer MS, Kirchner JA. Implementation science: What is it and why should I care? *Psychiatry Res*. 2020;283:112376. doi:10.1016/j.psychres.2019.04.025

105. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. *MIS Q Manag Inf Syst*. 2003;27(3):425-478. doi:10.2307/30036540

106. Braithwaite J, Marks D, Taylor N. Harnessing implementation science to improve care quality and patient safety: A systematic review of targeted literature. *Int J Qual Heal Care*. 2014;26(3):321-329. doi:10.1093/intqhc/mzu047

107. Handley MA, Gorukanti A, Cattamanchi A. Strategies for implementing implementation science: A methodological overview. *Emerg Med J*. 2016;33(9):660-664. doi:10.1136/emermed-2015-205461

108. Powell BJ, Fernandez ME, Williams NJ, et al. Enhancing the impact of implementation strategies in healthcare: A research agenda. *Front Public Heal*. 2019 Jan 22;7:3. doi:10.3389/fpubh.2019.00003

109. Wensing M. Implementation science in healthcare: Introduction and perspective. *Z Evid Fortbild Qual Gesundhwes*. 2015;109(2):97-102. doi:10.1016/j.zefq.2015.02.014

110. Edwards GF, Zagarese V, Tulk Jesso S, Jesso M, Harden SM, Parker SH. Designing healthcare for human use: Human factors and practical considerations for the translational process. *Front Heal Serv*. 2022;2. doi:10.3389/frhs.2022.981450

111. Lorden AL, Zhang Y, Lin SH, Côté MJ. Measures of success: The role of human factors in lean implementation in healthcare. *Qual Manag J*. 2014;21(3):26-37. doi:10.1080/10686967.2014.11918394

112. Epstein N. A perspective on wrong level, wrong side, and wrong site spine surgery. *Surg Neurol Int*. 2021;12. doi:10.25259/sni_402_2021

113. Klein EE, Drzymala RE, Purdy JA, Michalski J. Errors in radiation oncology: A study in pathways and dosimetric impact. *J Appl Clin Med Phys*. 2005;6(3):81-94. doi:10.1120/jacmp.2025.25355

114. Shafiq J, Barton M, Noble D, Lemer C, Donaldson LJ. An international review of patient safety measures in radiotherapy practice. *Radiother Oncol*. 2009;92(1):15-21. doi:10.1016/j.radonc.2009.03.007

115. Pennsylvania Patient Safety Advisory. Errors in radiation therapy. 2009; Available from: https://patientsafety.pa.gov/ADVISORIES/Documents/200909_87.pdf

116. Halvorsen PH, Cirino E, Das IJ, et al. AAPM-RSS Medical Physics Practice Guideline

9.a. for SRS-SBRT. *J Appl Clin Med Phys*. 2017;18(5):10-21. doi:10.1002/acm2.12146

117. de los Santos EF, Evans S, Ford EC, et al. Medical Physics Practice Guideline 4.a: Development, implementation, use and maintenance of safety checklists. *J Appl Clin Med Phys*. 2015;16(3):37-59. doi:10.1120/jacmp.v16i3.5431

118. Fourcade A, Blache JL, Grenier C, Bourgain JL, Minvielle E. Barriers to staff adoption of a surgical safety checklist. *BMJ Qual Saf*. 2012;21(3):191-197. doi:10.1136/bmjqs-2011-000094

119. Garnerin P, Arés M, Huchet A, Clergue F. Verifying patient identity and site of surgery: Improving compliance with protocol by audit and feedback. *Qual Saf Heal Care*. 2008;17(6):454-458. doi:10.1136/qshc.2007.022301

120. Kearns RJ, Uppal V, Bonner J, Robertson J, Daniel M, McGrady EM. The introduction of a surgical safety checklist in a tertiary referral obstetric centre. *BMJ Qual Saf*. 2011;20(9):818-822. doi:10.1136/bmjqs.2010.050179

121. Minnesota Department of Health. Adverse Health Events in Minnesota: 13th Annual Public Report. 2017; Available from: https://www.health.state.mn.us/patientsafety/ae/2017ahereport.pdf

122. Mallett R, Conroy M, Saslaw LZ, Moffatt-Bruce S. Preventing wrong site, procedure, and patient events using a common cause analysis. *Am J Med Qual*. 2012;27(1):21-29. doi:10.1177/1062860611412066

123. Shiraishi S, Grams MP, Fong de los Santos LE. Image-guided radiotherapy quality control: Statistical process control using image similarity metrics. *Med Phys*. 2018;45(5):1811-1821. doi:10.1002/mp.12859

124. Mody MG, Nourbakhsh A, Stahl DL, Gibbs M, Alfawareh M, Garges KJ. The prevalence of wrong level surgery among spine surgeons. *Spine (Phila Pa 1976)*. 2008;33(2):194-198. doi:10.1097/BRS.0b013e31816043d1

125. Groff MW, Heller JE, Potts EA, Mummaneni P V., Shaffrey CI, Smith JS. A survey-based study of wrong-level lumbar spine surgery: The scope of the problem and current practices in place to help avoid these errors. *World Neurosurg*. 2013;79(3-4):585-592. doi:10.1016/j.wneu.2012.03.017

126. Rattan R, Kataria T, Banerjee S, et al. Artificial intelligence in oncology, its scope and future prospects with specific reference to radiation oncology. *BJR Open*. 2019;1(1):20180031. doi:10.1259/bjro.20180031

127. Weidlich V, Weidlich GA. Artificial intelligence in medicine and radiation oncology. *Cureus*. 2018;10(4). doi:10.7759/cureus.2475

128. Feng M, Valdes G, Dixit N, Solberg TD. Machine learning in radiation oncology: Opportunities, requirements, and needs. *Front Oncol*. 2018 Apr 17;8:110. doi:10.3389/fonc.2018.00110

129. Luk SMH, Ford EC, Phillips MH, Kalet AM. Improving the quality of care in radiation oncology using artificial intelligence. *Clin Oncol*. 2022;34(2):89-98. doi:10.1016/j.clon.2021.11.011

130. Lessmann N, van Ginneken B, de Jong PA, Išgum I. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Med Image Anal*. 2019;53:142-155. doi:10.1016/j.media.2019.02.005

131. Xia L, Xiao L, Quan G, Bo W. 3D Cascaded convolutional networks for multi-vertebrae

segmentation. *Curr Med Imaging Former Curr Med Imaging Rev*. 2020;16(3):231-240. doi:10.2174/1573405615666181204151943

132. Forsberg D, Sjöblom E, Sunshine JL. Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data. *J Digit Imaging*. 2017;30(4):406-412. doi:10.1007/s10278-017-9945-x

133. Löffler MT, Jacob A, Scharr A, et al. Automatic opportunistic osteoporosis screening in routine CT: Improved prediction of patients with prevalent vertebral fractures compared to DXA. *Eur Radiol*. 2021;31(8):6069-6077. doi:10.1007/s00330-020-07655-2

134. Ito S, Ando K, Kobayashi K, et al. Automated detection of spinal schwannomas utilizing deep learning based on object detection from magnetic resonance imaging. *Spine (Phila Pa 1976)*. 2021;46(2):95-100. doi:10.1097/brs.0000000000003749

135. Roggen T, Bobic M, Givehchi N, Scheib SG. Deep learning model for markerless tracking in spinal SBRT. *Phys Medica*. 2020;74:66-73. doi:10.1016/j.ejmp.2020.04.029

136. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems 25*. 2012.

137. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology imaging. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1159-1161. doi:10.1016/j.ijrobp.2018.05.070

138. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiother Oncol*. 2018;129(3):421-426. doi:10.1016/j.radonc.2018.05.030

139. Qu B, Cao J, Qian C, et al. Current development and prospects of deep learning in spine

image analysis: A literature review. *Quant Imaging Med Surg*. 2022;12(6):3454-3479. doi:10.21037/qims-21-939

140. Penney GP, Weese J, Little JA, Desmedt P, Hill DLG, Hawkes DJ. A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE Trans Med Imaging*. 1998;17(4):586-595. doi:10.1109/42.730403

141. Romero M, Interian Y, Solberg T, Valdes G. Targeted transfer learning to improve performance in small medical physics datasets. *Med Phys*. 2020;47(12):6246-6256. doi:10.1002/mp.14507

142. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2015. doi:10.48550/arXiv.1409.1556

143. Srivastava N, Hinton G, Krizhevsky A, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research.* 2014;15(1):1929-1958.

144. Bengio Y. Practical recommendations for gradient-based training of deep architectures. *Lect Notes Comput Sci*. 2012;7700:437-478. doi:10.1007/978-3-642-35289-8_26

145. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR), 2015.* doi:10.48550/arXiv.1412.6980

146. Sendelbach S, Funk M. Alarm fatigue. *AACN Adv Crit Care*. 2013;24(4):378-386. doi:10.4037/nci.0b013e3182a903f9

147. Ruskin KJ, Hueske-Kraus D. Alarm fatigue. *Curr Opin Anaesthesiol*. 2015;28(6):685-690. doi:10.1097/ACO.0000000000000260

148. Deb S, Claudio D. Alarm fatigue and its influence on staff performance. *IIE Trans*

Healthc Syst Eng. 2015;5(3):183-196. doi:10.1080/19488300.2015.1062065

149. Reijnders-Thijssen P, Geerts D, van Elmpt W, Pawlicki T, Wallis A, Coffey M. Prevalence of software alerts in radiotherapy. *Tech Innov Patient Support Radiat Oncol*. 2020;14:32-35. doi:10.1016/j.tipsro.2020.04.002

150. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition.* 2010:248-255. doi:10.1109/cvpr.2009.5206848

151. Morid MA, Borjali A, Del Fiol G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput Biol Med*. 2021;128:104115. doi:10.1016/j.compbiomed.2020.104115

152. Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of TrueBeam developer mode for imaging QA. *J Appl Clin Med Phys*. 2015;16(4):322-333. doi:10.1120/jacmp.v16i4.5363

153. Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning-A review of recent innovations. *Br J Radiol*. 2018;91(1092). doi:10.1259/bjr.20180270

154. Covington EL, Chen X, Younge KC, et al. Improving treatment plan evaluation with automation. *J Appl Clin Med Phys*. 2016;17(6):16-31. doi:10.1120/jacmp.v17i6.6322

155. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. *Semin Radiat Oncol*. 2019;29(3):185-197. doi:10.1016/j.semradonc.2019.02.001

156. Perrier L, Morelle M, Pommier P, et al. Cost of prostate image-guided radiation therapy:

Results of a randomized trial. *Radiother Oncol*. 2013;106(1):50-58. doi:10.1016/j.radonc.2012.11.011

157. Gevaert T, Verellen D, Tournel K, et al. Setup accuracy of the Novalis ExacTrac 6DOF system for frameless radiosurgery. *Int J Radiat Oncol*. 2012;82(5):1627-1635. doi:10.1016/j.ijrobp.2011.01.052

158. Schmidhalter D, Malthaner M, Born EJ, et al. Assessment of patient setup errors in IGRT in combination with a six degrees of freedom couch. *Z Med Phys*. 2014;24(2):112-122. doi:10.1016/j.zemedi.2013.11.002

159. Charters JA, Bertram P, Lamb JM. Offline generator for digitally reconstructed radiographs of a commercial stereoscopic radiotherapy image-guidance system. *J Appl Clin Med Phys*. 2022;23(3):e13492. doi:10.1002/acm2.13492

160. Yoo TS, Ackerman MJ, Lorensen WE, et al. Engineering and algorithm design for an image processing API: A technical report on ITK - The Insight Toolkit. *Studies in Health Technology and Informatics*. 2002:586-592. doi:10.3233/978-1-60750-929-5-586

161. Sun Y, Zhu L, Wang G, Zhao F. Multi-input convolutional neural network for flower grading. *J Electr Comput Eng*. 2017;2017. doi:10.1155/2017/9240407

162. Petragallo R, Bertram P, Halvorsen P, et al. Development and multi-institutional validation of a convolutional neural network to detect vertebral body mis-alignments in 2D x-ray setup images. *Med Phys*. 2023;50(5):2662-2671. doi:10.1002/mp.16359

163. Potters L, Gaspar LE, Kavanagh B, et al. American Society for Therapeutic Radiology and Oncology (ASTRO) and American College of Radiology (ACR) practice guidelines for image-guided radiation therapy (IGRT). *Int J Radiat Oncol Biol Phys*. 2010;76(2):319-

325. doi:10.1016/j.ijrobp.2009.09.041

164. Pillai M, Adapa K, Das SK, et al. Using artificial intelligence to improve the quality and safety of radiation therapy. *J Am Coll Radiol*. 2019;16:1267-1272. doi:10.1016/j.jacr.2019.06.001

165. Hadley SW, Kessler ML, Litzenberg DW, et al. SafetyNet: Streamlining and automating QA in radiotherapy. *J Appl Clin Med Phys*. 2016;17(1):387-395. doi:10.1120/jacmp.v17i1.5920

166. Luximon DC, Ritter T, Fields E, et al. Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. *Med Phys*. 2022;49(10):6410-6423. doi:10.1002/mp.15927

167. Luximon DC, Neylon J, Ritter T, et al. Results of an AI-based image review system to detect patient misalignment errors in a multi-institutional database of CBCT-guided radiotherapy treatments. *Int J Radiat Oncol*. Published online March 12, 2024. doi:10.1016/j.ijrobp.2024.02.065

168. Charters JA, Luximon D, Petragallo R, Neylon J, Low DA, Lamb JM. Automated detection of vertebral body misalignments in orthogonal kV and MV guided radiotherapy: Application to a comprehensive retrospective dataset. *Biomed Phys Eng Express*. 2024;10(2):025039. doi:10.1088/2057-1976/ad2baa

169. Potters L, Ford E, Evans S, Pawlicki T, Mutic S. A systems approach using big data to improve safety and quality in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2016;95(3):885-889. doi:10.1016/j.ijrobp.2015.10.024

170. Huser V, Cimino JJ. Impending challenges for the use of big data. *Int J Radiat Oncol Biol*

Phys. 2016;95(3):890-894. doi:10.1016/j.ijrobp.2015.10.060

171. Rosenstein BS, Capala J, Efstathiou JA, et al. How will big data improve clinical and basic research in radiation therapy? *Int J Radiat Oncol Biol Phys*. 2016;95(3):895-904. doi:10.1016/j.ijrobp.2015.11.009

172. Lustberg T, Van Soest J, Jochems A, Deist T, Van Wijk Y, Walsh S. Big Data in radiation therapy: Challenges and opportunities. *Br J Radiol*. 2017;90. doi:10.1259/bjr.20160689

173. Mazur LM, Mosaly PR, Jackson M, et al. Quantitative assessment of workload and stressors in clinical radiation oncology. *Int J Radiat Oncol*. 2012;83(5):e571-e576. doi:10.1016/j.ijrobp.2012.01.063

174. Johnson J, Ford E, Yu J, Buckey C, Fogh S, Evans SB. Peer support: A needs assessment for social support from trained peers in response to stress among medical physicists. *J Appl Clin Med Phys*. 2019;20(9):157-162. doi:10.1002/acm2.12675

175. Chen E, Arnone A, Sillanpaa JK, Yu Y, Mills MD. A special report of current state of the medical physicist workforce — results of the 2012 ASTRO Comprehensive Workforce Study. *J Appl Clin Med Phys*. 2015;16(3):399-405. doi:10.1120/jacmp.v16i3.5232

176. Mutic S, Brame RS, Oddiraju S, et al. Event (error and near-miss) reporting and learning system for process improvement in radiation oncology. *Med Phys*. 2010;37(9):5027-5036. doi:10.1118/1.3471377

177. Pillai M, Adapa K, Das SK, et al. Using artificial intelligence to improve the quality and safety of radiation therapy. *J Am Coll Radiol*. 2019;16:1267-1272. doi:10.1016/j.jacr.2019.06.001

178. Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning

applications in radiation oncology. *Phys Imaging Radiat Oncol*. 2021;19:13-24. doi:10.1016/j.phro.2021.05.007

179. Su C, Tu T, Han P, Lakshminarayanan P, Mcnutt T. Deep learning approaches for anomalies detection of bladder CT contours in prostate cancer patients. *J Mech Med Biol*. 2022;22(8):2240033. doi:10.1142/S0219519422400334

180. McNutt TR, Moore KL, Wu B, Wright JL. Use of big data for quality assurance in radiation therapy. *Semin Radiat Oncol*. 2019;29(4):326-332. doi:10.1016/j.semradonc.2019.05.006

181. McNutt TR, Moore KL, Quon H. Needs and challenges for big data in radiation oncology. *Int J Radiat Oncol*. 2016;95(3):909-915. doi:10.1016/j.ijrobp.2015.11.032

182. Sari ABA, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: Retrospective patient case note review. *BMJ*. 2007;334(7584):79. doi:10.1136/bmj.39031.507153.ae

183. Arnold A, Delaney GP, Cassapi L, Barton M. The use of categorized time-trend reporting of radiation oncology incidents: A proactive analytical approach to improving quality and safety over time. *Int J Radiat Oncol Biol Phys*. 2010;78(5):1548-1554. doi:10.1016/j.ijrobp.2010.02.029

184. Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys*. 2011;81(2):545-551. doi:10.1016/j.ijrobp.2010.11.030

185. Zhu X, Ge Y, Li T, Thongphiew D, Yin F-F, Wu QJ. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys*. 2011;38(2):719-726.

doi:10.1118/1.3539749

186. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Med Phys*. 2020;47(9):e929-e950. doi:10.1002/mp.14320

187. Ge Y, Wu QJ. Knowledge-based planning for intensity-modulated radiation therapy: A review of data-driven approaches. *Med Phys*. 2019;46(6):2760-2775. doi:10.1002/mp.13526

188. Bohoudi O, Bruynzeel AME, Senan S, et al. Fast and robust online adaptive planning in stereotactic MR-guided adaptive radiation therapy (SMART) for pancreatic cancer. *Radiother Oncol*. 2017;125(3):439-444. doi:10.1016/j.radonc.2017.07.028

189. Furhang EE, Dolan J, Sillanpaa JK, Harrison LB. Automating the initial physics chart-checking process. *J Appl Clin Med Phys*. 2009;10(1):129-135. doi:10.1120/jacmp.v10i1.2855

190. Holdsworth C, Kukluk J, Molodowitch C, et al. Computerized system for safety verification of external beam radiation therapy planning. *Int J Radiat Oncol Biol Phys*. 2017;98(3):691-698. doi:10.1016/j.ijrobp.2017.03.001

191. Yang J, Veeraraghavan H, Armato SG, et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Med Phys*. 2018;45(10):4568-4581. doi:10.1002/mp.13141

192. Zabel WJ, Conway JL, Gladwish A, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. *Pract Radiat Oncol*. 2021;11(1):e80-e89. doi:10.1016/j.prro.2020.05.013

193. Brouwer CL, Boukerroui D, Oliveira J, et al. Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. *Phys Imaging Radiat Oncol*. 2020;16:54-60. doi:10.1016/j.phro.2020.10.001

194. Zhang X, Li X, Quan EM, Pan X, Li Y. A methodology for automatic intensity-modulated radiation treatment planning for lung cancer. *Phys Med Biol*. 2011;56(13):3873-3893. doi:10.1088/0031-9155/56/13/009

195. Hazell I, Bzdusek K, Kumar P, et al. Automatic planning of head and neck treatment plans. *J Appl Clin Med Phys*. 2016;17(1):272-282. doi:10.1120/jacmp.v17i1.5901

196. Mitchell RA, Wai P, Colgan R, Kirby AM, Donovan EM. Improving the efficiency of breast radiotherapy treatment planning using a semi-automated approach. *J Appl Clin Med Phys*. 2017;18(1):18-24. doi:10.1002/acm2.12006

197. Hansen CR, Bertelsen A, Hazell I, et al. Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans. *Clin Transl Radiat Oncol*. 2016;1:2-8. doi:10.1016/j.ctro.2016.08.001

198. Kisling K, Zhang L, Simonds H, et al. Fully automatic treatment planning for external-beam radiation therapy of locally advanced cervical cancer: A tool for low-resource clinics. *J Glob Oncol*. 2019;2019(5):1-8. doi:10.1200/jgo.18.00107

199. Lee H, Lee E, Kim N, et al. Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region. *Front Oncol*. 2019 Apr 9;9:239. doi:10.3389/fonc.2019.00239

200. Fogliata A, Reggiori G, Stravato A, et al. RapidPlan head and neck model: The objectives and possible clinical benefit. *Radiat Oncol*. 2017;12(1):1-12. doi:10.1186/s13014-017-

0808-x

201. Kusters JMAM, Bzdusek K, Kumar P, et al. Automated IMRT planning in Pinnacle: A study in head-and-neck cancer. *Strahlentherapie und Onkol*. 2017;193(12):1031-1038. doi:10.1007/s00066-017-1187-9

202. Yoder T, Hsia AT, Xu Z, Stessin A, Ryu S. Usefulness of EZFluence software for radiotherapy planning of breast cancer treatment. *Med Dosim*. 2019;44(4):339-343. doi:10.1016/j.meddos.2018.12.001

203. Liew C. The future of radiology augmented with Artificial Intelligence: A strategy for success. *Eur J Radiol*. 2018;102:152-156. doi:10.1016/j.ejrad.2018.03.019

204. Neri E, de Souza N, Brady A, et al. What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imaging*. 2019;10(1). doi:10.1186/s13244-019-0738-2

205. Olubunmi Afolabi M, Oyedepo Oyebisi T. Pharmacists' perceptions of barriers to automation in selected hospital pharmacies in Nigeria. *J Pharm Pract*. 2007;20(1):64-71. doi:10.1177/0897190007302894

206. Glasgow RE, Vinson C, Chambers D, Khoury MJ, Kaplan RM, Hunter C. National institutes of health approaches to dissemination and implementation science: Current and future directions. *Am J Public Health*. 2012;102(7):1274-1281. doi:10.2105/ajph.2012.300755

207. Krosnick JA. Survey research. *Annual review of psychology*. 1999;50(1):537-567.

208. Krosnick JA, Berent MK. Comparisons of party identification and policy preferences: The impact of survey question format. *Am J Pol Sci*. 1993;37(3):941. doi:10.2307/2111580

209. Kamoen N, Holleman B, Mak P, Sanders T, van den Bergh H. Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Process*. 2011;48(5):355-385. doi:10.1080/0163853X.2011.578910

210. Visser PS, Krosnick JA, Marquette J, Curtin M. Mail surveys for election forecasting? An evaluation of the Columbus Dispatch Poll. *Public Opin Q*. 1996;60(2):181. doi:10.1086/297748

211. Peytchev A. Consequences of survey nonresponse. *Ann Am Acad Pol Soc Sci*. 2013;645(1):88-111. doi:10.1177/0002716212461748

212. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org

213. Chen W, Li Y, Dyer BA, et al. Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat Oncol*. 2020;15(1):176. doi:10.1186/s13014-020-01617-0

214. Ahn SH, Yeo AU, Kim KH, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiat Oncol*. 2019;14(1):1-13. doi:10.1186/s13014-019-1392-z

215. Kaderka R, Gillespie EF, Mundt RC, et al. Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. *Radiother Oncol*. 2019;131:215-220. doi:10.1016/j.radonc.2018.07.013

216. Habraken SJM, Sharfo AWM, Buijsen J, et al. The TRENDY multi-center randomized trial on hepatocellular carcinoma – Trial QA including automated treatment planning and benchmark-case results. *Radiother Oncol*. 2017;125(3):507-513.

doi:10.1016/j.radonc.2017.09.007

217. Skitka LJ, Mosier K, Burdick MD. Accountability and automation bias. *Int J Hum Comput Stud*. 2000;52(4):701-717. doi:10.1006/ijhc.1999.0349

218. Mosier KL, Skitka LJ. Automation use and automation bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 1999;43(3):344-348. doi:10.1177/154193129904300346

219. Gopan O, Zeng J, Novak A, Nyflot M, Ford E. The effectiveness of pretreatment physics plan review for detecting errors in radiation therapy. *Med Phys*. 2016;43(9):5181-5187. doi:10.1118/1.4961010

220. Autor DH. Why are there still so many jobs? The history and future of workplace automation. *JEP*. 2015;29(3):3-30. doi:10.1257/jep.29.3.3

221. Wajcman J. Automation: is it really different this time? *Br J Sociol*. 2017;68(1):119-127. doi:10.1111/1468-4446.12239

222. Brehm JO. (1993). *The Phantom Respondents: Opinion Surveys and Political Representation*. Ann Arbor: The University of Michigan Press.

223. Moore KL, Kagadis GC, McNutt TR, Moiseenko V, Mutic S. Vision 20/20: Automation and advanced computing in clinical radiation oncology. *Med Phys*. 2014;41(1):010901. doi:10.1118/1.4842515

224. Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. *Br J Radiol*. 2019;92(1100). doi:10.1259/bjr.20190001/

225. Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opin Q*. 1987;51(2):201. doi:10.1086/269029

226. Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): A tool to measure the value of quality control checks in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2012;84(3):e263-e269. doi:10.1016/j.ijrobp.2012.04.036

227. Azmandian F, Kaeli D, Dy JG, et al. Towards the development of an error checker for radiotherapy treatment plans: A preliminary study. *Phys Med Biol*. 2007;52(21):6511-6524. doi:10.1088/0031-9155/52/21/012

228. Xia J, Mart C, Bayouth J. A computer aided treatment event recognition system in radiation therapy. *Med Phys*. 2013;41(1):011713. doi:10.1118/1.4852895

229. Olsen LA, Robinson CG, He GR, et al. Automated radiation therapy treatment plan workflow using a commercial application programming interface. *Pract Radiat Oncol*. 2014;4(6):358-367. doi:10.1016/j.prro.2013.11.007

230. Yang D, Moore KL. Automated radiotherapy treatment plan integrity verification. *Med Phys*. 2012;39(3):1542-1551. doi:10.1118/1.3683646

231. Yang D, Wu Y, Brame RS, et al. Technical Note: Electronic chart checks in a paperless radiation therapy clinic. *Med Phys*. 2012;39(8):4726-4732. doi:10.1118/1.4736825

232. Ford E, Conroy L, Dong L, et al. Strategies for effective physics plan and chart review in radiation therapy: Report of AAPM Task Group 275. *Med Phys*. 2020;47(6):e236-e272. doi:10.1002/mp.14030

233. Schofield DL, Conroy L, Harmsen WS, et al. AAPM task group report 275.S: Survey strategy and results on plan review and chart check practices in US and Canada. *J Appl Clin Med Phys*. 2023;24(4):e13952. doi:10.1002/acm2.13952

234. Xia P, Sintay BJ, Colussi VC, et al. Medical Physics Practice Guideline (MPPG) 11.a: Plan and chart review in external beam radiotherapy and brachytherapy. *J Appl Clin Med Phys*. 2021;22(9):4-19. doi:10.1002/acm2.13366

235. Smock C, Alemagno S. Understanding health care provider barriers to hospital affiliated medical fitness center facility referral: A questionnaire survey and semi structured interviews. *BMC Health Serv Res*. 2017;17(1):1-6. doi:10.1186/s12913-017-2474-y

236. Seedat F, Hargreaves S, Friedland JS. Engaging new migrants in infectious disease screening: A qualitative semi-structured interview study of UK migrant community health-care leads. *PLoS One*. 2014;9(10):e108261. doi:10.1371/journal.pone.0108261

237. Ismayilova M, Yaya S. What can be done to improve polycystic ovary syndrome (PCOS) healthcare? Insights from semi-structured interviews with women in Canada. *BMC Womens Health*. 2022;22(1):157. doi:10.1186/s12905-022-01734-w

238. Burr O, Berry A, Joule N, Rayman G. Inpatient diabetes care during the COVID-19 pandemic: A Diabetes UK rapid review of healthcare professionals' experiences using semi-structured interviews. *Diabet Med*. 2021;38(1):e14442. doi:10.1111/dme.14442

239. Watt JA, Fahim C, Straus SE, Goodarzi Z. Barriers and facilitators to virtual care in a geriatric medicine clinic: A semi-structured interview study of patient, caregiver and healthcare provider perspectives. *Age Ageing*. 2022;51(1):1-9. doi:10.1093/ageing/afab218

240. Neylon J, Luximon DC, Ritter T, Lamb JM. Proof-of-concept study of artificial intelligence-assisted review of CBCT image guidance. *J Appl Clin Med Phys*. Published online May 10, 2023:e14016. doi:10.1002/acm2.14016

241. Lumivero (2020). *NVivo* (Version 14). www.lumivero.com

242. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa

243. Campbell KA, Orr E, Durepos P, et al. Reflexive thematic analysis for applied qualitative health research. *Qual Rep*. 2021;26(6):2011-2028. doi:10.46743/2160-3715/2021.5010

244. Bowen PA, Edwards PJ, Cattell K. Corruption in the South African construction industry: a thematic analysis of verbatim comments from survey participants. *Constr Manag Econ.* 2012;30(10):885-901. doi:10.1080/01446193.2012.711909

245. Shafi S, Mallinson DJ. The potential of smart home technology for improving healthcare: A scoping review and reflexive thematic analysis. *Hous Soc*. Published online 2021. doi:10.1080/08882746.2021.1989857

246. Corr L, Rowe H, Fisher J. Mothers' perceptions of primary health-care providers: thematic analysis of responses to open-ended survey questions. *Aust J Prim Health*. 2015;21(1):58. doi:10.1071/py12134

247. Semaan A, Audet C, Huysmans E, et al. Voices from the frontline: Findings from a thematic analysis of a rapid online global survey of maternal and newborn health professionals facing the COVID-19 pandemic. *BMJ Glob Heal*. 2020;5(6). doi:10.1136/bmjgh-2020-002967

248. Karavadra B, Stockl A, Prosser-Snelling E, Simpson P, Morris E. Women's perceptions of COVID-19 and their healthcare experiences: A qualitative thematic analysis of a national survey of pregnant women in the United Kingdom. *BMC Pregnancy Childbirth*. 2020;20(1):600. doi:10.1186/s12884-020-03283-2

249. Clarke v, Braun V, Hayfield N. (2015). Thematic Analysis. In JA Smith (Ed.). *Qualitative Psychology: A Practical Guide to Research Methods* (3rd ed., 222-248). SAGE Publications Ltd.

250. Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Heal*. 2019;11(4):589-597. doi:10.1080/2159676X.2019.1628806

251. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nurs Health Sci*. 2013;15(3):398-405. doi:10.1111/nhs.12048

252. Braun V, Clarke V. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qual Res Psychol*. 2021;18(3):328-352. doi:10.1080/14780887.2020.1769238

253. Terry G, Hayfield N, Clarke V, Braun V. (2017). Thematic Analysis. In C Willig & W Stainton-Rogers (Eds.). *The SAGE Handbook of Qualitative Research in Psychology* (2nd ed., 17-37). SAGE Publications Ltd.

254. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field methods*. 2006;18(1):59-82. doi:10.1177/1525822X05279903

255. Burmeister JW, Busse NC, Cetnar AJ, et al. Academic program recommendations for graduate degrees in medical physics: AAPM Report No. 365 (Revision of Report No. 197). *J Appl Clin Med Phys*. 2022;23(10):e13792. doi:10.1002/acm2.13792

256. Mazur LM, Mosaly PR, Jackson M, et al. Quantitative assessment of workload and stressors in clinical radiation oncology. *Int J Radiat Oncol Biol Phys*. 2012;83(5):e571-

e576. doi:10.1016/j.ijrobp.2012.01.063

257. Johnson J, Ford E, Yu J, Buckey C, Fogh S, Evans SB. Peer support: A needs assessment for social support from trained peers in response to stress among medical physicists. *J Appl Clin Med Phys*. 2019;20(9):157-162. doi:10.1002/acm2.12675

258. Petragallo R, Bardach N, Ramirez E, Lamb JM. Barriers and facilitators to clinical implementation of radiotherapy treatment planning automation: A survey study of medical dosimetrists. *J Appl Clin Med Phys*. 2022;23(5):e13568. doi:10.1002/acm2.13568

259. Samei E, Pawlicki T, Bourland D, et al. Redefining and reinvigorating the role of physics in clinical medicine: A report from the AAPM Medical Physics 3.0 Ad Hoc Committee. *Med Phys*. 2018;45(9):e783-e789. doi:10.1002/mp.13087

260. Abravan A, Correia D, Gasnier A, et al. Qualitative study on Diversity, Equity, and Inclusion within radiation oncology in Europe. *Int J Radiat Oncol Biol Phys*. 2023;116(2):246-256. doi:10.1016/j.ijrobp.2023.02.009

261. Paradis KC, Ryan KA, Schmid S, et al. A qualitative investigation of resilience and well-being among medical physics residents. *J Appl Clin Med Phys*. 2022;23(3):e13554. doi:10.1002/acm2.13554